ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ БЮДЖЕТНОЕ УЧРЕЖДЕНИЕ НАУКИ
ИНСТИТУТ ВЫЧИСЛИТЕЛЬНОЙ МАТЕМАТИКИ
И МАТЕМАТИЧЕСКОЙ ГЕОФИЗИКИ
СИБИРСКОГО ОТДЕЛЕНИЯ РОССИЙСКОЙ АКАДЕМИИ НАУК

НЕКОММЕРЧЕСКОЕ ПАРТНЕРСТВО
“НАЦИОНАЛЬНЫЙ ЭЛЕКТРОННО-ИНФОРМАЦИОННЫЙ КОНСОРЦИУМ”

С. В. Бредихин, А. Ю. Кузнецов, Н. Г. Щербакова

АНАЛИЗ ЦИТИРОВАНИЯ В БИБЛИОМЕТРИИ

Новосибирск, Москва, 2013

УДК 001.12+303.2
ББК 72+78

Бредихин С. В., Кузнецов А. Ю., Щербакова Н. Г. Анализ цитирования в библиометрии. Новосибирск: ИВМиМГ СО РАН, НЭИКОН, 2013. 344 с.

Книга посвящена изучению взаимных связей, возникающих в процессе акта цитирования между цитирующим и цитируемым акторами. Рассмотрены социальные и психологические аспекты цитирования. Определены базовые свойства и характеристики цитат. Приведены основные методы анализа цитирования и построения карт науки. Представлен обзор современных исследований в области *h*-технологий.

Книга снабжена актуальным библиографическим материалом, некоторые ее главы могут быть использованы в качестве учебного пособия и (или) справочника по курсу “библиометрия”. Предназначена для библиотечных работников, ученых и администраторов, занимающихся планированием научного труда, а также специалистов по теории и практике распределенных информационных систем.

Bredikhin S. V., Kuznetsov A. Yu., Scherbakova N. G. Citation analysis in bibliometrics. Novosibirsk: ICM&MG SB RAS, NEIKON, 2013. 344 p.

The book is devoted to studying of the links arising during the process of citation between the citing and the cited actors. The social and psychological aspects of citing behavior are considered. The basic properties and characteristics of citations are defined. The main methods of citation analysis and science maps generation are provided. In the final part the overview of the modern researches in the field of application of *h*-related technologies are presented.

Up-to-date reference lists are included in the book. Some of the chapters can be used as a textbook or a reference book for the bibliometrics course. The book is intended to the library workers, scientists and managing officers that deal with the planning of the scientific work process, as well as any specialists interested in the questions of distributed information systems theory and practice.

ISBN 978-5-91907-009-2

Работа выполнена в рамках государственного контракта № 07.551.11.4002 между Министерством образования и науки Российской Федерации и Некоммерческим партнерством “Национальный электронно-информационный консорциум” по теме “Поддержка и расширение системы обеспечения новыми информационными технологиями участников Федеральной целевой программы “Исследования и разработки по приоритетным направлениям развития научно-технологического комплекса России на 2007–2013 гг.”.

ПРЕДИСЛОВИЕ

Издавна в научном мире авторы “общались” с помощью библиографических ссылок, которые являются неотъемлемой частью научных документов. Как только в свет выходит научная работа, появляется вероятность того, что она будет прочитана и, возможно, процитирована. Для выполнения акта цитирования цитирующий автор не должен спрашивать у цитируемого автора разрешения. На этом основан феномен цитирования, состоящий в том, что в результате акта цитирования между цитирующим и цитируемым авторами вне зависимости от желания последнего возникает взаимная связь, которую в западной литературе принято называть коммуникацией.

Настоящая книга посвящена вопросам эволюции научных коммуникаций, порождаемых библиографическими ссылками. В библиометрии этот раздел исследований, основанный на интерпретации данных о цитировании, называют анализом цитирований (АЦ). Различные методы исследований, применяемые для АЦ, можно условно разделить на две части: анализ статистики цитирования публикаций и анализ коммуникаций цитирования. Статистика цитирования позволяет выявлять закономерности развития науки, вероятные темпы развития и “прорывы”. Анализ коммуникаций цитирования – библиометрический инструмент, позволяющий, с одной стороны, выявлять дисциплинарную структуру науки и обнаруживать зарождающиеся направления исследований, с другой – проводить количественную оценку научных исследований.

Пик интереса к АЦ следует отнести ко второй половине XX в., когда границы предмета АЦ были существенно расширены, он преобразовался в нечто большее, чем “арифметика числа цитирований”, появилась надежда создать “теорию цитирования”,

разработаны механизмы ранжирования научных документов. Основным инструментом АЦ с 1961 г. является индекс научного цитирования *Science Citation Index* (*SCI*), который с 2006 г. стал называться *ISI Web of Knowledge* (*WoK/ISI*). *WoK/ISI* на протяжении 40 лет владел монополией на рынке библиографических баз данных (ББД) и информационных услуг на их основе, вплоть до развития веб-технологий, оказавших существенное влияние на исследования в области АЦ. Во многом благодаря этим технологиям появились международные информационные системы, основанные на ББД: *Science Direct, Medline, MathSciNet* и т. п., – а также национальные системы в Китае, Японии, Индии и РФ. На сегодняшний день наиболее авторитетной международной ББД является *Web of Science* (далее WoS).

По мере того как международные и национальные индексы научного цитирования заработали в "промышленном" режиме, результаты АЦ стали востребованы администраторами науки и руководителями научных фондов как один из инструментов принятия решений относительно их важности в рассматриваемой области исследований. Таким образом, в конце XX в. возникла ситуация, когда ученые, труды которых были проиндексированы в ББД указанных информационных систем и которые имели к ним доступ, могли в автоматическом режиме получить ответ на вопрос: "Сколько раз процитировали мою работу *A* и кто именно это сделал?". Безусловно, ответ на этот вопрос имеет субъективный характер, поскольку справедливо предположить, что существуют публикации, авторы которых процитировали работу *A*, однако их публикации не были проиндексированы ни в одной ББД и поэтому не были учтены.

Далеко не все считают, что АЦ является основанием для дискуссии о научной работе. Критики справедливо указывают на манипулируемость и непрозрачность механизмов вычисления индек-

сов цитирования, а также на главные проблемы. Например, омография, т. е. невозможность разделить цитирования независимых ученых, которые имеют одинаковые фамилии и инициалы, намеренное взаимное цитирование между коллегами или самоцитирование и т. п. Защитники АЦ, в свою очередь, утверждают, что, несмотря на неоднозначность, результаты АЦ являются достаточно объективной мерой производительности труда ученого [Гарфилд, 1982].

Современного научного работника характеризует его когнитивная система, в которую входят научные данные, теоретические представления и методы исследований [Кара-Мурза, 1989]. Согласно трудновыполнимому требованию к ученому, сформулированному Р. Мертоном, "… он (научный работник) должен как можно быстрее передавать свои научные результаты коллегам, но он не должен торопиться с публикациями". Как соответствовать этому требованию, когда над тобой висит как "дамоклов меч" тезис Р. Мертона: "Публикуйся или умри"? Возникшее противоречие разрешается, например, применением принципа Черной Королевы:[*] "… нужно бежать изо всех сил, только для того, чтобы остаться на месте", что в нашем случае означает продолжать интенсивно публиковаться, а следовательно, и цитировать.

И. В. Маршакова-Шайкевич в энциклопедии [ЭЭиФН, 2009] дает следующее определение термина "цитирование": "… это факт бытия науки и объект науковедческого исследования, надежность которого определяется самой традицией науки как социального института. Феномен цитирования является неоспоримо важной этической нормой в науке, общенаучным регулятором (на языке

[*]Героиня сказки "Приключения Алисы в Стране чудес" (*Alice's Adventures in Wonderland* , часто используется сокращенный вариант "Алиса в Стране чудес"), написанной английским математиком, поэтом и писателем Чарльзом Лютвиджом Доджсоном под псевдонимом Льюис Кэррол и изданной в 1865 г. (перевод Н. М. Демуровой, 1966).

философии науки) и, безусловно, одним из важных средств научной коммуникации". Именно с этой точки зрения в данной книге рассматриваются вопросы коммуникаций, возникающие в результате акта цитирования между взаимодействующими акторами.

В этой книге в форме обзора рассматриваются проблемы анализа цитирования в библиометрии. Она состоит из семи глав, каждая из которых представляет аналитический обзор по одной из тем АЦ и может рассматриваться как самостоятельная часть. Но если читатель осилит все главы подряд, то сможет почувствовать связь, распространяющуюся от главы к главе, и обнаружить, что он держит в руках нечто отличное от справочника. По крайней мере, именно так мы стремились построить изложение. Материал книги не претендует на полноту, а результат выбора тем для глав соответствует афоризму К. Пруткова[**].

В первой главе даны определения терминов "ссылка" и "цитирование", рассмотрены их основные функции и свойства. Представлено видение проблемы АЦ, построенное на суждениях Ю. Гарфилда, Й. Николайсена, Г. Смолла, Б. Кронина, Л. Лаудана и др. Далее цитирование рассматривается как психологический процесс, в основу положены работы И. В. Маршаковой-Шайкевич, С. Вулгара, Б. Латура, а также психологов бихевиористской и гештальтской школ. Затем рассматривается "нормативная теория цитирования", в основе которой лежат идеи Дж. Раветца, Н. Каплана, С. Балди и Р. Мертона.

Во второй главе процесс цитирования рассматривается как социальное действие. Вопрос "Почему именно эта работа (среди прочих равных) была выбрана цитирующим автором в качестве библиографической ссылки (иными словами, была признана релевантной)", является фундаментальной нерешенной проблемой

[**] "Никто не обнимет необъятного". Из собрания мыслей и афоризмов Козьмы Пруткова «Плоды раздумья», 1854.

в области анализа цитирования. Ряд авторов посвятили свои работы изучению этой проблемы с точки зрения мотивации цитирования. В работах М. Моравчика и П. Мургесана, Д. Чубина и С. Мойтры предприняты попытки создания схемы классификации библиографических ссылок. Существенный вклад в изучение проблемы цитирования внес Б. Кронин своей монографией "Процесс цитирования", однако окончательная ясность в этой проблеме не была достигнута.

Намерения цитирующего автора также изучены недостаточно. М. Вейншток считает, что причины, по которым авторы цитируют работы других исследователей, можно разделить на две категории: "серьезные" и "незначительные". Позднее Ф. Торн составляет свой перечень причин цитирования, который можно рассматривать как дополнение к работе М. Вейнштока. К концу 70-х гг. XX в. в работах М. Лайна и А. Сандисона, Дж. и С. Коулов, Дж. Гильберта и Г. Бавелас были созданы предпосылки для создания шкалы оценок мотивации цитирования.

В основе ряда процедур АЦ лежат предположения относительно свойств цитирования и проблем, касающихся использования данных о цитировании. Наиболее известным из них является принадлежащий Л. Смит тезис: "Все цитаты равны", – который ставит "точку" во многих дискусах. Рассматриваются вопросы распределения цитирований для работ, выполненных в соавторстве и при самоцитировании. Обсуждаются аргументы "за" и "против" теории цитирования, негласным модератором этой темы является П. Воутерс.

Третья глава посвящена вопросам продуцирования цитат и их статистическим характеристикам, которые эмпирическим путем получены Д. Прайсом в середине XX в. Далее обсуждается вопрос о том, что измеряет цитирование. Особое внимание заслуживает мнение Х. Моеда, рассматривающего этот вопрос с нескольких

точек зрения: физической, социологической, психологической, исторической и информационной. Здесь же представлен анализ различных точек зрения авторитетных ученых в области библиометрии на проблему взаимосвязи между ссылками и цитированиями и рассмотрена модель цитирования Симкина – Ройчодхури.

В четвертой главе представлен основной аналитический инструментарий, используемый в процессе АЦ. Во-первых, это графы, матрицы цитирования и некоторые результаты М. Кохена, обеспечивающие формальное представление таким часто используемым понятиям, как количество соавторов, количество публикаций за определенный период и т. п. Во-вторых, векторная модель для формального представления текстовой информации. В-третьих, фундаментальные методы библиографического сочетания (М. Кесслер) и коцитирования (И. В. Маршакова, Г. Смолл). Подробно рассмотрена круговая модель мультицитирования. Завершают эту главу вопросы анализа истории коцитирования и контекста цитирования.

Пятая глава полностью посвящена кластерному анализу – методу классификации исследуемых объектов по “осмысленным” группам на основе информации об их цитировании. Основное внимание сосредоточено на представлении “сырых” данных в виде матрицы, описывающей объекты исследования и их параметры. Далее рассматриваются прямые, теоретико-множественные (косинусная мера, индексы включаемости и Жаккара), вероятностная (коэффициент ассоциативности) меры подобия документов, формально определены функции сходства и расстояния. В заключение приведена структура системы автоматической кластеризации и ранжирования множества документов, являющихся результатом извлечения информации из библиографической базы данных.

В шестой главе рассматривается процесс “картографирования науки”, который позволяет изучать отрасли науки и междисцип-

линарные научные коммуникации на основе разнообразных форм анализа цитирования. Приведены этапы построения карты науки: извлечение данных, определение единиц анализа, выбор метрик и мер подобия. Далее рассматриваются методы анализа и размещения данных, среди которых особое место занимают методы преобразования данных с целью понижения их размерности: сингулярное разложение, факторный анализ и метод главных компонент, многомерное шкалирование, латентный семантический анализ, сети *PFnet,* самоорганизующиеся карты Кохена, техника триангуляции, силовые алгоритмы. Еще один параграф посвящен картам *ISI*; в нем рассматриваются иерархия кластеров, комбинированная мера сходства, позиционирование объектов и создание общего координатного пространства с помощью системы *SciViz*. Обсуждаются вопросы сравнения карт науки. В заключение приведены краткие сведения о пакетах программ визуализации многомерных данных *Pajek* и *VOSviewer.*

В седьмой главе представлено современное состояние работ в области библиометрии, основанной Х. Хиршем: аксиоматическое определение *h*-индекса и его основные свойства; три модели, интерпретирующие этот индекс, авторами которых являются Х. Хирш, А. Шуберт и В. Гланзель, Л. Егге и Р. Руссо. Завершает главу материал об *h*-последовательностях – современном аналитическом инструменте, позволяющем расширить прикладные границы изучаемой области.

В качестве читателей мы видим тех, у кого хватило терпения ознакомиться с нашей предыдущей работой [Бредихин, Кузнецов, 2012], в которой, помимо прочего, достаточно подробно рассмотрен ряд метрик ранжирования периодических изданий, в том числе *Eigenfactor*, *SNIP*, *SJR*. По этой причине здесь эти метрики не рассматриваются.

С. В. Бредихин, А. Ю. Кузнецов, Н. Г. Щербакова

Глава 1. Социальные аспекты цитирования

История науки, подобно истории
всех человеческих идей, есть история
безотчетных грез, упрямства и ошибок.

Карл Поппер [*Поппер, 1983*]

1.1. Ссылки и цитирования: свойства и функции

Термины “ссылка” и “цитирование” являются главными в данной работе. Поэтому изложение материала начинается с определения этих терминов.

В современном англо-русском переводе слова *Citation, Citations* звучат довольно официально. Например, слово *Citation* имеет значение: а) документ или речь, содержащие восхваление некоторого лица за совершение смелого или другого особо значимого поступка; б) вызов в суд или требование предстать перед законом; в) цитирование какого либо письменного документа. Слова *Cite, Cites, Citing, Cited* (формы глагола “цитировать”) также носят официальный характер.

Цитировать (*Cite*) что-то означает: а) вы привлекаете внимание к этому, в частности в качестве примера или подтверждения того, что вы утверждаете; б) вы цитируете письменный труд, в частности в качестве примера или подтверждения того, что вы утверждаете.

Цитировать (*Cite*) кого-то означает: а) официально вызвать его в суд; б) официально упомянуть его в отчете или другом документе за совершение смелого или другого особо значимого поступка.

Автор работы [Smith, 1981] в процессе изучения трансформации ссылки в соответствующее цитирование выдвигает предположение, что все цитаты обладают единым качеством: “Все цитаты равны”. Это удобное качество позволяет вводить цитаты из

той или иной статьи, в результате чего возникает так называемая частота цитирования. Знак цитирования не разделяет данное свойство со своим материнским знаком ссылки. Ссылки не равны: они выполняют разные функции относительно цитируемого текста, различается и лежащая в их основе мотивация.

Главная функция ссылки – функция указателя. Другое дело – цитата. Данный знак может функционировать одновременно как указатель и как количество, но именно последняя функция, а не первая, наиболее важна. Кроме того, с точки зрения цитирования, это делается в силу выполняемой им роли фундаментальной единицы частоты цитирования. Другими словами, указание играет вспомогательную роль по отношению к индексу при анализе цитирования. Может быть измерена не только частота цитирования определенной статьи, данная частота может быть также суммирована на более высоких уровнях агрегирования для получения частот цитирования исследовательских групп, учреждений, журналов, стран и даже дисциплин и областей науки в целом.

Поскольку может быть измерена частота цитирования каждой статьи – если она не цитируется, ее частота цитирования равна нулю – любая статья может быть сравнена с любой другой, независимо от их предмета. Если бы цитирование не имело данного качества, индекс цитирования мог бы быть менее полезным в качестве библиографического инструмента.

Роль цитирования можно также сравнить с ролью денег, в частности, если принять во внимание оценочное использование библиометрии. Если ценность той или иной статьи выразить в виде частоты ее цитирования, цитирование становится “валютой науки”. Автор работы [Wouters, 1999b] приводит следующий пример. В 1997году в Техническом университете Анкары была установлена прямая зависимость между частотой цитирования и деньгами. Исследователи могли зарабатывать деньги за счет

цитирования. Если кто-то набрал определенный объем цитирования, он получал право занять должность профессора.

Другой причиной, по которой индексы цитирования настолько полезны, является их независимость от тематических дескрипторов [Egghe, Rousseau, 1990]. В данном случае центральную роль играет универсальность, достигаемая двояким путем. Сначала, при преобразовании ссылки в цитату, устраняется локальный контекст цитируемого документа. Затем устраняется локальный контекст индексируемого учреждения, хотя и не полностью.

Цитата имеет и другие общие свойства с денежными знаками: функционировать должным образом только среди других цитат. В этой связи цитирование требует массового продуцирования. Цитата в единственном экземпляре бессмысленна. Она приобретает свои функциональные свойства, главным образом, в связи с другими цитатами.

Если попытаться составить карту науки или оценить ее, потребуется большой объем данных цитирования. Это свойство цитат хорошо известно, в работе [Luukkonen,1990] говорится об этом следующее: «Базовая проблема ... состоит в том, что “качество” труда измеряется его цитированием, т. е. на основе индикатора, требующего проверки и исследования. Это предполагает доказательство по принципу замкнутого круга: большинство наиболее цитируемых ученых являются наиболее цитируемыми по той причине, что они являются наиболее цитируемыми, т. е. хорошо потому, что часто цитируется».

Как уже отмечалось, цитата в единственном экземпляре ничего не значит. Если рассматривать ее как изолированный знак, она не обладает многими качествами. Практически единственное качество – это ее фундаментальное равенство со всеми другими цитатами. Более интересны свойства, возникающие при взаимодействии между цитатами, а также между цитатами и ссылками.

Многие библиометрические исследования посвящены раскрытию и интерпретации потенциальных моделей цитирования и сетей цитирования в науке.

Одна из этих тем – способ, с помощью которого "время" присутствует в представлении науки при цитировании. И здесь следует подчеркнуть важное преимущество индекса цитирования в качестве поискового инструмента: если взять публикацию, можно проследить ее "потомков" (статьи, в которых цитируется данная публикация) до последнего времени. Если прослеживание ссылок лишь уводит все дальше и дальше в прошлое (поскольку данный метод позволяет выйти лишь на "предков", т. е. статьи, процитированные в соответствующей публикации), индекс цитирования приближает нас к настоящему. Причина этого очевидна: цитирующий текст, естественно, является более свежим, чем соответствующий цитируемый текст. Пожалуй, единственным исключением из этого правила являются ссылки на ожидаемые публикации.

По этой причине частота цитирования той или иной статьи является динамическим свойством. Она может измениться в любой момент. В результате возникает два вопроса. Во-первых, сети цитирования всегда ретроспективны; они по определению обращены в прошлое. Во-вторых, существующая литература, из которой возникают цитаты, по определению статична. Любая статья, независимо от времени ее появления, в принципе может цитироваться. Время зримо присутствует в сети цитирования не как что-то текущее или движение сети, но как отсутствие определенных отношений цитирования (отсутствие в старых статьях цитат из более свежих статей) - другими словами, как структура. Это ограничение, налагаемое ссылками на цитирование.

Научная традиция требует, чтобы ученый в статье ссылался на более ранние работы, относящиеся к теме исследования. В основе

продуцирования цитат лежит предположение, что ссылки идентифицируют предшествующих авторов, чьи концепции, методы, инструменты и т. д. используются автором данной работы.

Ссылка и цитирование. В общем случае, следует различать термины "ссылка" (*Reference, R*) и "цитирование" (*Citation, C*). Если документ d_2 содержит упоминание и библиографические признаки, описывающие документ d_1, то d_2 содержит ссылку (R) на d_1, а d_1 имеет цитирование (C) от d_2 [Price, 1970]. Другими словами, ссылка – это признательность по отношению к другому автору, а цитирование – это признательность, получаемая автором от других. Т. е. ссылка означает взгляд назад, а цитирование – взгляд вперед (рис. 1.1.1).

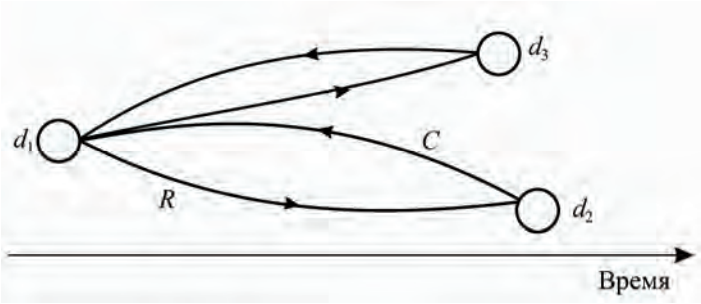

Рис. 1.1.1. Ссылка и цитирование

На рис. 1.1.1. знак "←" (стрелка влево, взгляд назад) означает "ссылается на" или "цитирует"; знак "→" (стрелка вправо, взгляд вперед) означает "процитирован". Документ d_1 содержится в списке ссылок документов d_2 и d_3 (стрелки влево); документ d_1 цитируется в документах d_2 и d_3 (стрелки вправо). Ссылки различны, они выполняют разные функции и имеют разную мотивацию.

Итак, следует различать знаки ссылки и цитирования. Последнее происходит из первого. Это означает, что семиозис цитирования является операцией второго порядка в отношении создания ссылки ученым. Вследствие этого различные вмешательства в

данный процесс продуцирования будут влиять на результат. Во-первых, очевидно, невозможно обработать каждую ссылку, сделанную в каждой научной статье в мире. Следовательно, необходим избирательный подход, который, естественно, будет влиять на полученный в результате индекс.

Для большинства целей действительно важно, представляет ли цитата “корректную” инверсию ссылки: семиозис цитирования является точной операцией. Вовлеченные в нее акторы, как специалисты по библиометрии, так и составители индекса, обычно контролируют идентичность ссылки и цитирования: они должны соответствовать друг другу. С точки зрения данного исследования, важна не идентичность, а “достоверная инверсия”, т. е. инверсия без любых других изменений. Ни одна операция не может пройти идеально (даже с привлечением цифрового копирования), что также касается определения индекса цитирования. Поэтому индексы неизбежно содержат “ошибки”. Насколько они серьезны, зависит от использования индекса. Любой анализ цитирования с обработкой больших объемов данных цитирования содержит значительное количество таких “ошибок”, которые сложно идентифицировать. В этой связи постоянно ведется дискуссия о “качестве данных” в области библиометрии. Фактически в рамках общей темы рассматривается три разных проблемы: качество ссылки, выбор ссылки и целостность инверсии.

Приведем пример установления библиографической связи, природой которого является акт цитирования. Библиографическая связь представляет собой совокупность библиографических сведений о документе d_1, на который есть ссылка в документе d_2, необходимых для его идентификации и поиска (документа d_1). Публикация документа d_2, например автора *Петрова*, содержащая библиографическую ссылку на документ d_1 автора *Иванова*,

представляет собой адресованное научному сообществу сообщение о том, что автор *Петров* в процессе создания документа d_2, был знаком с документом d_1 *Иванова* или “слышал” о нем и не более того, если дополнительные разъяснения *Петрова* на этот счет отсутствуют. В этом случае принято говорить, что документ d_2 библиографически связан с документом d_1 или что *Петров* в документе d_2 процитировал документ d_1 автора *Иванова*.

Столь изысканная форма общения в мире науки практикуется уже не один век, и не видно предпосылок тому, чтобы она изменилась. При такой “широковещательной” технологии коммуникаций, принятой в научном сообществе, очевидно, что вероятность того, что *Иванов* когда-либо узнает, что кто-либо его цитировал, мала. Не будем обсуждать, нужна ли эта информация *Иванову*, будем считать, что она ему необходима, в противном случае дальше книгу можно не читать.

1.2. Видение проблемы

В 1963 г. Ю. Гарфилд (*Eugene Garfield*) создал инструмент анализа научной информации − “Указатель научного цитирования” (*SCI, Science Citation In dex*), пригодный для практического использования. Тем самым он косвенно открыл путь как минимум трем новым областям исследований на базе цитирования: а) поиску информации на основе цитирования; б) организации знаний на основе анализа библиографических связей и совпадающих цитирований; в) оценке исследований на основе подсчетов цитирований.

Объектом исследования в этих областях является множество научных документов и их библиографических ссылок. Однако специалисты по анализу цитирования обычно фокусируют свое внимание на определенных частях этого множества, которые образуют зоны конкретных исследований или областей знаний. Не

смотря на то, что большинство специалистов ощущают потребность в качественной теории при создании умозаключений на основе таких связей, они обычно подходят к этим проблемам посредством простой арифметики определения количества цитирований. Качественный анализ данных цитирования следует строить на теоретической системе взглядов, которая объясняет, почему авторы цитируют именно таким образом, как они это делают. Однако сегодня приходится констатировать, что несмотря на нескончаемые дебаты и многочисленные попытки, удовлетворительной теории цитирования до сих пор не существует.

По мнению Й. Николайсена [Nicolaisen, 2003], современная библиометрия, основанная на изучении цитирования, "…подвержена влиянию методологического индивидуализма, который представляет знания, скорее, как отдельные психические состояния в противоположность знаниям как социальному процессу". Г. Смолл [Small, 1978] рассматривает ссылки в качестве понятийных символов: "Для понимания важности библиографических ссылок нам требуется исследовать когнитивные процессы, вовлеченные в генерирование связного письменного текста". Д. Чубин и С. Мойтра [Chubin, Moitra, 1975] считают цитирование частным процессом, а Б. Кронин [Cronin, 1 984] − частной практикой и внутренним явлением. Автор работы [Van Raan, 1998] высказывает мнение, что такие индуктивные подходы едва ли пригодны для создания теорий цитирования: "Бóльшая часть такого теоретизирования фиксируется на роли цитирующего автора и присущих ему (ей) библиографических ссылок. Это не очень успешный подход к объяснению последствий с точки зрения цитируемого автора". Там же далее: " Некоторые отдельные характеристики цитируемых авторов интересны, но функции распределения этих характеристик представляют собой оценку той части мира, которая наиболее важна для библиометрии".

Справка. Когнитивная наука (*Cognitive science*) – комплекс наук, изучающих познание и высшие мыслительные процессы на основе применения теоретико-информационных моделей. Включает исследования, проводимые в таких областях, как эпистемология, когнитивная психология, лингвистика, психолингвистика, психофизиология, нейробиология и компьютерная наука. Основания когнитивной науки были заложены исследованиями математика А. Тьюринга по конечным автоматам (1936). Ему удалось показать, что для проведения любого вычисления достаточно повторения элементарных операций. Тем самым, открылись перспективы для проверки и реализации известной идеи Т. Гоббса и Д. Буля, что мышление есть вычисление (http://dic.academic.ru/).

Альтернативой методологическому индивидуализму является методологический коллективизм. С этой точки зрения связующим звеном всех человеческих поступков являются системы социальных значений. Чтобы сделать возможным понимание, объяснение или прогноз человеческой деятельности, потребуется, во-первых, войти в социальный мир, в котором располагается индивидуум. Только выполнив это, можно получить возможность анализировать, интерпретировать и объяснять деятельность отдельных личностей. Б. Хйорланд [Hjørland, 1997] приводит веские аргументы в пользу методологического коллективизма. Следуя тщательному изучению теорий познания менталистов и их недостатков в отношении проблем, касающихся человеческой коммуникации, он предлагает альтернативу на основе теории деятельности. Важной особенностью теории деятельности является то, что она считает когнитивную способность адаптацией к экологическому и социальному окружениям.

Из понимания того, что свойственная человеку познавательная способность не развивается отдельно в каждом индивидууме в результате сенсорного ввода в мозг, следует, что информатике следует прекратить принимать структуры знаний индивидуумов в качестве своей исходной точки, а вместо этого сфокусироваться

на областях знаний, дисциплин или иных коллективных структур знаний. Адекватная теория цитирования должна воздержаться от рассмотрения действия цитирования с психологической точки зрения, а вместо этого признать его встроенным в социокультурные конвенции дисциплин.

Справка. Методологический индивидуализм − концепция и проблема социальной эпистемологии, обсуждавшаяся, наряду с методологическим коллективизмом, в "понимающей" социологии М. Вебера, в формальной социологии Г. Зиммеля, в работах К. Поппера, Дж. Уоткинса (автор данного термина) и с позиций аналитической философии А. Данто. Сегодня эта проблема приобретает новые смыслы и акценты. Суть вопроса состоит в следующем: являются ли социальные процессы результатом деятельности отдельных людей, или они развиваются по своим законам, увлекая людей, включенных в них? Глубинной предпосылкой данной проблемы является вечный спор между номинализмом и реализмом; при этом первый утверждает, что действительным существованием обладают единичные вещи, чувственное и интуитивное познание фиксирует их реальное бытие, а универсалии, абстракции − это только имена, существующие в языке и мышлении. Сегодня эта проблема в учениях об обществе приняла форму методологической оппозиции индивидуализма − общество состоит только из людей и их действий, − и холизма − законы социальных целостностей не сводимы к действиям индивидов [ЭЭиФН, 2009].

Л. Лаудан [Laudan, 197 7] считает, что научные дисциплины представляют собой виды социальных институтов, фундаментальная функция которых – решать проблемы и документировать свои решения в виде создания текста. Научная традиция требует, чтобы ученые при документировании решений в текстовой форме ссылались на предыдущие исследования, связанные с предметом обсуждения решений, по которым дается отчет. Считается, что эти библиографические ссылки определяют более ранних исследователей, чьи концепции, теории, методы, оборудование и т. п. вдохновляли или направляли автора в процессе решения научных проблем. Таким образом, общая теория цитирования

должна заниматься исследованиями решения научных проблем и научной документацией.

В работе [Nickles, 1981] предлагается модель решения научных проблем. Автор аргументирует, что проблема включает не только ее решение, но и все условия или ограничения на это решение, а также требования того, чтобы данное решение (объект, удовлетворяющий этим ограничениям) было найдено. Согласно этой модели научная проблема представляет собой потребность в том, чтобы была достигнута основная цель в сочетании с ограничениями на способ, которым достигается эта цель, т. е. условия адекватности решения проблемы. Данное определение противоречит распространенной ранней "логике вопросов и ответов", согласно которой вопрос определяется как набор приемлемых ответов плюс запрос на ответ, удовлетворяющий определенным условиям по количеству, определенности и полноте. Однако, по мнению Т. Никлза, определение вопроса как набора допустимых ответов на него и, аналогично, проблемы по ее набору приемлемых решений логически создает следующую простую дилемму: либо возможные решения проблемы известны, либо они неизвестны. Если решения известны, то проблемы нет как таковой, поскольку уже имеются все решения. Если они неизвестны, опять-таки проблемы нет, поскольку, как можно знать, что это такое?

Эта дилемма на самом деле представляет собой вариант парадокса, приведенного в трактате Платона "Менон": "Либо человек знает, что он ищет, или нет. Если он действительно знает, то он это уже имеет, таким образом, поиск бессмыслен. Если человек не знает, он не смог бы распознать это, даже если бы он об него случайно споткнулся; таким образом, поиск опять невозможен". Отметим, что в математическом образовании парадокс Менона формулируется следующим образом: "Человек, никогда не учив-

шийся математике, с помощью хорошо подобранных вопросов может сам открывать даже геометрические истины".

Тем не менее Т. Никлз утверждает, что данная дилемма может быть решена. Поиски возможны, потому что изложение реального вопроса или проблемы представляет собой, кроме требования, простое описание потенциальных подходов к ответу – искомому объекту. Если известно достаточно ограничений, чтобы сделать проблему ясной и хорошо определенной для исследователей, это означает, что они знают, что можно было бы считать решением, если бы им удалось наткнуться на него. Эти ограничения не только говорят им, что они нашли это решение, но и определяют область пространства проблемы, в которой следует выполнять поиск, потому что каждое ограничение вносит свой вклад в описание характеристик проблемы, помогая исключать некоторые возможные решения как недопустимые. В самом реальном смысле постановка проблемы представляет собой половину ее решения!

Автор монографии [Laudan, 1977] предполагает, что сама цель построения теорий – обеспечить решение эмпирических проблем и избежать концептуальных проблем и аномалий. Ссылаясь на различные эпизоды в истории науки, Л. Лаудан указывает на то, что построение теорий происходит в пределах границ традиционных научных исследований (междисциплинарных областях). Он заявляет, что традиция научных исследований предлагает набор рекомендаций по разработке конкретных теорий. Часть этих рекомендаций образует онтологию, которая определяет типы фундаментальных сущностей, существующих в области или областях, в пределах которых построена данная традиция научных исследований. Таким образом функция конкретных теорий – объяснять эмпирические проблемы в конкретной области, уменьшая их до онтологии традиции научных исследований. В дополнение Л. Лаудан пишет, что традиция научных исследований

указывает на определенные процедуры, которые представляют собой легитимные методы поиска, открытые для исследователей в рамках данной традиции. Таким образом, он приходит к выводу, что традиции научных исследований представляют собой наборы онтологических и методологических предписаний и запретов.

Справка. Онтология − раздел философии, изучающий бытие. Онтология в классическом понимании есть знание о предельно общем. Термин был предложен Р. Гоклениусом в 1613 г. в его "Философском словаре". В практическом употреблении термин был закреплен Х. Вольфом, явно разделившим семантику терминов "онтология" и "метафизика".

Онтология (в информатике) − это попытка всеобъемлющей и детальной формализации некоторой области знаний с помощью концептуальной схемы. Обычно такая схема состоит из структуры данных, содержащей все релевантные классы объектов, их связи и правила (теоремы, ограничения), принятые в этой области. Этот термин в информатике является производным от древнего философского понятия "онтология". [Википедия]

Социальные конструктивисты Г. Коллинз (*H. Collins*), Б. Латур (*B. Latour*) и другие, подобно И. Канту, задававшемуся вопросами о границах познания, изымают понятие внешнего мира самого по себе из модели объяснения научного знания и говорят об этом как о методологическом приеме, направленном на преодоление определенных проблем. Научное знание не индуцируется и не выводится рационально из наблюдений а, наоборот, является продуктом социального договора [Collins, 1986] и "внешний мир" появляется только по достижении консенсуса между учеными [Latour, 1987]. Однако в своих работах они заявляют об исключительно социальной природе научного знания как онтологического "открытия". Конечно, подходы к решению научных проблем социальных конструктивистов существенно отличаются от подходов Т. Никлза и Л. Лаудана. Предполагается, что такая радикальная релятивистская позиция относительно решения научных

проблем опирается на философский тезис Дюгема – Куайна. Однако не ясно, почему из этого тезиса должно следовать, что социология несет единоличную ответственность за заполнение пробела между наблюдениями и теорией.

Справка. Тезис Дюгема – Куайна (*Duhem − Quine thesis*) − точка зрения французского философа науки Пьера Дюгема (*Pierre Maurice Marie Duhem,* 1861−1916) и американского логика Уилларда Куайна (*Willard Van Orman Quine*, 1908−2000), согласно которой наука состоит из сложной сети суждений, понятий, гипотез и теорий, оцениваемых "в целом" без возможности выделить индивидуальные суждения из всей системы убеждений [Википедия].

Л. Лаудан [Laudan, 1996] говорит о том, что, вероятно, может быть дан социологический отчет о том, почему ученые принимают что-либо за истину, а иное − за ложь. Он указывает, что для достижения такого заключения потребуется интерпретация тезиса Дюгема – Куайна. Приверженцы социального конструктивизма часто ссылаются на разнообразие эмпирических исследований как свидетельство своих радикальных утверждений. Однако доказательный вес этих исследований оспаривается. Так, Дж. Коул [Cole, 1992] тщательно исследовал ряд таковых и пришел к заключению, что их создатели не смогли создать ни одного примера или исследования конкретного случая, который демонстрирует, что социальные процессы действительно влияют на конкретные результаты науки. Его вывод оставался фактически не оспоренным в течение длительного времени. С. Шапин [Shapin, 1995] в своем обзоре и защите социального конструктивизма цитирует возражение Дж. Коула, но не делает попытки подвергнуть сомнению его достоверность. В согласии с Дж. Коулом следует признать, что социальным конструктивистам неоднократно удавалось демонстрировать тот факт, что процесс "производства" науки имеет место в социальной среде, однако следует признать

и то, что им не удалось продемонстрировать, как конкретные результаты этого процесса зависят от социальных процессов.

Научные тексты являются предлагаемыми решениями эмпирических и концептуальных проблем. По существу, они создаются не в вакууме, а являются продуктами традиций научных исследований. При документировании своих исследований ученые применяют ссылки, которые, как подразумевается, не используются без достаточных оснований. Поступая так, они ссылаются на предыдущие работы для установления связи с рассматриваемым исследованием. Поскольку к заданной проблеме можно подходить с точки зрения различных теоретических направлений и с использованием различных методов, можно надеяться, что набор цитат будет релевантен выбранному направлению и методу исследования. Как правило, отчеты о результатах передовых исследований (ПИ) предоставляют ограниченный набор ссылок на документы, которые выбираются для включения в перечни библиографических ссылок и научные отчеты. Таким образом, ПИ отражают ситуацию в отношении теории, направленной на понимание, объяснение и предсказание распределения цитат в сети научных документов.

Следуя Т. Гирину [Giery n, 1978] и Г. Цукерман [Zuckerman, 1987], область исследований складывается из ряда связанных, хотя и отдельных проблем, а ряд связанных областей проблем составляет отрасль науки (специальность). Научные дисциплины покрывают набор связанных специальностей, а программа научных исследований в любой дисциплине в любой момент времени представляет собой подмножество теорий и проблем, над которыми можно, с большими или меньшими трудностями, работать. Из этого следует, что ПИ определяют, какие проблемы будут выбираться, из каких направлений, посредством каких методов. Принимая во внимание вывод предыдущего абзаца, получаем,

что распределение цитирований в любой сети цитирований отражает исследовательскую активность по различным проблемам, теориям и методам.

Влияние социальных факторов почти не вызывает сомнений. Например, работа [Merton, 1970] демонстрирует влияние ряда экономических и военных вопросов, с которыми столкнулось их общество, на выбор проблем для исследования британскими учеными XVII в. Социальные факторы также "управляют" учеными при формулировке гипотез научных исследований. Особенно это касается ограниченных по времени научно-исследовательских проектов, например работы над диссертацией.

1.3. Цитирование как психологический процесс

Историки науки датируют рождение традиций библиографических ссылок чрезвычайно широко, начиная с XII в., и всякий раз для этого находят хорошую аргументацию. Однако существует мнение, что до XVI в. авторы часто дублировали работы своих предшественников без надлежащего признания. На этом основании конец XVI в. может быть принят в качестве начальной точки истории современной библиографической ссылки. На этот счет существуют и другие точки зрения. Так, в работе [Прайс, 1966] говорится: "По подшивкам многих научных журналов можно видеть, что примерно в 1850 г. возникает традиция открыто ссылаться на работы предшественников…", а И. В. Маршакова-Шайкевич в определении "Цитирование в науке" [ЭЭиФН, 2009] сообщает: "Цитирование становится стандартной этической нормой в науке примерно в середине 20-го столетия, когда научный журнал начинает рассматриваться как социальный институт и неотъемлемый инструмент общения между учеными".

Следует считать, что акт цитирования важен для научной документации, поскольку все авторы, практически без исключения,

действуют как представители традиций научного документирования, встроенных в социокультурные и эпистемологические правила научных дисциплин. В середине 1970-х гг. социальные конструктивисты (см., например, [Woolgar, 1976], [Latour, 1987]) выдвинули тезис "процесса убеждения", который привел ряд аналитиков к приравниванию цитирования к процессу убеждения. Опираясь на него, автор работы [Gilbert, 1977] полагает, что "ученому воздается должное путем признания за создаваемые результаты, которые рассматриваются как новые, важные и истинные". Далее он замечает: "Не все важные статьи, которые могли бы быть процитированы, имеют равное значение для обеспечения такой поддержки" – и делает вывод: "…следовательно, авторы статей будут стремиться цитировать "важные и корректные" статьи, могут цитировать "ошибочные" статьи, чтобы бросить им вызов, и будут избегать цитирования "тривиальных" и "не относящихся к делу" статей". Нельзя не отметить, что существует критика радикального заявления Дж. Гильберта [Gilbert, 1977] о том, что процесс убеждения является основной мотивацией цитирования, например, в работе [Zuckerman, 1987]. Иными словами, тезис процесса убеждения нуждается в аналитической поддержке.

Й. Николайсен [Nicolaisen, 2002] считает, что если взаимосвязь между частотой цитирования и качеством научного исследования действительно существует, то не похоже, что эта взаимосвязь является линейной. Вероятно, будут цитироваться чаще весьма плохие научные исследования по сравнению с посредственными работами. Это заявление Й. Николайсен сделал на основе анализа коллекции рецензий на монографии, публикуемые в социологическом журнале, и количества цитирований, которые монография получила в пятилетний период после данного критического анализа.

Для понимания поведения людей, занимающихся написанием научных документов, весьма полезно ознакомиться с базовыми положениями трех направлений (школ) в области научной психологии, которые имеют релевантные результаты исследований:

− бихевиористская школа рассматривает окружающие условия как причину того или иного поведения человека;

− гештальтская школа исследует, как мы экстраполируем получаемую (входящую) информацию для создания завершенных мысленных образов;

− когнитивная психология изучает, как человеческое мышление контролирует поведение.

Все, что относится к психоаналитической школе (психоанализу), в нашей работе не рассматривается.

Бихевиоризм (*Behavior* − поведение) − буквально − наука о поведении. Основателем данного направления в психологии является американский психолог Д. Уотсон (*John Broadus Watson*; 1878−1958). Его кредо выражалось в том, что предметом психологии является поведение, а не сознание. Важнейшими категориями бихевиоризма являются: стимул (любое воздействие на организм со стороны среды, в том числе и данная, наличная ситуация), реакция и подкрепление (в качестве этой категории для человека может выступать реакция окружающих людей).

Гештальтская школа (ГШ) психологии базируется на утверждении о существовании свойства, называемого “гештальтом”, означающего различие между звуками и музыкой или цветом и живописью, и это определенное свойство такое же реальное, как тон и цвет. Основание ГШ приписывают трем немецким психологам: М. Вертеймеру (*Max We rtheimer*, 1880 −1943), сформулировавшему принципы “гештальтской психологии” в 1912 г.; В. Кёлеру (*Wolfgang Kohler*, 1887−1967) и К. Коффке (*Kurt Koffka*, 1886−1941). В 1910 г. М. Вертеймер поставил эксперимент,

убедивший наблюдателей в том, что они видели движение, которого на самом деле не было. С помощью стробоскопа он доказал, что человеческое восприятие может отличаться от действительности. Психологи ГШ полагали, что три понятия: “ ощущение”, “обучение” и “внимание” − недостаточны для описания воспринимаемого нами окружающего мира. Они объясняли это экстраполяцией, основанной на собственных ощущениях. Если для нас нарисована какая-то незаконченная картина, мы мысленно пытаемся завершить ее, представляя и (или) предполагая “остальное”, и делаем это бессознательно.

Теоретики ГШ не соглашались с бихевиоризмом, поскольку рассматривали сознание как нечто важное (бихевиористы почти не придавали ему значения). Сторонники ГШ допускали, что мозг “незамедлительно, спонтанно и неизбежно” организовывает все, что мы воспринимаем. Мы осуществляем это через психические процессы, зачастую добавляющие к реальной информации воображаемую, чтобы сделать собственную картину яснее. Более того, мы часто допускаем ошибки в этом процессе “додумывания”. Приведем две основные идеи теории ГШ, применяемые в современных исследованиях социальной деятельности:
а) существует “целое” с реальными качествами помимо его составляющих, другими словами, целое больше, чем сумма его частей; б) элементы, создающие целое, взаимозависимы и приобретают значимость посредством своей роли в целом.

Когнитивная психология (КП) изучает познавательные процессы человеческого сознания. Исследования в этой области обычно связаны с вопросами памяти, внимания, чувств, представления информации, логического мышления, воображения, способности к принятию решений. В информационных единицах когнитивные способности человека не очень велики и, по экспериментальным данным работы [Лившиц, 1 976], составляют

120 бит/чел. час. Например, самообучение − это лишь одна из многих частей когнитивных способностей человека. Многие положения КП возникли и формировались под влиянием информационного подхода, на аналогии между преобразованием информации в вычислительном устройстве и осуществлением познавательных процессов у человека. Психолог Р. Аткинсон (*Richard C. Atkinson*, родился в 1929 г.) в своих исследованиях ориентировался на "компьютерную метафору", проводящую параллель между познавательными процессами человека и преобразованием информации в вычислительном устройстве.

Ряд исследователей, по всей видимости, разделяет мнение, что цитирование лучше всего понимать как психологический процесс. В качестве примера обратимся к работе о манере цитирования − теории С. Хартера [Harter, 1992] , который считает, что акт цитирования является динамическим, комплексным и когнитивным действием: "Релевантные библиографические ссылки, обнаруживаемые исследователем, вызывают когнитивное изменение. По мере продвижения исследования библиографические ссылки (и обнаруживаемые в них знания) вносят свое влияние на концептуальную основу данной работы, выбор проблем и методов, а также на интерпретацию результатов. Наконец, когда исследование завершено, релевантные библиографические ссылки, или те ссылки, которые привели к особенно важным источникам, будут внесены в список библиографических ссылок в конце опубликованной работы, которая представляет результаты этого исследования. Автор, который включает конкретные ссылки в свой перечень библиографических ссылок, извещает читателей об исторической важности этих ссылок для данного исследования; каждая ссылка определяется автором как важная в некоторой точке исследования или процесса написания работы". Результаты исследования манеры цитирования, выполненные в работе [White,

Wang, 1997], в целом подтвердили теорию С. Хартера, однако было отмечено, что учет важности, на котором базируется теория, оказывается проблематичным условием.

Психолог У. Найссер (*Ulric Neisser*; 1928−2012) в своей монографии [Найссер, 1981] сравнивает действия, выполняемые компьютером, с познавательными процессами человека и высказывает предположение о том, что "их можно исследовать и даже, может быть, понять". В этом случае человек рассматривается как система, имеющая устройства ввода, хранения, вывода информации с учетом ее пропускной способности. У. Найссер склонялся к мнению, что КП не в состоянии эффективно решать повседневные проблемы и объяснять особенности человеческого поведения. Ответственность за такую ситуацию он возлагает на практически полную ориентацию исследований на лабораторные методы эксперимента, предполагающие низкую валидность получаемых результатов. У. Найссер отмечает, что у КП не много шансов реализовать свой потенциал без использования результатов работ Дж. Гибсона (*James Jerome Gibson,* 1904−1979) по восприятию информации высокоуровневыми процессами, происходящими в головном мозге.

1.4. Нормативная теория цитирования*

Автор монографии [Ravetz, 1973] высказывает мнение, что цитирование регулируется профессиональной этикой, основанной на неформальном и естественном профессиональном знании. Согласно работе [Kaplan, 1965] основной функцией практики подстрочных сносок было повторное подтверждение базовых норм научной манеры поведения. Нормативная теория цитирования основывается на предположении о том, что наука является нормативным институтом, управляемым внутренними наградами и санкциями. Она полагает, что ученые обмениваются информацией

* В параграфах 1.4.–1.7. представлен обзор работы [Nicolaisen, 2007].

(в форме публикаций) для признания (в форме наград и цитирования). Эта точка зрения предполагает, что цитирование является способом признания интеллектуального долга и, таким образом, подвержено значительному влиянию со стороны осознанной ценности, а также когнитивного, методологического или локального контекста цитируемых статей [Baldi, 1998].

В первых работах по социологии науки предполагалось, что консенсус в науке определялся специфическим научным этосом, т. е. набором правил поведения в науке, которые не имеют статута юридических законов. Их действенность, по Р. Мертону, связана с ориентацией членов научного сообщества на определенный комплекс ценностей и норм, который характерен для этого социального института. Нормы выражаются в форме позволений, запрещений, предписаний, предпочтений и т. п. Эта точка зрения в наиболее краткой и влиятельной формулировке была представлена Р. Мертоном в его работе [Merton,1973], где он сумел выявить и сформулировать четыре этических принципа поведения в науке, признанных классическими. Эта четверка получила аббревиатуру *CUDOS*: *C – Communalism* (коллективизм: научные результаты должны стать достоянием всего общества, быть доступными для всех); *U – Universalism* (универсализм – норма, требующая, чтобы оценка научного результата всецело основывалась на внеперсональном уровне); *D – Disinterestedness* (бескорыстность: на результаты исследования не должны влиять ненаучные интересы религиозного, политического, экономического или иного характера); *OS – Organised Skepticism* (организованный скептицизм: исследователи обязаны быть критичными как по отношению к работе других, равно как и к собственной работе). В работе [Мирская, 2005] содержится толкование принципов *CUDOS*.

Императив “коллективизм” имеет явно директивный характер, он предписывает ученому незамедлительно передавать плоды

своих трудов в общее пользование, т. е. сообщать о своих открытиях другим ученым тотчас после проверки, свободно и без предпочтений. Научные открытия являются продуктом социального сотрудничества и принадлежат сообществу. Они образуют общее достояние, в котором доля индивидуального "производителя" весьма ограничена. Права собственности в науке фактически не существует.

Универсализм порождается внеличностным характером научного знания. Поскольку утверждения науки относятся к объективно существующим явлениям и взаимосвязям, то они универсальны и в том смысле, что они справедливы везде, где имеются аналогичные условия, и в том смысле, что их истинность не зависит от того, кем они высказаны. Ценность научного вклада не зависит от национальности, классовой принадлежности или личных качеств ученого.

Бескорыстность – стремление ученых к приоритету – создает в науке своего рода конкурентные условия. Такая ситуация может толкать на какие-то особые действия, предпринимаемые специально, чтобы затмить соперников. Эти действия способны исказить нормальный ход исследования и, соответственно, его результаты. В качестве "противоядия" указанным побуждениям выдвигается требование бескорыстной деятельности. Эта норма предписывает ученому строить свою деятельность так, как если бы кроме постижения истины у него не было никаких других интересов. Р. Мертон излагал требование бескорыстности как предостережение от поступков, совершаемых ради достижения более быстрого или более широкого профессионального признания внутри науки. В трактовке американского социолога Б. Барбера эта норма направлена на осуждение ученых, использующих исследования как способ достижения финансового успеха или приобретения престижа вне профессионального сообщества.

В общем, императив бескорыстности (это ориентационная норма) в наиболее широком толковании утверждает, что для ученого недопустимо приспосабливать свою профессиональную деятельность к целям личной выгоды.

Организованный скептицизм − это одновременно и методологическая, и институциональная норма. Р. Мертон рассматривал именно первый аспект – организованный скептицизм метода естественных наук, требующего по отношению к любому предмету детального объективного анализа и исключающего возможность некритического приятия. Для науки нет ничего “святого”, огражденного от критического анализа. В то же время норма организованного скептицизма является и директивным требованием по отношению к ученым. Поскольку работа каждого ученого-естественника строится на результатах предшествующих исследований, умышленное или неумышленное отступление от истины является преступным по отношению к развитию науки. Отсюда следует, что никакой вклад в знание не может быть допущен без тщательной, всесторонней проверки. Норма скептицизма предписывает ученому подвергать сомнению как свои, так и чужие открытия и выступать с публичной критикой любой работы, если он обнаружил ее ошибочность. Требование публичной критики любой замеченной ошибки создает уверенность в надежности и правильности тех работ, включение которых в архив науки не сопровождалось критической реакцией. Императив организованного скептицизма создает атмосферу ответственности, институционально подкрепляет профессиональную честность ученых, предписываемую им нормой бескорыстия.

Императивы *CUDOS*, передаваемые наставлением и примером и подкрепленные санкциями, составляют исторически сложившийся “этос науки”, являющийся образцом профессионального поведения ученого, который должен делать то, что полезно для

науки. Безусловно, Р. Мертон останется в социологии ученым, сумевшим выявить и сформулировать “кодекс строителя научных знаний”.

Нормативная теория цитирования постулирует, что исследователи цитируют те материалы, которые подтвердили свою ценность для них, таким образом, имеет место императив “*C*”. Теория также констатирует, что ученые при оценке работ коллег ведут себя универсально; другими словами, на их решения о том, что цитировать, не влияют функционально не относящиеся к делу характеристики, такие как пол, раса, религия или ранг научного автора (императив “*U*”). Более того, теория цитирования предполагает, что ученые бескорыстны и не стараются получить персональных выгод, льстя другим или цитируя самих себя (императив “*D*”). Также она утверждает, что ученые рассматривают собственные работы с тем же скептицизмом, что и работы других (императив “*OS*”). На основе этих положений в работе [Smith, 1981] сформулировано пять основных предположений, лежащих в основе АЦ (см. 2.2). В работе [Merton, 1995] высказывается убеждение, что авторы в общем случае цитируют материалы, которые подтвердили свою ценность для них, из-за социальных механизмов контроля науки.

Следует отметить, что нормативная теория цитирования неоднократно подвергалась критике. Например, авторы работы [Cole, Cole, 1972] указывали, что “ссылка на ключевого предшественника некоей работы не приводится”. Данное упущение редко бывает результатом прямого злого умысла со стороны автора, но более часто – недосмотром или недостатком осведомленности. В работе [MacRoberts, MacRoberts, 1986] утверждается: “Ученые цитируют источники, оказывающие влияния на читателя”. В работах Т. Брукса [Brooks, 1985], [Brooks, 1986] приведены результаты следующего эксперимента. В Университете штата

Айова (*University of Iowa*) был проведен опрос группы ученых-исследователей, которые пояснили мотивы указания ссылок в своих недавно опубликованных работах, расставив по ранжиру семь основных мотивов цитирования. Результаты опроса показали, что авторы цитируют по многим причинам, при этом наименьшую важность имела мотивация цитирования возвращением долга. Обнаружено, что 70 % цитат были мотивированы несколькими причинами. Кроме того, было сделано заключение: “Мы не можем более наивно предполагать, что авторы цитируют только заслуживающие внимание выдержки в положительной манере … Авторы являются адвокатами собственных точек зрения, которые используют цитируемые документы в расчетливой попытке оправдать себя”.

1.5. Социальная конструктивистская теория цитирования

В этом разделе феномен цитирования рассматривается с точки зрения социальных конструктивистов, которые полагают, что научная закрытость является результатом некоего переговорного процесса, в котором одна сторона убеждает другую сторону. Авторы работы [Latour, Woolgar, 19 86] отстаивают точку зрения, что наука является искусством убеждения. С их точки зрения, успешные ученые – это те, кто наиболее искусно справляется с убеждением других в том, что их не просто убеждают, а что отсутствует посредничество между тем, что говорится, и истиной.

В искусстве убеждения нет запрещенных приемов. Согласно социальным конструктивистам при отчете об исследовании успешный ученый пользуется многими убеждающими ходами. Когда авторы цитируют, они располагают цитируемые документы таким образом, чтобы убедить читателей в правильности своих заявлений. Таким образом, убеждение, а не желание отдать должное там, где следует, является основной мотивацией цитирова-

ния. На эту позицию, прямо противоположную нормативной теории цитирования, большое влияние оказала статья [Gilbert, 1977], в которой утверждается, что уважаемые работы могут цитироваться с целью “придать сияния” отчету, даже если они кажутся не очень связанными с реальным содержанием данного отчета.

Б. Латур [Latour, 1987], размышляя в этом направлении, говорит: “Чтобы воздвигнуть убедительный фасад, авторы научных документов привлекают крючкотворство. Во-первых, многие ссылки могут быть неправильными или неправильно процитированными; во-вторых, многие статьи, на которые они ссылаются, могут не иметь какого-либо отношения к сути документа и могут быть там только для показухи”. Б. Латур не считал такие действия нелогичными, наоборот, он отстаивал мнение, что если бы читатели смогли выяснить, что на самом деле происходит (например, цитаты используются только для “показухи”) результат оказался бы пагубным для таких авторов.

Г. Уайт [White, 2004] представил тщательный анализ гипотезы убеждения, доказывая, что она состоит из двух частей. Первая часть должна иметь дело с тем, чтó говорят цитирующие авторы о цитируемых работах или, более точно, с контекстом, в рамках которого они их обсуждают. Он называет эту часть “убеждением с помощью фальсификации” и считает, что “цитирующие авторы часто искажают работы, на которые они ссылаются, подгоняют их значения под свои нужды”. Вторая часть гипотезы убеждения должна иметь дело с выбором самих цитируемых работ вне зависимости от того, что о них говорится. Г. Уайт называет эту часть “убеждением с помощью упоминания громких имен вскользь” и отмечает, что она в большей или меньшей степени не зависит от контекста: “Цитирующие авторы несоразмерно цитируют работы признанных авторитетов, чтобы добиться правдоподобия по ассоциации”.

Исследователи обычно проверяли либо первую, либо вторую часть гипотезы убеждения. Согласно первой части гипотезы убеждения (убеждение с помощью фальсификации) цитирующие авторы часто искажают работы, на которые они ссылаются, подгоняя их мнение под свои нужды (см. высказывание Б. Латура). Таким образом, первая часть гипотезы убеждения представляет собой отрицание нормативного предположения: “ Цитируемый документ связан по контексту с цитирующим документом” [Smith, 1981]. Нормативная теория подразумевает, что ссылки непосредственно указывают на семантические взаимосвязи между цитирующими и цитируемыми работами.

Например, автор работы [Garfield, 1979] предположил, что цитата есть точное, однозначное представление предмета. Однако в серии проверок этого предположения на реальных массивах данных были получены противоречивые результаты, см. [Harter, Nisonger, Weng, 1993]. Таким образом, невозможно сделать однозначное заключение о том, до каких пределов цитируемые и цитирующие документы связаны семантически. Б. Кронин [Cronin, 1994] отметил, что тексты могут цитироваться на разных уровнях глубины детализации или агрегации. Этот факт влияет на множества подобий и объясняет, почему степень сходства предмета между парами цитируемых и цитирующих документов часто оказывается различной.

Согласно второй части гипотезы убеждения (убеждение с помощью упоминания громких имен вскользь), авторы несоразмерно цитируют работы признанных авторитетов, чтобы добиться правдоподобия по ассоциации. Авторы работы [M acRoberts, MacRoberts, 1996] утверждают: “Основной целью автора является не цитирование своих источников влияния, а представление настолько авторитетного аргумента, насколько это возможно”. Однако, существуют исследования, в которых высказываются

сомнения в действительности данного заявления (см., например, [Moed, Garfield, 200 3]). Авторов этой работы заинтересовал вопрос: “ Как меняется относительная частота, с которой авторы цитируют “авторитетные” документы в перечнях библиографических ссылок в своих статьях, в зависимости от числа ссылок, которые эти статьи содержат?” Было обосновано, что “если эта пропорция уменьшается по мере того, как перечни библиографических ссылок становятся короче, можно сделать вывод, что цитирование документов авторитетных авторов менее важно, чем другие типы цитат, и не является основным мотивом цитирования”.

1.6. Концепция гандикапа

Й. Николайсен в своей диссертации [Nicolaisen, 2004] выполнил теоретическое обоснование манеры цитирования, которое основано на концепции гандикапа − нетривиальной версии эволюционной теории. Он сравнил библиографические ссылки с “сигналами угрозы” (*Threat signals*). Подобную аналогию предлагает Б. Латур [Latour, 1987] в книге “Наука в действии”. Оба автора разделяют идею, что библиографические ссылки являются сигналами угрозы, они не согласны с тем, как авторы могут их использовать.

Б. Латур говорит: “ Автор может превратить факт в домысел или домысел в факт, просто добавляя или удаляя библиографические ссылки”. Однако, чтобы выполнить такое преобразование, автор должен знать и применять правильные стратегии. Б. Латур исследовал две такие стратегии: накопления (*Stacking*) и модализирования (*Modalizing*). Он защищает точку зрения, что наличие или отсутствие библиографических ссылок в научном тексте означает, является ли текст “серьезным и крепким”. Чтобы создать такой текст, автору следует процитировать ряд документов, эту практику Б. Латур называет “накоплением значительного количе-

ства библиографических ссылок". Накопление может оказаться эффективным средством преобразования домысла в факт, поскольку оно требует, чтобы потенциальный читатель читал или был ознакомлен с цитируемыми документами, чтобы он мог определить силу и точность цитирующего текста, а эта задача трудновыполнима, если цитируется большое количество документов. Однако Б. Латур замечает, что накопление большого количества библиографических ссылок не является достаточным условием для того, чтобы выглядеть серьезно и прочно, если автору противостоит "сильный оппонент". Такой оппонент мог бы просто отследить все библиографические ссылки и проверить степень их привязанности к аргументу автора: " Если читатель достаточно сильный, то результат может быть пагубным для автора". Поэтому требуется еще один ход для выполнения этого преобразования. Автор должен изменить статус цитированных документов (выполнить модальное преобразование), т. е. модифицировать библиографическую ссылку, чтобы подогнать ее как можно ближе к своим собственным аргументам.

Большинство журнальных статей сталкивается с двумя потенциальными группами читателей: теми, кто читает их до публикации, и теми, кто читает после публикации. Потенциальные читатели до публикации включают, кроме прочих, редакторов и рецензентов, участвующих в типичной экспертной оценке академических журналов до публикации. Среди потенциальных читателей после публикации – другие, работающие в этой области и, возможно, более важные авторы, которые цитируются в данной статье. Как и редакторы и рецензенты, большое количество читателей после публикации обладает экспертными знаниями в данной области и ее литературе. Авторы, которые подгоняют предшествующую литературу под собственные цели и таким образом совершают умышленные поступки мошенничества, рискуют

быть разоблаченными своими потенциальными читателями как мошенники, которыми они и являются. Честным авторам, которые надлежащим образом цитируют свои источники информации и вдохновения, не следует бояться подобного разоблачения. Оба типа авторов (честные и другие) могут апеллировать к одному и тому же набору источников в качестве поддержки своих аргументов, но с разной потенциальной стоимостью, которая более высока для мошенничающих авторов, чем для честных.

Таким образом, возникает ситуация, отвечающая концепции гандикапа, согласно которой, дорогая цена важна для эволюции честности. Честные сигналы используются, потому что они принимают форму, которая требует на свое производство значительных затрат, − условие, которое могло бы привести к неэффективной коммуникации, если отправитель не смог бы понести такие расходы.

Справка. Концепция гандикапа. В 1975 г. израильский зоолог А. Захави (*Amotz Zaha vi*) сформулировал концепцию гандикапа (*handicap*), согласно которой информацию о качестве генома самца могут нести лишь вредные для выживаемости признаки. Так, размер хвоста павлина является мерой качества его генома, поскольку с длинным хвостом сложно улетать от хищников и только очень высоко приспособленный самец (с хорошими генами) сможет с большим хвостом дожить до момента размножения. Точно так же яркая окраска оперения и громкие песни самцов птиц делают их более заметными для хищников [Википедия].

Согласно работе [Nicolaisen, 2004] цитируемые документы цитирующего текста представляют собой символ уверенности. Набор библиографических ссылок представляет собой гандикап, который может себе позволить только честный автор. Он сравнивает “модализованные” ссылки с голосом “блефующего соперника” и справедливо отмечает: “ Опытный соперник немедленно обнаружит фальшивую ноту и будет знать, куда направить удар”.

Потенциальная цена такого хода может заставить автора пересмотреть свое мошенническое поведение и склонить его отказаться от риска потери репутации. Однако в соответствии с теорией А. Захави наивно предполагать, что все библиографические ссылки были честными, поскольку имеется достаточно случаев мошенничества и обмана в науке, чтобы опровергнуть такое предположение. Тем не менее, можно сделать предположение о том, что принцип гандикапа гарантирует, что цитирующие авторы честно отдают должное своим вдохновителям и источникам до некоторой удовлетворительной степени, достаточной, чтобы предохранить систему научной коммуникации от коллапса.

1.7. Рефлексивная теория цитирования

В данной теории ключевую роль играют символьные характеристики цитат. Б. Ван дер Веер Мартинс [Van der Veer Martins, 2001] считает: “ Современным “Святым граалем” библиометрии должна быть, скорее, разработка теорий индикаторов, а не разработка теории манер цитирования”. Теории индикаторов связаны с символьными характеристиками цитирования и его индикативными возможностями. Теоретики этой школы пытаются понять, как цитаты отражают и представляют науку, а не исследовать причины цитирования. Важным примером является работа Г. Смолла [Small, 1978] о весьма цитируемых документах как “концептуальных символах”. Одним из пылких защитников этого направления исследований является П. Воутерс, чье представление рефлексивной теории цитирования отражено в диссертации [Wouters, 1999b] и статьях [Wouters, 1998], [Wouters, 1999a].

П. Воутерс считает тупиковым дальнейшее развитие теории цитирования на основе попыток объяснить цитирование, связывая его с манерой цитирования. Он предлагает сосредоточить усилия на изучении символьных характеристик цитирования.

П. Воутерс рассматривает цитаты в качестве индикаторов, создающих “формализованное представление” науки, однако для интерпретации этих формализованных представлений требуется приписать значения индикаторам. Согласно П. Воутерсу атрибут значения следует основывать не на манере цитирования цитирующих ученых, а на том, как цитаты отражают характеристики науки. В своей диссертации П. Воутерс назвал это “рефлексивной теорией цитирования”.

Эта теория основывается на различной интерпретации терминов “библиографическая ссылка” и “цитирование” [Price, 1970]. Многие авторы отмечали техническое различие между ними, но П. Воутерс считает это различие фундаментальным: “Библиографическая ссылка полностью определяется цитирующим текстом, к которому она принадлежит, и цитируемым текстом, на который она указывает. С точки зрения семиотики библиографическая ссылка представляет собой символ – элементарную единицу репрезентативной системы с цитируемым текстом в качестве его объекта ссылки. … Цитирование является зеркальным отражением библиографической ссылки. С точки зрения организации библиографические ссылки не соответствуют текстам, которым они принадлежат, а соответствуют текстам, на которые они указывают, т.о. они становятся атрибутами цитируемого, а не оригинального, цитирующего текста”.

Итак, П. Воутерс рассматривает цитирование как новый сигнал, отличающийся от библиографической ссылки, на которой он строится. В отличие от библиографической ссылки цитирование не имеет “значения” и приобретает его только в руках специалиста по анализу цитирования: другими словами, местом рождения цитаты является рабочий стол специалиста по индексации документов, а не ученого. Поэтому научные исследования манеры цитирования ученых способствуют объяснению моделей библио-

графических ссылок, а не моделей цитирования: “Цитирование и библиографическая ссылка имеют разные объекты ссылок и являются на самом деле зеркальным отражением друг друга. Более того, это различие является преимуществом, так как поиск теории цитирования в библиометрии и социологии науки разделяется на две отличающиеся, аналитически независимые исследовательские проблемы: с одной стороны, модели в манере цитирования ученых, социологов и филологов в гуманитарных областях, с другой − теоретическое основание анализа цитирования” [Wouters, 1999b].

Теория П. Воутерса отражает основную идею “информационной семантики”, семейства теорий, ищущих естественно-научное и упрощенное объяснение семантических свойств мышления и языка. В основном, информационный подход объясняет обусловленное истиной содержание в терминах казуальной, номинальной или просто обычной корреляции между представлением (сигналом) и состоянием дел (ситуацией). Сигналы могут уверенно соотноситься (коррелировать) с конкретными ситуациями и, таким образом, указывать на эти ситуации.

Справка. Информационная семантика − это направление в моделировании смысла фраз на естественном языке, основанное на анализе количества переданной информации. В информационном подходе к этой проблеме случайный текст связывается с произвольной, ничего не значащей информацией − “статистическим шумом”. Значимую информацию несут закономерности в чередовании букв и слов в тексте. При отсутствии априорной информации, пожалуй, единственный способ идентификации этих закономерностей состоит в регистрации повторений фрагментов текста. Эта задача может быть решена (с заданной точностью) с помощью методов проверки статистических гипотез (http://www.wikiznanie.ru/).

В работе [Wouters, 19 99a] автор приводит два определения термина “информация”. Первое принадлежит американскому инженеру и математику К. Шеннону (*Claude Elwood Sh annon*; 1916−2001), работы которого являются синтезом математических

идей и конкретного анализа чрезвычайно сложных проблем их технической реализации. К. Шеннон обобщил идеи американского ученого-электронщика Р. Хартли (*Ralph Vinto n Lyon Hartley* , 1888−1970) и ввёл понятие информации, содержащейся в передаваемых сообщениях. В 1928 г. Р. Хартли предложил в качестве меры информации передаваемого сообщения *M* использовать логарифмическую функцию $H=\log |M|$, которая называется *хартлиевским количеством информации*. С этой точки зрения информация является исчислимой сущностью. Объем информации, заключенной в определенном сообщении, равен среднему количеству знаков, необходимых для ее кодирования [Шеннон, 1963] . Эта концепция информации не имеет отношения к ее смыслу. В 1979 г. философ Г. Бейтсон (*Gregory Bateson*) предложил диаметрально противоположную концепцию [Bateson, 1980], которая ставит в центр смысл − то, что отбросила теория К. Шеннона. Его определение информации: "Любое отличие, которое создает отличие". Сигнал ("известие об отличии") не является информацией до тех пор, пока он не означает что-либо или делает что-либо ("создает отличие").

Далее, П. Воутерс предлагает определение науки как информационного цикла, включающего цикл экспертной оценки и цикл цитирования. Получение научных знаний можно представить в виде циклического процесса, в котором определенные затраты, такие как деньги и рабочая сила, преобразуются в определенную продукцию, такую как научные статьи и формулы знания. Различные этапы этого процесса были тщательно проанализированы в научных исследованиях на микро-уровне [Knorr-Cetina, 1981] , [Latour, Woolgar, 1986] и макро-уровне [Jasanoff, 1990], в результате была отмечена важная особенность − постоянная оценка научных знаний. Эта оценка используется для задания новых целей и написания планов новых исследовательских работ. Экспертная

оценка в различных формах является центральной для данного цикла, по этой причине он называется циклом экспертной оценки.

Сам цикл экспертной оценки является продуктом довольно сложным и запутанным (рис. 1.7.1). Его можно представить в виде сети с узлами: задание приоритетов, предложение, экспертная оценка, исследование, представление в журнал, публикации, библиографические ссылки − и многочисленными связями между узлами включая циклические. Цикл экспертной оценки не следует рассматривать как часть науки, поскольку на него влияют соображения политики в области научных исследований [Cozzens, et al., 1990].

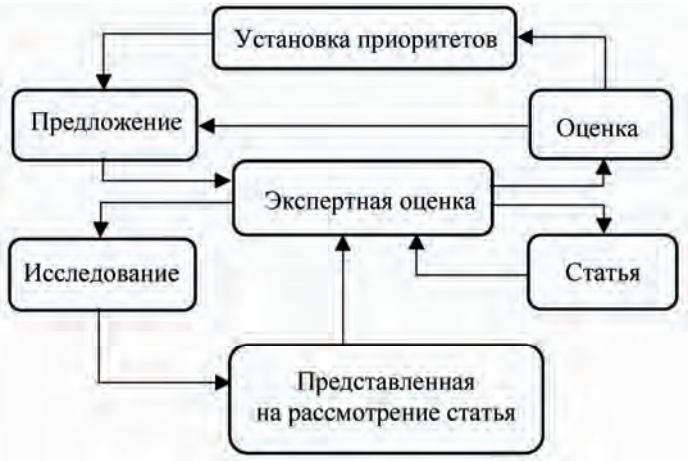

Рис. 1.7.1. Цикл экспертной оценки

Пояснение к рис. 1.7.1: оценка (*Evaluation*), установка приоритетов (*Priority setting*), экспертная оценка (*Peer review*), предложение (*Proposal*), исследование (*Research*), статья (*Article*), представленная на рассмотрение статья (*Submitted article*).

Описание данного цикла, как правило, начинается с написания предложения на научное исследование, следующим шагом является оценка предложения привлекаемыми экспертами, использующими научные критерии. Затем выполняется исследование,

которое, вероятно, ведет к представлению статьи. Вторая форма экспертной оценки, организуемой редактором журнала, при положительном решении приводит к публикации. Третьей формой экспертной оценки является серия оценок от исследовательских групп, научных институтов и университетов, наконец, оценка национального вклада в отрасль науки в целом. Эти оценки также основываются на экспертных оценках. Результаты разнообразных оценок в этой серии вносят свой вклад в расстановку приоритетов на различных уровнях, приводя к новым предложениям на исследования, организационным трансформациям или приоритетным программам. Согласно анализу Р. Уитли [Whitley, 1984], совокупность всех экспертных оценок является частью системы контроля репутации. Доминирующую роль в этом цикле играет "экспертная оценка", которая отражает центральную позицию научной экспертизы.

С появлением библиометрических индикаторов доминирующая роль экспертизы теряет свои позиции. Научная публикация может измеряться с помощью анализа цитирования или позиционироваться с помощью анализа совпадающих слов (*Co-word analysis*). Следовательно, специалист в данной области больше не является единственным источником экспертной оценки. Что касается представления информационного цикла, библиометрические индикаторы появляются в форме нового, добавленного цикла. Этот цикл обрабатывает информацию о первичном информационном цикле, т. е. о цикле экспертной оценки. Очевидно, что этот вторичный цикл (цикл цитирования) создает и преобразовывает информацию об информации, т. е. метаинформацию. Следовательно, этот вторичный цикл также является информационным циклом (рис. 1.7.2).

Первый и опорный этап цикла цитирования – семиотическая инверсия библиографической ссылки в цитирование. Индикаторы

цитирования представляют собой фундамент, на котором строятся карты науки. На основе индикаторов цитирования строятся индикаторы совпадающих цитирований (коцитирований) для карты науки. В результате другого процесса трансляции статья "переводится" в набор ключевых слов (ключевые слова, определяемые индексатором, избранные слова из названия или выборки из полного текста). Затем эти слова используются для создания индикаторов совпадающих слов. Для построения карт науки индикаторы коцитирования и совпадающих слов могут использоваться совместно.

Рис. 1.7.2. Цикл цитирования

Если бы анализ цитирования был цифровым отражением заключений экспертной оценки, ничего нового бы не произошло. Однако перевод ссылок на используемую литературу в цитирование создает дополнительную степень свободы в обращении с цитированиями. Это означает, что функционально эквивалентные индикаторы могут быть получены несколькими способами, из чего не следует, что данные индикаторы должны иметь одинаковое значение.

Поскольку цитирование играет ключевую роль во втором цикле, этот цикл можно назвать "циклом цитирования". Сделаем некоторые пояснения. Д. Прайс (*Derek de Solla Price* , 1922–19 83) был первым, кто предложил эту концепцию [Price, 1979]. Предлагаемый П. Воутерсом цикл представляет собой, в отличие от предложенного Д. Прайсом, динамический цикл, в котором

обрабатывается информация о создании знаний. Результатом исполнения этого цикла является информация о науке в форме информации о производительности ученых и других действующих лиц, представленной в виде индикаторов, карт науки, динамических значений рейтингов научной периодики и т. п.

Анализируя информационные потоки в циклах экспертной оценки и цитирования, П. Воутерс делает вывод, что научные индикаторы создают “формализованное представление” науки, которое изначально игнорирует значение. Конечно, для интерпретации этих представлений снова потребуется атрибутивное значение. Ключевым моментом является то, что данное приписывание значения может быть отложено, что позволяет манипулировать “лишенными значения” символами, такими как цитирование. Цитирование знака является объектом, таким же, как шенноновская информация: безразмерным, лишенным значения, исчисляемым. Поэтому формализованное представление науки использует не бейтсоновскую, а шенноновскую концепцию информации. Если данная цепь рассуждений верна, то наиболее важным аспектом библиометрии может быть ее численный характер. Это действительно и для индикаторов науки. Однако заметим, что не все библиометрические результаты имеют численную природу. В частности, карты науки являют собой геометрические объекты, хотя в основе их построения лежат вычисления.

Необходимо отметить, что рефлексивная теория цитирования П. Воутерса страдает от той же неразрешимой проблемы, что и информационная семантика, поскольку она не может работать с ложноположительными результатами, т. е. ссылками, которые на самом деле указывают не на то, на что следует указывать.

Пожалуй, самыми примечательными являются попытки продемонстрировать, что цитирования представляют собой индикаторы качества. Данный подход постулирует, например, что число

цитирований и качество научных исследований находятся в зависимости друг от друга, например, линейной. Этот факт подтверждает эмпирическое исследование Р. Борнштейна [Bornstein, 1991] о *J*-образной взаимосвязи между качеством научного исследования и подсчетами цитирований. Однако, по мнению автора работы [Nicolaisen, 2002], многие работы узко сфокусированы на крайних противоположных пределах распределениях цитирований и не подтверждают гипотезу. Таким образом, можно сделать заключение, что цитирования не являются индикаторами качества. Но на что же тогда они указывают?

Ю. Гарфилд отвечает на этот вопрос в своей книге[Garfield, 1979]: "Цитирования не указывают четкости, важности, качества или значимости. Скорее, они являются индикаторами полезности и влияния. … Весьма цитируемой работой является та, которая считается полезной для относительно большого количества людей или в относительно большом количестве экспериментов. … Подсчет цитирований конкретной научной работы не обязательно говорит что-либо об ее относительной важности для развития науки или общества". Итак, согласно Ю. Гарфилду цитирование указывает на то, что ссылка на цитируемую работу была сделана и использована цитирующей работой – ничего больше и ничего меньше.

1.8. Теория индикаторов

В заключительной части статьи [Wouters, 1999a] автор говорит о теории индикаторов (как суммы теорий), под которой он понимает работы, которые пытаются придать цитированиям определенный смысл. Эти теории можно условно разделить на четыре группы: группа "наука о науке", " социологическая" группа, "семиотическая" группа и группа "информационной науки".

1.8.1. *Группа "науки о науке"*. Ее основателем является Д. Прайс, выдвинувший идею рефлексивного использования

науки для измерения ее самой. По его мнению, "можно возлагать большие надежды на объективное толкование структуры фронта научных исследований, автоматическое создание карт полей активности с оценкой всех их прорывов и их центральных исследователей и автоматическое регулирование посредством анализа цитирования" [Price, 1961]. Однако на основе этой идеи не было создано детальной и непротиворечивой теории цитирования.

1.8.2. "*Социологическая" группа.* Исследователи этой группы использовали частоту цитирования как меру научной продуктивности и как интересную (с точки зрения социологии) связь между авторами и публикациями. Они разделяли предпосылку о том, что библиографическая ссылка и цитирование идентичны. Это позволяло использовать интуитивный подход к значению индикатора цитирования.

Ю. Гарфилд (*Garfield*) и И. Шер (*Irving Sher*) дали обоснование "мертоновским" нормам науки [Merton, 1973]. С этой точки зрения, цитирование рассматривается как воплощение представления признания, которое обязан делать ученый. Поскольку это ведет к симметричному позиционированию цитирования, это означает, что полученное количество цитирований прямо пропорционально полученному признанию. Попытка перевести образцы цитирования в бихевиористские характеристики является наиболее общим подходом в теориях цитирования. Таким образом социологическая группа является самой большой группой в теориях цитирования.

Мертоновская парадигма создала ряд интерпретаций цитирования, каждая из которых имеет свои отличия, но все они объединены центральным понятием специфических норм и правил науки. С более общей точки зрения, мертоновские теории индикаторов пытаются объяснить цитирование, привязывая его к манере цитирования ученого. Это было сделано также посредст-

вом конкурирующих социологических парадигм. Например, цитаты интерпретировались как форма убедительного аргумента [Gilbert, 1977].

Основной причиной того, что данная точка зрения не создала требуемой всеохватывающей теории цитирования, является многообразие бихевиористских характеристик, лежащих в основе образцов цитирования. Именно поэтому невозможно единственным образом связать цитирование символа с конкретной бихевиористской характеристикой в отношении цитирования. Тем не менее, хотя эта группа исследователей получила значительное количество знаний о культуре цитирования в науке, в поисках теории цитирования это, по мнению П. Воутерса, тупик.

1.8.3. “ *Семиотическая*” *группа.* В этой группе, созданной М. Каллоном (*Michel C allon*), центральным “актантом” является не цитата, а совпадающее слово. Оно, как и цитата, создается из слова (обычно из ключевого слова или слова из названия работы) посредством удаления из него какого-либо значения. Эти совпадающие слова впоследствии используются для создания карт науки, которые отражают поля исследований и сети взаимоотношений. Межтекстовые взаимоотношения между совпадающими словами используются для отражения расстановки сил в науке. Уместно заметить, что М. Каллон является одним из авторов акторно-сетевой теории.

Справка. Акторно-сетевая теория (*Actor-Network The ory, ANT*) представляет собой подход к пониманию научной практики, основанный на тщательном наблюдении за лабораторной деятельностью; особое внимание уделяется активному участию объектов, инструментов, материалов и т. п. в конструировании научного знания. С середины 90-х гг. методы *ANT* активно используются в социологических исследованиях информационных систем и генной инженерии. Пионерами работ по созданию акторно-сетевой теории считают социологов М. Каллона (*Michel Callon*), Б. Латура (*Bruno Lat our*) и Дж. Ло (*John Law*). [Wikipedia].

1.8.4. *Группа "информационной науки"*. Основателем этой группы следует считать российского математика и философа В. В. Налимова (В. В. Налимов, 1910−1997), который рассматривал науку как самоорганизующийся процесс обработки информации [Налимов, 1966]. Десятилетие спустя Ф. Нарин [Narin, 1976] высказал обоснованную точку зрения, что отношения цитирования между журналами отражают обмен информацией между ними. Некоторые работы Ю. Гарфилда также можно отнести к данной группе, в которой цитирование связывается с процессами обмена информацией в науке. Отличительная особенность этой группы теорий заключается в том, что она представляет науку в виде информационного процесса и абстрагируется от вопросов, связанных с существом дела. Другими словами, она использует шенноновскую информационную концепцию в качестве отправной точки исследований.

Исследователи всех четырех групп создают формализованную основу теории индикаторов в некотором парадигматическом представлении науки. Из-за разницы между библиографической ссылкой и цитированием (основной постулат П. Воутерса!), легитимация анализа цитирования должна аналитически отличаться от исследования манеры цитирования в науке. Такой вывод, без сомнения, подтверждает правомерность применения анализа цитирования. Каждая существующая теория индикаторов может рассматриваться как воплощение одного возможного типа связи в пределах области всех возможных взаимоотношений.

Итак, рост разнообразия формализованных представлений науки, а также взаимосвязь между формализованными и парадигматическими представлениями науки, можно считать основным результатом появления анализа цитирования и теории индикаторов науки.

Глава 2. Мотивация и проблемы цитирования

Ничто так не портит цель, как попадание.
Приписывается Н. Фоменко

Цитирование является частью формального научного процесса, используемого для оценки качества, важности, оригинальности, глубины, а также очевидности индивидуального или коллективного труда. Перечень библиографических ссылок (далее БС), указанный в публикации, создает своеобразный контекст работы, дает первое представление о тех проблемах, которые в ней рассматриваются, является ключом к пониманию идей, заложенных в публикации. Гипотеза о том, что ссылки представляют собой символы научных концепций, и составляет, по мнению Ю. Гарфилда, теоретическую основу указателей цитирования [Гарфилд, 1982].

Научная традиция требует, чтобы ученые при документировании собственных научных исследований ссылались на более ранние труды, связанные с предметом. Несмотря на то что эта традиция иногда называется такой же старой, как и сама наука (см., например, [Price, 1963]), историки науки расходятся в вопросе происхождения БС. Одако они солидарны с тем, что авторы научных работ издавна используют БС, чтобы придать своим текстам более убедительный вес. Развитие научных связей привело к тому, что БС стали считаться необходимыми для “эффективных” и “разумных” коммуникаций по научным и техническим вопросам [Garfield, 1977], а процесс цитирования стал “второй натурой” любого ученого, пишущего научные документы.

Цитирование в науке может включать три различные смысловые компоненты, поэтому в целом оно отражает: 1) когнитивную связь между публикациями (научными работами), 2) моду, как бы моральную необходимость ссылки на работы предшественников,

3) социальный фактор. Разделить эти компоненты часто бывает трудно. Так считает И. В. Маршакова-Шайкевич (см. определение термина “Цитирование в науке” [ЭЭиФН, 2009]).

2.1. Мотивация цитирования

В данном параграфе представлены известные нам варианты ответа на вопрос: “Что движет автором при совершении акта цитирования?” Существует несколько причин цитирования авторами работ других исследователей. В некоторых случаях имеется возможность обосновать мотив цитирования, но в большинстве случаев нельзя точно объяснить намерения автора. Всегда будет существовать расхождение между тем, почему автор делает ссылку, и нашим мнением на этот счет.

Авторы работы [Moravcsik, Murugesan, 1975] создали схему классификации БС по мотивациям цитирующего автора и использовали ее для распределения по категориям 706 БС в 30 статьях в области теоретической физики высоких энергий, опубликованных в период 1968–1972 гг. В результате было выявлено, что 41 % БС несущественные, а 14 % − негативные. Авторы предприняли попытку формально определить мотивации цитирующих авторов. Для этого были определены четыре группы цитирований $m_1 - m_4$ и выдвинуто требование, что конкретное цитирование не обязательно должно принадлежать одной группе, но не к обеим категориям одной группы. Группы и соответствующие им мотивации (список “*M*”):

m_1 − концептуальное или техническое (*Conceptual or operational*);

m_2 − органичное или формальное (*Organic or perfunctory*);

m_3 − эволюционное или сопоставимое (*Evolutionary or juxtapositional*);

m_4 − одобряющее или отрицающее (*Confirmative or negational*).

Принадлежность к группе m_1 позволяет отличить идею от технического решения; m_2 – существенное от несущественного; m_3 – развитие идеи от альтернативной точки зрения; m_4 указывает на корректность работы.

Вместе с работой М. Моравчика и П. Муругесана обычно рассматривают работу [Chubin, Moitra, 1975], в которой предложена своя взаимоисключающая классификационная схема БС (рис. 2.1.1).

Рис. 2.1.1 Схема классификации цитирований (*Chubin-Moitra*)

Существенные (*Essential*) цитирования считаются основными (*Basic*), если цитируемая работа определяется как центральная, на которую опирается цитирующий автор. Если результат в цитируемой работе не является непосредственно связанным с цитирующей работой, но является важным для автора, цитирование рассматривается как вспомогательное (*Subsidiary*). Дополнительное (*Supplementary*) цитирование – это цитирование работы, содержащей независимое суждение, с которым автор соглашается. К формальным (*Perfunctory*) относятся цитирования, связанные с работой, но оставленные без комментариев. Негативные (*Negative*) цитирования не нуждаются в комментариях. Авторы распределили по категориям БС, взятые из массива статей по физике, опубликованных в период 1968–1969 гг. Результаты анализа

показали, что только 20 % БС были формальными, примерно 5 % – негативными, при этом не было ни одного полного отрицания работы (список “*C*”).

Причины цитирования могут быть серьезными или незначительными. М. Вейншток [Weinstock, 1971] идентифицировал 15 специальных функций ссылок (далее − список “*W*”), все их можно классифицировать как “серьезные”:

w_1 − выражение уважения к первооткрывателям (*Homage to pioneers*);

w_2 − отдание должного родственной работе (*Credit*);

w_3 − идентификация методологии, оборудования и т. п. (*Methodology*);

w_4 − указание на предпосылки (*Background*);

w_5 − корректировка собственной работы (*Self correcting*);

w_6 − корректировка работы другого исследователя (*Correcting*);

w_7 − критика чужой работы (*Criticizing*);

w_8 − заявление о своих правах (*Claims*);

w_9 − указание на предстоящую работу (*Alerting to forthcoming*);

w_{10} − указание на нецитируемую работу (*Leads to poorly disseminated, indexed*);

w_{11} − идентификация данных (*Classes of fact-physical constants*);

w_{12} − идентификация оригинальной работы, в которой обсуждалась идея или концепция (*The original publication with ideas or concepts*);

w_{13} − идентификация оригинальной работы, где объясняются поименованные концепции или термины (*The original publication with terms*);

w_{14} − отрицание работы или идеи других исследователей (*Disclaiming*);

w_{15} − полемика с утверждениями других авторов (*Disputing*).

Сложность выбора характерной мотивации при использовании перечня причин, составленного М. Вейнштоком, заключается в том, что на первый взгляд они выглядят как вполне хорошее объяснение намерений автора, но фактически они являются не более чем поверхностными объяснениями. В лучшем случае определительный процесс может предложить только внешнюю обоснованность. Например, конкретное цитирование в тексте может оповестить читателя о предстоящей работе (одна из причин М. Вейнштока), но читатель, не полностью знакомый с темой, к которой относится статья, не может знать, была ли цитируемая статья наиболее подходящим выбором, имея в виду тему и ориентацию цитирующей статьи. Более того, в большинстве случаев он не сможет судить о предвзятости автора в выборе ссылок. Для полного знания факторов, оказывающих влияние на авторов в практике цитирования, от читателя следует требовать обладания энциклопедическими знаниями.

В работе [Thorne, 1977] приводится (список "*T*"), содержащий 19 скрытых причин цитирования, которые могут рассматриваться как дополнение к списку "*W*":

t_1 − серийная публикация − части целого печатаются раздельно (*Serial publication*);

t_2 − многократные публикации, т. е. вариации проекта, разосланные в разные журналы (*Multiple publications*);

t_3 − "снять шляпу", т. е. выразить глубокую признательность видным фигурам (*Hat-tipping*);

t_4 − излишняя детализация (*Over-detailed*);

t_5 − излишняя тщательность (*Over-elaborate*);

t_6 − цитирование в случае очевидных фактов (*Evidentiary validity*);

t_7 − самоцитирование (*Self-serving*);

t_8 − умышленная игра (*Deliberate premeditation*);

t_9 – поиск грантов (*Searching out grant funding*);

t_{10} – привлечение внимания путем цитирования лучших образцов (*Finding support*);

t_{11} – цитирование по принципу предпочтений, которых придерживается печатное издание (*Editorial preferences*);

t_{12} – предубежденное цитирование (*Projective behavior*);

t_{13} – взаимное цитирование (*Conspiratorial cross-referencing*);

t_{14} – "под давлением", т. е. соответствие ожиданиям читающей публики (*Pandering to pressures*);

t_{15} – политика редакционной коллегии (*Editorial publication policies*);

t_{16} – незнание новых авторов (*Non-recognition*);

t_{17} – внутрипрофессиональная вражда (*Intra-professional feuding*);

t_{18} – устаревшие цитирования (*Obsolete citations*);

t_{19} – политические соображения (*Political considerations*).

Необходимость введения шкалы оценок обоснована в работе [Brooks, 1985], посвященной исследованию результатов опроса группы ученых, которым было предложено оценить мотивации ссылок в своих работах (список "B"). Рассматривалось семь шкал, калиброванных от нуля до трех. Значение "0" означает, что предложенная шкала не отвечает мотивациям, а значение "3" указывает, что данная шкала соответствует первичной мотивации. Ниже приведены название шкал, их история и соответствие спискам "*W*" и "*T*".

b_1 – актуальность (*Up-to-dateness*). В работе [Line, Sandison, 1974] предлагалось престиж документа определять в зависимости от того, насколько актуальными являются последние публикации, цитируемые автором.

b_2 – негативная оценка (*Negative credit*). Приводится информация об авторах, давших негативную оценку другим работам,

имеются в виду критика, корректировка, отрицание, дискуссия. В работе [Cole, Cole, 1971] отстаивается позиция, что незначимую работу не будут и критиковать, но, с другой стороны, сильно критикуемая работа может содержать плодотворную ошибку, вносящую значительный вклад в соответствующую область. Пункты w_5, w_6, w_7 и w_{14} могут попасть в эту категорию.

b_3 − техническая информация (*Operational information*). В работе [Moravcsik, Murugesan, 1975] техническое цитирование − это ссылка на техническое средство, используемое в цитирующей работе; соответствует w_3.

b_4 − убедительность (*Persuasiveness*). В работе [Gilbert, 1977] научные публикации рассматриваются как средство убеждения: автору нужно убедить коллег в правильности своих методов и результатов.

b_5 − положительная оценка (*Positive credit*). Например, w_1 и w_2 являются положительными оценками.

b_6 − указание читателям (*Reader alert*) на новую, отличающуюся и, возможно, непонятную работу (см. t_4, t_{10}, t_{12} и t_{13}).

b_7 − достижение консенсуса (*Social consensus*). В работе [Bavelas, 1978] подчеркивается, что многие ссылки делаются из неясных соображений и для достижения согласия между учеными в рассматриваемой области (см. t_{14}).

Далее Т. Брукс, используя статистические методы, провел анализ результатов опроса. Выяснилось, что имеется существенное различие между мотивациями цитирования. Более половины обозреваемых мотиваций относились к разряду убедительных (*Persuasive*), причем намного больше, чем ожидалось, со значением “3”, и намного меньше, чем ожидалось, со значением “0”. Негативная оценка (*Negative credit*) получила намного больше, чем ожидалось, значений “0”, и намного меньше значений “3”.

В целом получилось три группы: убедительность (*Persuasiveness*), получившая наивысший средний балл; вторую группу составляют положительная оценка (*Positive credit*), современность (*Up-to-dateness*), вниманию читателей (*Reader alert*) и техническая информация (*Operational information*); в группу с наименьшими средними баллами попали достижение консенсуса (*Social consensus*) и негативная оценка (*Negative credit*). Однако наблюдается значительное различие между распределениями для точных и гуманитарных наук, что указывает на различие в мотивациях для разных областей науки.

Еще одной важной работой, исследующей мотивации цитирующего автора, является работа [Vinkler, 1987], посвященная изучению степени научного давления, побуждающего к цитированию. Наименьшее значение этой величины называется “порогом цитирования”. В своей работе П. Винклер сообщает, что этот порог, прежде всего, зависит от того, насколько близко в профессиональном смысле находится цитируемая работа к исследованию, рассматриваемому в цитирующей работе. Работа П. Винклера комментируется в статье [Garfield, 1989].

2.2. Предположения Л. Смит

Рассмотрим предположения относительно свойств цитирования, лежащие в основе процедур анализа цитирования, и проблемы, касающиеся использования данных о цитировании. Эти предположения представлены в работе [Smith, 1981], поэтому мы пометили их буквой “*S*” и порядковым номером.

S_1 – *цитирование того или иного документа предполагает использование этого документа цитирующим автором*. Данное предположение фактически состоит из двух частей: 1) автор ссылается на все или, по меньшей мере, наиболее важные документы, использованные при подготовке своего труда; 2) все перечис-

ленные документы были действительно использованы, т. е. автор ссылается на некоторый документ только если этот документ внес вклад в его работу. Несоблюдение этих двух условий приводит к тому, что определенные документы недооцениваются, поскольку не все использованные статьи были процитированы, тогда как другие документы переоцениваются, поскольку не все процитированные статьи были использованы. В отношении недооценки автора, написавшего ту или иную статью, ясно, что цитирование не обязательно полностью и добросовестно отражает использование его труда. Часто то, что цитируется, представляет собой лишь небольшой процент того, что прочитано; не все, что было найдено полезным, цитируется. В работе [Davies, 1970] предлагается ввести “фундаментальный закон о ссылках”: “…совершенно необходимо самому прочесть или хотя бы просмотреть ссылку перед ее использованием”. Если не изучить текст как цитирующего, так и цитируемого документа, невозможно судить о том, действительно ли конкретное цитирование свидетельствует об использовании данного материала в цитирующем документе.

S_2 – *цитирование того или иного документа отражает достоинства* (*качество, важность, влияние*) *этого документа*. Базовое предположение при использовании индекса цитирования в качестве индикатора качества состоит в наличии высокой положительной корреляции между количеством цитирований конкретного документа (автора, журнала и т.п.) и качеством этого документа [Edwards, 1973]**.** Использование анализа цитирования для целей оценки – наиболее дискуссионный вопрос. Если авторы работы [Bayer, Folger, 1966] отмечают, что параметры, определенные с помощью индекса цитирования, сами по себе вполне правомерны, то автор работы [Thorne, 1977] считает, что индекс цитирования имеет сомнительную ценность, поскольку докумен-

ты могут цитироваться по причинам, не имеющим отношения к их достоинствам. Тем не менее, данное предположение было проверено и поддержано в ряде исследований включая исследование научных статей, журналов и ученых.

В каждом случае некоторые небиблиометрические параметры качества необходимо сравнивать с библиометрическими параметрами на основе индекса цитирования. Трудность состоит в том, что качество является сложным атрибутом, а общепринятые небиблиометрические параметры отсутствуют. Кроме того, нельзя автоматически предполагать, что документ (автор, журнал и т. п.), цитируемый нечасто, не имеет достоинств. В случае журналов, например, полезность цитирования как параметра качества журнала отличается в зависимости от функции журнала; новостные журналы могут иметь высокое качество, но цитируются нечасто. До тех пор пока не будет достигнуто большее понимание относительно причин цитирования, индекс цитирования может в лучшем случае рассматриваться как индикатор качества в грубом приближении. Небольшие различия в индексах цитирования не следует интерпретировать как значительные, но крупные различия могут быть интерпретированы как отражение различий качества и влияния. Результаты индекса цитирования должны сравниваться с альтернативными индикаторами качества с применением корреляционных методов.

S_3 – *цитаты приводятся из наилучших работ*. Можно лучше понять характер цитирования, если знать “генеральную совокупность”, из которой выбираются цитаты. Если полагать, что цитируются наилучшие из возможных трудов, нужно представить, что авторы просеивают все возможные документы, которые можно было бы процитировать, и тщательно выбирают те из них, которые считаются лучшими. Однако исследователи использования научной информации полагают, что доступность может быть

столь же важным фактором, как и качество, при выборе источника информации.

В работе [Soper, 1976] сообщается о результатах исследования эффекта “физической доступности” документа на выбор и использование ссылок. Обнаружено, что наибольшая часть документов, цитируемых в последних статьях авторов, находилась в личных собраниях, а незначительная часть находилась в библиотеках учреждений, к которым принадлежали респонденты, и самая незначительная часть располагалась в библиотеках других городов и стран. Таким образом, статья вполне могла быть процитирована, поскольку она в нужное время оказалась на письменном столе цитирующего, а не потому, что это идеальная статья для цитирования.

Доступность того или иного документа может быть обусловлена его формой, местом происхождения, годом написания и языком. Если это журнальная статья, ее доступность может определяться тиражом журнала, политикой перепечатки, наличием служб индексирования и реферирования. Также как некоторый документ может быть более или менее доступным, исследователь может быть более или менее заметным. Тот или иной автор, скорее всего, будет более осведомлен о трудах своих коллег. Как и в случае документов, цитируемые исследователи не обязательно являются наиболее выдающимися учеными в некоторой конкретной области.

S_4 – *насколько верны предположения о связи документов по содержанию*: а) цитируемого и цитирующего документов; б) документов, находящихся в состоянии библиографического сочетания (БиС); в) документов, находящихся в состоянии коцитирования. Ответ для а) дан, например, в работе [Barlup, 1969], в которой сообщается о результатах следующего эксперимента: авторов статей просили оценить, соотносится ли цитирование с их

собственным трудом. Авторы полагают, что 72 % определенно соотносится, и 5 % − определенно не соотносится. Ответ для б) можно найти в работе [Martyn, 1964], автор которой утверждает, что БиС не является действительной единицей измерения, поскольку не известно, насколько два документа, цитирующие третий, цитируют идентичные единицы информации в нем. Таким образом, БиС является всего лишь показателем существования вероятности (возможно, нулевой) зависимости двух документов по их содержанию. То же касается и случая в); факт коцитирования двух статей не гарантирует связи между их содержаниям.

S_5 − *все цитаты равны*. Зависимость между цитируемым и цитирующим документами весьма проблематична. Поскольку имеется целый ряд причин, благодаря которым существует цитирование, может существовать и целый ряд причин, по которым цитирующий автор не провел связь с некоторыми другими документами. Хотя наиболее очевидной причиной является то, что предыдущий документ нерелевантен для данной работы, это может быть вызвано также тем фактом, что автор не знаком с документом, не смог его достать или не знаком с языком, на котором он был опубликован. Как отмечает автор работы [Kochen, 1974]: “не удивительно, что есть значительная доля произвола в том, как авторы выбирают ссылки для своей библиографии. Несомненно, многие документы, которые надо было бы процитировать, потеряны; а многие документы, которые автор цитирует, релевантны лишь в незначительной степени”.

Несмотря на неопределенность, связанную с характером отношений цитирования, цитирование является привлекательным объектом исследования благодаря своей ненавязчивости и доступности. В отличие от данных, полученных в результате интервью и на основании анкеты, цитирование представляет ненавязчивый параметр, не требующий сотрудничества респондента и

сам по себе не засоряющий ответ [Webb, 1966]. Цитирование представляет собой указательные знаки, остающиеся после использования информации и в качестве таковых предоставляющие данные, на основании которых можно построить картину поведения пользователя без конфронтации с самим пользователем. Любая совокупность документов содержит списки литературы, которые могут послужить сырьем для анализа цитирования, и индекс цитирования на основе данной совокупности документов будет точен и объективен.

2.3. Проблемы

Следует отметить, что цитирование является личным процессом, несмотря на то что его результат становится общедоступным. Значительная субъективность акта цитирования означает, что можно только предполагать, по каким причинам автор сделал те или иные ссылки. Текстовый анализ цитирующей статьи не может выявить того, почему автор выбрал конкретные ссылки, хотя и можно указать правдоподобные причины. Отсутствие удовлетворительной теории цитирования частично объясняет, почему авторы часто прибегают к метафоре при попытке прояснения того, что не является самоочевидным. Метафора, подобно аналогии, имеет неуместную тенденцию к увеличению осведомленности за счет непонимания. По логике, использование ссылок в качестве основы ценных суждений должно предполагать, что среди авторов существует универсально признанная договоренность. Однако для этого нет достаточного основания, поскольку отсутствует согласованное понимание назначения ссылок. Так, Дж. Равец [Ravetz, 1971] интерпретирует ссылки как форму премирования, в то время как Дж. Гилберт [Gilbert, 1977] предпочитает рассматривать их как инструменты для подтверждения обоснованности суждений. С другой стороны, Г. Смолл [Small,

1978a] говорит о ссылках как о метках или символах. Поскольку процесс цитирования является субъективным и не поддающимся стандартизации, необходимо полагаться на очевидные причины цитирования или причины, которые могут быть аргументированы из контекста цитирующей работы.

Авторы работы [May, Talbot, 1967] оспаривали общепринятую точку зрения о том, что ссылки дают точную или правильную картину интеллектуальных связей между публикациями. С их точки зрения, существует значительная величина отклонения, полученная в результате "…недостаточности памяти, отсутствия самоанализа, небрежности, плагиата ссылок других людей без фактического их использования", кроме того, широко распространенного обычая не цитировать очевидные источники. Нельзя отрицать существование других причин, например все следствия такого простого факта, что автор выбирает ссылки для обслуживания своих научных, политических и персональных целей, а не для того, чтобы указать своих интеллектуальных предшественников.

Сознавая социальную значимость подсчета цитирований, Дж. Вейнер [Weiner, 1977] дал совет честолюбивому автору о способах получения цитирований при отсутствии таковых. Конечно, фетишизация ссылок может быть оправданна как не слишком серьезная демонстрация человеческого тщеславия. Даже если это допускается, то открытым остается вопрос о предумышленном злоупотреблении цитированием.

Примером подобного злоупотребления является так называемый скандал К. Бурта, который возник на почве того, что психолог К. Бурт (*Cyril Burt*) в 1937 г. предпринял попытку овладеть приоритетом на результаты английского психолога Ч. Спирмена (*Charles Edward Spearman*, 1863-1945) по корреляционному анализу. В 1951 г. К. Бурт опубликовал библиографию, содержащую ссылки на свои несуществующие публикации, включенные

в документ для добавления иллюзорного правдоподобия и веса. О. Джиль в работе [Gillie, 1980] ставит перед научной общественностью два вопроса: "…не появилась ли патологическая черта К. Бурта раньше?" и "… не занимался ли К. Бурт плагиатом?". Разумеется, это самая предельная иллюстрация нечестной практики цитирования. Тем не менее линия раздела между обманом и тем, что Дж. Равец [Ravetz, 1971] называет способностью "недостаточного цитирования без результатов воровства, или избыточного цитирования с эффектом раздувания ценности результатов коллеги" или тем, что Р. Мертон [Merton, 1973b] называет "криптомнезией" (непреднамеренный плагиат), временами является до некоторой степени размытой.

Возражения по поводу использования счетчиков цитирований обусловлены сложностью определения мотиваций цитирования. Объединим основные возражения в группы и снабдим их краткими характеристиками.

2.3.1. *Соавторство*. Цитируемые статьи, указанные в индексах цитирования, как правило, включают лишь авторов, указанных первыми. Для того чтобы найти все цитирования публикаций какого-либо автора включая те, в которых он не является первым автором, необходима работа с библиографией его трудов либо просмотр всей статьи указателя. Существует также проблема распределения зачетов в работах с соавторами [Lindsey, 1980]. Кроме того, следует ли рассматривать такие труды в индексе цитирования так же, как и труды одного автора, или применять пропорциональный зачет? Следует ли учитывать порядок следования имен авторов при таком зачете, поскольку данная последовательность часто является показателем вклада каждого автора в труд, о котором сообщается в публикации?

Предположим, что в цитируемом документе первым всегда указывается "главный" автор, а поиск в указателе цитирования

организован по первому автору, таким образом, получение информации о других авторах затруднено. Возникает первый вопрос: насколько серьезной является потеря информации о соавторах? Второй вопрос: как оценить соавторство при наличии информации о каждом авторе?

Вообще говоря, существует три пути подсчета: прямолинейный (*Straight*), когда работа относится к первому автору без учета соавторов; урегулированный (*Adjusted*), когда каждому автору приписывается доля; простой подсчет (*Normal*), когда все соавторы получают по единице.

В работе [Cole, Cole, 1973] предлагается использовать прямолинейный подсчет как наиболее соответствующий сущности явления. Кроме того, это сокращает издержки, так как не нужно разбираться в распределении долей значимости, и сокращает работу по сбору данных. В работе [Lindsey, 1980] указывается, что такой подход можно использовать в качестве стратегии. Однако выясним, имеются ли для этого предпосылки.

Прямолинейная процедура предполагает, что множество публикаций, в которых автор указан первым, является репрезентативной выборкой для всего множества работ этого автора. Тогда резонно считать, что порядок фамилий отражает степень значимости авторов. Если авторы указываются в алфавитном порядке, то указание на значимость отсутствует. В работе [Rudd, 1977] исследовались публикации 1500 химиков, и было получено следующее распределение по диапазонам первых букв в фамилиях первых авторов A–F, G–M, N–Z имеют распределение 56,8, 29,9, 13,3 % соответственно. Кроме того, в работе [Lindsey, 1978] при исследовании случайных наборов публикаций в семи областях науки выяснилось: наиболее вероятно, что среди членов совместных групп те, кто имеет фамилии, начинающиеся с первых букв алфавита, появятся первыми и в списке авторов. В работах

[Zuckerman, 1968], [Garfield, 1982] показано, что выдающиеся ученые получают в два раза больше цитирований в качестве вторых авторов, чем в качестве первых. Это говорит о том, что они регулярно передают авторство молодым коллегам. Таким образом, можно заключить, что нет эмпирической очевидности и теоретической предпосылки для использования прямолинейных счетчиков.

Возникает вопрос: как распределять публикации и цитирования между соавторами? Как показано в работе [Lindsey, 1980], при использовании простого счетчика выигрывают авторы, публикующие большое количество работ с большим количеством соавторов. Кроме того, суммарное количество работ, приписываемых авторам, может оказаться больше количества самих работ. Остается один выход: приписывать авторам равные доли от целого, т. е. если у работы два автора, то считается, что каждому принадлежит половина публикации. Это урегулированный подсчет. Вместе с авторами работ [Lindsey, 1980], [Long, et al., 1980], [Garfield, 1979] мы делаем вывод, что единственно правильный путь получения адекватных счетчиков – иметь всеобъемлющую библиографию и использовать урегулированные счетчики.

К сожалению, все указанные исследования имеют серьезный недостаток: они не могут объяснить, почему цитируемая ссылка была признана релевантной. Этот факт обусловлен отсутствием меры релевантности документов, т. е. формального механизма сравнения степени близости документов. Поэтому автор часто не осознает или не может распознать собственные причины цитирования конкретного источника и неиспользования для цитирования другого источника. Таким образом, опрос автора о его мотивах цитирования или нецитирования не может выявить действительных причин, почему автор процитировал так, как он это сделал на самом деле. Согласно работам датского психолога

Хйорланда [Hjørland, 2000], [Hjørland, 2002] существуют важные заповеди науки, которые настолько интегрированы в жизнь и культуру исследователей, что авторы частично или совсем не знают о них. Однако незнание этих заповедей не отменяет их существования.

2.3.2. *Самоцитирование*. Термин “самоцитирование” (*Self-citations*) может использоваться в различных значениях. Если цитирующая статья написана одним или несколькими авторами и цитирует автора или авторов из этого же набора, это считается самоцитированием. Однако в случае, когда цитирующая статья публикуется в том же журнале, что и цитируемая, это тоже считается самоцитированием относительно журнала.

Другой тип самоцитирования, рассматриваемый в рамках изучения научной политики, относится к случаю, когда цитирующая и цитируемая работа принадлежат авторам, работающим в одной и той же организации или рабочей группе (см. [Moed, et al., 1985a], [Moed, et al., 1985b], [Vinkler, 1986]). В работе [Earle, Vickery, 1969] термин “самоцитирование” употребляется для обозначения одной и той же предметной области в противоположность термину “самоотвод” (*Self-derivation*), когда цитируемая работа относится к другой тематике или научной области. Поэтому первый вариант иногда называют авторским самоцитированием (*Author self-citation*), чтобы отличить его от других.

В работе [Tagliacozzo, 1977] проведено систематическое исследование авторского самоцитирования в области физиологии растений и неврологии. Основные выводы таковы. Во-первых, только небольшое количество публикаций совершенно не содержит самоцитирований, и количество самоцитирований на статью имеет широкий диапазон. Более того, самоцитирований зачастую больше, чем цитирований других авторов. Авторы часто ссылаются на группу своих публикаций. Таким образом, если считать,

что цитирование указывает на важность, получается, что собственные работы и являются самыми важными. Во-вторых, несмотря на ожидания, в результате исследования выяснилось, что количество самоцитирований не впрямую связано с размерами библиографии. Также не замечено прямой взаимосвязи между количеством самоцитирований и продуктивностью автора.

В работе [Meadow, 1974] сделано предположение, что количество самоцитирований зависит от “зрелости” раздела науки. В свете такого предположения можно считать, что количество и пропорция “не самоцитирований” говорит о зрелости ученого (см. [Porter, 1977]).

В работе [Lawani, 1982] различаются два типа самоцитирований: синхронные и диахронные самоцитирования, т. е. сколько самоцитирований делает автор и сколько ссылок получает. Синхронную долю цитирования можно получить, если количество ссылок во всех работах автора за рассматриваемый период, включая совместные работы, разделить на количество ссылок на работы автора. Диахронную долю цитирования можно получить, если количество всех цитирований работ автора за рассматриваемый период разделить на количество цитирований, сделанных в работах, в которых участвует автор. В работе [Lawani, 1982] подчеркивается важность изучения этих параметров. Например, доля синхронного самоцитирования для ученого может быть существенной, а диахронного цитирования − небольшой.

В работе [Porter, 1977] утверждается, что для исследования отрасли науки в целом самоцитирования не играют роли, в то время как подход, представленный в работе [Lawani, 1982], можно использовать для оценки авторов.

Факт самоцитирования в случае единственного автора публикации очевиден. Исключить учет такого самоцитирования из индекса цитирования, можно, но правильно ли это, является

вопросом, требующим специального рассмотрения. Для статей, написанных в соавторстве, сам термин “самоцитирование” не является прозрачным. Еще более трудная проблема состоит в исключении группового самоцитирования, т. е. ссылок любого члена (членов) той или иной исследовательской группы на любого другого члена (членов) той же исследовательской группы. В данном случае необходимо работать с источником информации, идентифицирующим всех членов исследовательской группы и, прежде всего, найти такой источник.

2.3.3. *Омонимы*. Некоторые авторы имеют совпадения в имени и фамилии. Для различения требуется дополнительная информация. Особенную трудность составляют китайский и японский языки [Cornell, 1982].

2.3.4. *Синонимы.* Требуется установить стандарт написания имен. Иначе возникает проблема упоминания одного автора с разным количеством инициалов, по-разному переведенной на иностранный язык фамилией и т. д., например, *Derek de Solla Price.*

2.3.5. *Типы источников.* В работе [Line, 1979] показано, что от типа источника может зависеть результат. Например, журналы и монографии могут демонстрировать различие в организации материала, размере списков ссылок и т. д.

2.3.6. *Неявные цитирования*. Неявное цитирование – это цитирование без указания источника, например закон Ципфа. В случае очень известного достижения вообще не принято указывать источник, например ссылаться на Ю. Гарфилда при использовании аббревиатуры *SCI*.

2.3.7. *Зависимость от времени.* Счетчики цитирования могут существенно различаться для разных лет. Попытка исследовать флуктуации цитирований журналов сделана в работе [Nieuwenhuysen, Rousseau, 1988].

2.3.8. *Зависимость от научной области*. Практика в сфере публикаций сильно зависит от научной области, это ведет к затруднениям при сравнении специалистов разных областей [Pinski, Narin, 1976].

2.3.9. *Неполнота библиографических баз данных*. На неполноту ББД, на основании которых строятся индексы цитирования, указывают, например, работы [Vehlo 1986], [Vehlo 1987], [Gaillard, 1989] и др.

2.3.10. *Доминирование английского языка*. На предпочтение цитирования англоязычных работ указано в работах [Vinkler, 1986], [Garfield, 1979a]. Этот “перекос” в большей степени наблюдается в общественных науках.

2.3.11. *Преобладание работ из США*. В ББД *WoS* наблюдается значительное смещение в сторону работ ученых из США. В работе [Cronin, 1981a] показано, что в среднем американский автор в 95 % случаев ссылается на работы авторов из США, а английский – приблизительно в 40 % случаев. Более того, американские журналы печатают работы американских авторов в семь раз чаще, чем английских. Английские авторы разделяют ссылки поровну между США и Объединенным королевством [Inhaber, Alvo, 1978].

2.3.12. *Гендерный вопрос*. В работе [Ferber, 1986] показано, что авторы часто предпочитают ссылаться на работы авторов того же пола. Это особенно заметно в тех областях, где исследователи-мужчины составляют большинство.

2.3.13. *Ошибки*. Индексы цитирования часто основываются на плохо подготовленных данных, в которых могут встречаться ошибки; в работе [Goodrich, Roland, 1977] показано, что такое случается не так редко. Кроме ошибок, которые можно назвать случайными, встречаются и систематические ошибки, например, могут быть проиндексированы документы, помеченные грифами “в печати” или “не опубликовано”.

В последнее время проблема оценки научного влияния с помощью библиометрических методов приобрела особую актуальность. Исследователи разделились на два лагеря [Bornmann, Daniel, 2008]. Одни считают, что вычислительный анализ подходит для оценки научного вклада, и изучение материалов показывает, что, например, оценка по количеству цитирований автора адекватно соответствует оценке коллег, проявляющейся в различного рода наградах. Другие сомневаются в том, что счетчики цитирования, являющиеся функцией многих переменных, верно отражают важность научной деятельности. Они считают, что вероятность цитирования работы зависит от многих факторов, не связанных с принятыми соглашениями по научным публикациям. Перечислим основные факторы:

− *время* (в связи с увеличением объемов научной продукции вероятность цитирования любого документа со временем возрастает, кроме того, наблюдается явление “успех порождает успех” (см. эффект Матфея));

− *область деятельности* (наблюдается различная культура цитирования; кроме того, узкие научные области привлекают меньшее число цитирований);

− *место публикации* (влияние частоты публикации работ, связанных с рассматриваемой публикацией; расположение статьи в журнале (первая статья привлекает больше цитирований); кроме того, влияет качество и престиж журнала);

− *тип публикации* (известно, что по степени цитируемости различаются методологические статьи, обзоры, исследования, письма, заметки, книги; существует также зависимость от количества авторов, размеров публикации (чем больше публикация, тем больше причин для ее цитирования));

− *связь между автором и читателем* (язык публикации, персональное знакомство);

– *доступность публикации* (физическая доступность, например возможность доступа в реальном времени, влияет на количество цитирований);

– *технические проблемы* (некорректные ссылки, синонимы, омонимы).

В работе Б. Кронина [Cronin, 1982] указывается, что центральной проблемой использования счетчиков цитирования для оценки научной деятельности является отсутствие формализации норм и соглашений по цитированию. Использование счетчиков цитирования в качестве оценки вклада может быть адекватным лишь в том случае, если цитирующие авторы использовали только действительно лучшие работы.

Сложно постичь все особенности предвзятости и субъективности, будучи удаленным от образа действий при цитировании, хотя прогресс в исследовании искусственного интеллекта, в конечном счете, может освободить авторов от необходимости решать, где и на что делать ссылки при написании статей о формальных исследованиях. Еще в 1964 г. Ю. Гарфилд обсуждал возможность цитирования, которое генерируется компьютером автоматически, без участия автора [Garfield, 1965]. Тем не менее, до сих пор нет ответа на вопрос об использовании количественной оценки в качественном измерении, который бы полностью удовлетворял всех. Нумерология (гадание по числам) … интересна, но она должна использоваться с предельной осторожностью. В конкретных случаях она может вводить в заблуждение.

2.4. Необходимость теории цитирования

Третья глава монографии Б. Кронина “Процесс цитирования” [Cronin, 1984] начинается с метафоры: “Цитирования представляют собой застывшие отпечатки на ландшафте научных достижений, которые являются свидетельством развития идей”.

И далее: “На основе этих отпечатков можно сделать выводы о направлении исследований; по конфигурации и глубине оставленных следов можно воссоздать изображение тех, кто был известен, в то время как распространение и многообразие обеспечивают ключ к разгадке того, было ли данное продвижение вперед методичным и систематическим. Таким образом, цитирования обеспечивают реальное выражение процесса инноваций в части того, что касается роста и развития научного знания, и, если они правильно выстроены, могут предоставить исследователю аналитический инструмент с привлекательной мощностью и многосторонностью”. Относительно объективная информация, предоставляемая ссылками, была неоднократно и впечатляюще использована в различных приложениях. Например, можно использовать ссылки как “количественную меру” для сравнения результатов, полученных исследователями-одиночками или научными коллективами в пределах одной отрасли науки, варьируя степень доверия.

Совокупность знаний, существующая в любой научной области, представляет собой накопление отфильтрованных взглядов, теоретических построений, экспериментально полученных данных и эмпирических наблюдений. Литература, опубликованная в определенной тематической области, представляет собой выборочные, отредактированные и одобренные фонды знаний, которые при разумной систематизации могут отображать генеалогию достижений в данной области исследований. Индексация цитирования, получаемая из перекрестных ссылок, открыла новые горизонты и возможности социологических исследований в науке. Историк науки может использовать анализ цитирований для установления происхождения идей и объяснения цепочек научного взаимодействия.

Индексирование цитирований основано на том предположении, что библиографические ссылки являются выражением связи

между двумя документами − цитирующим и цитируемым. Технология индексирования не является точным выражением этой связи, она является не чем иным, как прямым, хотя и достаточно сложным и дорогим с коммерческой точки зрения переоформлением публичных и свободно доступных данных. В основе этой технологии лежат два предположения. Первое: данные цитирования можно обрабатывать количественно. Второе: все ссылки имеют равную ценность. Разумеется, в действительности некоторые ссылки являются "более равными", чем другие, т. е. можно утверждать, что наличие ссылки может означать, что на автора *A* оказала воздействие работа автора *B*, но, основываясь только на этом, ничего нельзя сказать о степени или силе такого воздействия. Тем не менее, если каждой ссылке присвоить значение "единица", то можно легко ощутить преимущества этого метода.

Существует значительный объем данных, подтверждающих, что результаты, полученные на основе подсчета цитирований, коррелируют с различными субъективными и объективными средствами оценки научной деятельности. Наиболее важный вклад внесли Дж. и С. Коулы, которые показали, что высокие показатели цитирования имеют положительную корреляцию с такими признанными показателями качества, как почетные награды, Нобелевские премии и репутация [Cole, Cole, 1971]. Это открытие было поддержано в работе [Hagstrom, 1971], автор которой исследовал корреляцию цитирования с такими переменными, как качество профессорско-преподавательского состава и гранты, полученные факультетами высших учебных заведений США. В области научной политики Национальный научный фонд США (*National Science Foundation, NSF*) использует результаты анализа цитирования как один из показателей эффективности исследовательских программ, финансируемых фондом.

Тем не менее, наряду с практическими преимуществами существуют неразрешенные понятийные и методологические вопросы (см., например, [Edge, 1979]). Эта критика не осталась незамеченной. В статье [Garfield, 1980], которую можно рассматривать как “отклик” на критику, Ю. Гарфилд указал, что можно отказаться от подсчета цитирований и перейти к подсчету того, на скольких авторов оказала воздействие данная публикация. Однако до настоящего времени не произошло никаких изменений процедуры подсчета цитирований.

2.5. Аргументы pro et contra

Ряд исследователей задает вопрос: нужна ли теория цитирования вообще? Более того, было сделано предложение прекратить “теоретизирование” и вернуться к позиции логического позитивизма: “Полагаю, что текущее состояние нашей области требует больше эмпирической и практической работы, а все это теоретизирование может подождать, пока не будет создан большой задел – выше порога эмпирических знаний”, − заявляет автор работы [Arunachalam, 1998]. Создается мнение, что некоторые специалисты по анализу цитирования согласились с тем, что авторы часто не отдают должное там, где это требуется. Это заявление, по их мнению, не отрицает анализа цитирования. Например, автор работы [Small, 1987] считает: “Проблема не в том, можем ли мы полагаться на перечни библиографических ссылок в отдельных случаях как на полные комплекты источников влияния (мы не можем), а, скорее, в том, можно ли использовать библиографические ссылки с точки зрения статистики, как индикатор влияния”. Авторы работы [Nederhof, Van Raan, 1987] пустили в ход тот же аргумент: “При рассмотрении библиографических ссылок, содержащихся в одной отдельной статье, можно найти большое количество “неправильностей”, таких как пропущенные

ссылки на важные статьи или на работы авторов, которые сделали важный вклад в совокупность знаний в этой области. Таким образом, когда для этой цели используется одна конкретная работа, может быть получена ошибочная картина основных источников влияния в конкретной области".

Актуальной темой в библиометрии является интерпретация, использование и создание новых индикаторов науки на основе результатов цитирования. Общая теория цитирования могла бы сделать смысл индикаторов более прозрачным, что, несомненно, подняло бы уровень доверия к ним. Отсюда следует сохраняющаяся потребность в теории цитирования, которая до сих пор не разработана экспертами в области библиометрии, социологии и психологии. П. Воутерс [Wouters, 1999a] предполагает, что исчерпывающая теория цитирования должна быть способна: а) предоставить теоретический фундамент для анализа цитирования; б) подтвердить необходимость и целесообразность использования индикаторов в области научных исследований; в) предоставить теоретическое объяснение манеры цитирования ученых и филологов. Поэтому понимание манеры цитирования является ключевой предпосылкой теории цитирования. Результаты серии работ по исследованию списков цитируемой литературы в научных документах не привели к недвусмысленному объяснению цитирования [Cronin, 1981], [Smith, 1981], [Cronin, 1984], [Leydesdorff, 1987], [Cozzens, 1989], [MacRoberts, MacRoberts, 1989], [Luukkonen, 1990]. Таким образом, анализу цитирования недостает общего теоретического обоснования. Кроме того, использование анализа цитирования в качестве индикатора научных заслуг вызывает серьезное возражение, поскольку открывается доступ к оценке научной деятельности для "посторонних".

Необходимость в теории цитирования возникла в 1970-х гг., когда социологи и исследователи информатики осознали потреб-

ность в теории, объясняющей, почему авторы цитируют так, как они это делают. Первой работой, в которой прозвучало высказывание “за теорию”, видимо, следует считать работу [Mulkay, 1974], автор которой заявил: “Отсутствует ясность в вопросе о том, как библиографические ссылки отражают процесс научного влияния”. Этот факт привел автора к выводу: “На самом деле мы очень мало знаем о том, кто кого цитирует в науке и почему”. Затем прозвучало высказывание о необходимости создания “удобного и быстрого метода для обнаружения “природы” значимой связи, которую установил цитирующий автор”.

В 1981 г. три автора подняли эту проблему одновременно и независимо друг от друга, рассмотрев ее с разных позиций. С. Козенс [Cozzens, 1981] критически рассмотрел существующие теории цитирования с точки зрения социологии. Б. Кронин [Cronin, 1981] высказал аргументы за теорию цитирования с точки зрения информационного поиска. Л. Смит [Smith, 1981] пришла к выводу о недостаточности знаний о манере цитирования авторов и о важности таких знаний в поисках смысла в использовании анализа цитирования для конкретных применений.

Возможно, одной из первых работ, содержащих сравнительную оценку коллег с оценкой, основанной на цитировании, является работа [Virgo, 1977], преследующая две цели. Во-первых, показать, что в данной дисциплинарной области более цитируемые работы являются и более значимыми. Во-вторых, с помощью регрессионного анализа выявить факторы, ассоциирующиеся с важными работами.

Каждому участнику экспериментальной группы предлагалось рассмотреть библиографию, относящуюся к его определенному исследованию в области медицины. Участники оценивали степень уместности ссылок в библиографиях со своей точки зрения. Были оставлены только те работы, ссылки на которые были

признаны “очень уместными”. Затем с помощью *WoS* были получены индексы цитирования выбранных работ, причем рассматривались цитирующие работы за три года, последующих после публикации. Работы были ранжированы на основании количества цитирований. Затем выбрали пять наиболее цитируемых и пять наименее цитируемых работ. После этого на основании случайного выбора образовали пары: одна работа из группы с высоким цитированием, другая − из группы редко цитируемых работ. Пары были представлены участникам эксперимента в сопровождении серии вопросов, на которые нужно было ответить. Затем каждому участнику предлагалось назвать двух выдающихся ученых США, работающих с ним в одной области. Одного из двух выбранных ученых, согласившегося принять участие в эксперименте, просили оценить те же статьи, которые были предложены первому участнику. Оказалось, что в среднем частота цитирования лучше предсказывает важность работы, чем даже суждение второго судьи, что является аргументом в пользу анализа цитирования. Также было показано, что импакт-фактор журнала (среднее количество цитирований на статью, опубликованную в журнале, за определенный период времени), в котором напечатана работа, вместе с частотой цитирования в значительной степени способствует выявлению рейтингов публикаций.

Справка. Регрессионный (линейный) анализ − статистический метод исследования влияния одной или нескольких независимых переменных X_1, X_2, ..., X_p на зависимую переменную Y. Независимые переменные иначе называют регрессорами или предикторами, а зависимые переменные − критериальными. Терминология *зависимых* и *независимых* переменных отражает лишь математическую зависимость переменных, а не причинно-следственные отношения. Цели регрессионного анализа: а) определение степени детерминированности вариации критериальной (зависимой) переменной предикторами (независимыми переменными); б) предсказание значения зависимой переменной с помощью независимой (независимых); в) определение

вклада отдельных независимых переменных в вариацию зависимой переменной [Википедия].

Цитирования строго коррелируют с другими мерами важности. В работе [Myers, 1970] приводятся доводы в пользу анализа цитирования. Здесь рассматривается список наиболее цитируемых авторов в области психологии и 15 различных мер важности. Сделан вывод, что частота цитирования является хорошим показателем для оценки ученых. Аналогичные выводы сделаны и в работе [Narin, 1976], где проводилось сравнение библиометрических мер с другими способами оценки важности: коэффициент корреляции варьируется от 0,5 до 0,8.

Справка. Корреляция (корреляционная зависимость) – статистическая взаимосвязь двух или нескольких случайных величин (либо величин, которые можно с некоторой допустимой степенью точности считать таковыми). При этом изменения значений одной или нескольких из этих величин сопутствуют систематическому изменению значений другой или других величин. [Гмурман, 2004]. Математической мерой корреляции двух случайных величин служит корреляционное отношение η, либо коэффициент корреляции R. В случае если изменение одной случайной величины не ведет к закономерному изменению другой случайной величины, но приводит к изменению другой статистической характеристики данной случайной величины, то подобная связь не считается корреляционной, хотя и является статистической. Впервые в научный оборот термин “корреляция” ввел французский палеонтолог Жорж Кювье в XVIII в. [Википедия].

В работе [Broadus, 1977] проведено исследование тезиса о том, что “ученые сильно полагаются на цитирования при поиске библиотечных материалов”. Автор обнаружил существование соответствия между счетчиком цитирований и другими методами оценки ученых, несмотря на некоторые несообразности. В более позднем обзоре [Bensman, 1982] сделано заключение о росте согласия по поводу обоснованности использования анализа цитирования.

Первый номер журнала "*Scientometrics*", вышедший в 1998 г., был полностью посвящен обсуждению аргументов "за" и "против" теории цитирования. Инициировал эту дискуссию Л. Лейдесдорф [Leydesdorff, 1998] статьей с провокационным названием "Теории цитирования?", в которой он приводил доводы в пользу того, что всеобъемлющая теория так и не была сформулирована, хотя для анализа цитирования уже было апробировано большое разнообразие методов и информационных массивов. Здесь уместно ответить на вопрос: "Что принято считать теорией?". Приводим ответ популярной энциклопедии на этот вопрос.

Справка. Теория (от греч. − рассмотрение, исследование) − учение, система идей или принципов − является совокупностью обобщенных положений, образующих науку или ее раздел. Теория выступает как форма синтетического знания, в границах которой отдельные понятия, гипотезы и законы теряют прежнюю автономность и становятся элементами целостной системы. В теории каждое умозаключение выводится из других умозаключений на основе некоторых правил логического вывода. Способность прогнозировать − следствие теоретических построений. Теории формулируются, разрабатываются и проверяются в соответствии с научным методом. В "чистых" науках теория − произвольная совокупность предложений некоторого искусственного языка, характеризующегося точными правилами построения выражений и их понимания. Обычно считают, что стандартным методом проверки теорий является прямая экспериментальная проверка ("эксперимент − критерий истины"), результат которой уточняет или расширяет положения этой теории. Таким образом, прикладная цель науки − предсказывать будущее как в наблюдательном (аналитическом) смысле − описывать ход событий, на который мы не можем повлиять, так и в синтетическом − создавать посредством технологии желаемое будущее. Образно говоря, суть теории в том, чтобы связать воедино "косвенные улики", вынести вердикт прошлым событиям и указать, что будет происходить в будущем при соблюдении определенных условий [Википедия].

Введение Ю. Гарфилдом в 1963 г. "Указателя научного цитирования" (*SCI*) отметило важный этап в истории информатики.

Возможность поиска документов в соответствии с полученными цитированиями помогла существенно улучшить ранее существовавшие методы поиска на основании терминов. Более того, поиск на основе цитирований изменил понимание связанности документов. Однако, как правильно заметил Г. Смолл [Small, 2000], Ю. Гарфилд не изобрел цитирование, цитирование является зеркальным отражением библиографической ссылки. Таким образом, для понимания природы цитирования необходимо понять природу библиографической ссылки. А для понимания природы библиографической ссылки требуется теория цитирования, которая объясняет, почему авторы цитируют именно так, как они это делают. Игнорирование библиотечной ссылки (т. е. игнорирование истории цитирования) для понимания цитирования логически невозможно, к такому выводу пришел Х. Моед в монографии по анализу цитирования: “Теории библиографических ссылок и цитирования хотя и различны, основаны на понимании того, что ученые стремятся выразить в своих методах выполнения библиографических ссылок” [Moed, 2005].

2.6. Горизонты теории цитирования

Создание теории цитирования, которая дает объяснение цитированию, связывая его с манерой цитирования ученого, не является тупиком. Наоборот, это единственный путь вперед, если нам требуется понять полный потенциал анализа цитирования. Значительная часть предыдущих исследований манеры цитирования предоставила всего несколько деталей для решения головоломки цитирования. Исследования имели тенденцию покоиться на предположениях, что цитирование лучше всего понимается как физиологический процесс и что теории цитирования следует строить на исследованиях отдельных цитирующих авторов посредством опросов, размышления вслух или записи моделей поведения.

Это направление исследований создало ряд схем классификации, собравших различные причины цитирования. Но они, как правильно заметили С. Балди [Baldi, 1998] и Б. Кронин, имеют ограниченное использование. Мы поддерживаем мнение Б. Кронина [Cronin, 1994]: «Объясняя манеры цитирования, мы имеем дело с “черным ящиком”»

Й. Николайсен [Nicolaisen, 2003] рассматривает вопрос о том, как и какие компоненты и ограничения влияют на исследовательскую работу и, в конечном счете, на цитирование. Он считает, что для того, чтобы понять, объяснить и предсказать динамику сети цитирования, потребуется проникнуть в социальные миры отдельных авторов. Очевидно, что сделать это непросто. Необходима общая теория цитирования, которая позволяет нам изучать социальное действие. Такая теория должна не только объяснять, как ученые адаптируются к своим социальным окружениям, но и как адаптация влияет на форму научного исследования и, следовательно, на написанные работы и библиографические ссылки цитирующих авторов.

Й. Николайсен считает, что философские работы [Laudan, 1977] и [Nickles, 1981] предоставляют часть требуемого обоснования. Т. Никлз убедительно показывает, что, поскольку проблема заключает в себе все ограничения на ее решение вместе с требованием того, чтобы это решение было найдено, данные ограничения фактически представляют собой определяющую часть самой проблемы; они характеризуют проблему и задают ее структуру. Л. Лаудан опирается на очевидный факт, что явное требование того, что данное решение должно быть найдено, возникает из заданных показателей традиции научных исследований. Далее он демонстрирует, что традиции научных исследований характеризуются существованием набора теорий, иллюстрирующих данную традицию, методологических приверженностей

и истории (или традиции). Вместе эти характеристики устанавливают границы исследования и используемой методологии.

Уточнение проблемы дается в работах [Faigley, 1986] и [Bazeman, 1988], в которых показано, что авторы действуют как члены группы с определенными традициями написания работ, ограниченными рамками эпистемологических и социальных правил для дисциплин. Акт цитирования тесно связан с социальными соглашениями по поводу традиций написания работ.

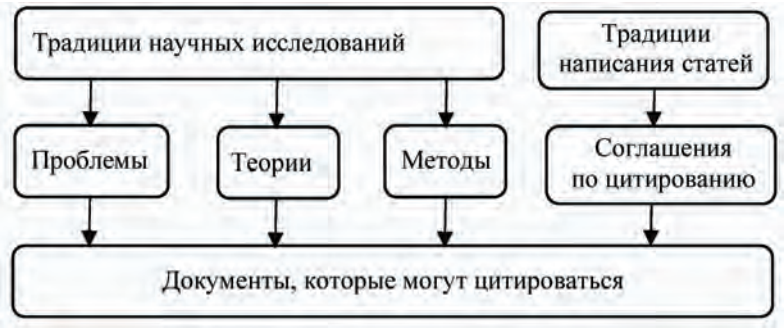

Рис. 2.6.1. Компоненты общей теории цитирования

Согласно Й. Николайсену, основными компонентами общей теории цитирования являются социально установленные дисциплины, традиции научных исследований, проблемы, теории, договоренности по цитированию и документы, которые противостоят цитирующим авторам. Эти компоненты взаимодействуют друг с другом множеством способов. Некоторые взаимодействия очевидны, в то время как другие более туманны. На рис. 2.6.1. показано влияние адаптации на форму научных исследований, а следовательно, на письменные работы и библиографические ссылки цитирующих авторов. По сути, рис. 2.6.1. можно рассматривать как наметки общей теории цитирования.

Рассмотрение цитирования, скорее, как социального действия, а не частной практики, не вынуждает принять заявления

радикального социального конструктивизма. Автор попытался добиться компромисса между наивным реализмом и радикальным релятивизмом, который, с одной стороны, признает ограничивающие силы коллективов, с другой − отвергает идею устойчивости взаимосвязей. Данный компромисс определяет типы фундаментальных сущностей, имеющихся в области изучения, их взаимодействие и возможности исследования этих влияний.

Итак, теория цитирования, базирующаяся на социальных аспектах процесса цитирования, может выгодно отличаться от нынешней группы теорий библиометрических индикаторов, рассматривающих цитирование как частную практику ранжирования ученых и научной периодики с использованием в качестве аргумента коллекции библиотечных ссылок. Завершим этот параграф констатацией факта из работы [Garfield, 1983]: “Анализ цитирования не является заменой или умалением оценки коллег; это отправная точка для тех, кто хочет использовать средства для более тщательной оценки своего труда”.

Глава 3. Продуцирование и характеристики цитат

"…когда смотришь на предмет,
важнее всего сам взгляд"

М. Пруст

Часть любого научного документа представляет собой список ссылок, указывающих на предыдущие публикации. Ссылки представляют собой подтверждение, которое один документ *дает* другому; цитирование представляет собой подтверждение, которое один документ *получает* от другого. В целом цитирование предполагает зависимость между той или иной частью или всем цитируемым документом и частью или всем цитирующим документом, считает автор [Malin, 1968]. Анализ цитирования представляет собой область библиометрии, исследующую данные зависимости. Развитие анализа цитирования было отмечено изобретением новых приемов и параметров, использованием новых инструментов и исследованием различных единиц анализа. Эти тенденции привели к стремительному росту количества и типов исследований, в которых используется анализ цитирования.

3.1. Развитие методов анализа

Начнем с цитаты [Прайс, 1966]: "Исходным пунктом нашего рассуждения будут эмпирические свидетельства статистики, почерпнутые из множества справочников по различным отраслям и аспектам науки. Все они с удивительной настойчивостью и постоянством показывают, что если найден более или менее удовлетворительный способ измерить какой-либо достаточно большой сегмент науки, то этот сегмент в нормальных условиях растет экспоненциально". Эту мысль подтверждает рис. 3.1.1, на котором показан рост общего числа научных и реферативных журналов в период с 1665 по 2000 гг. В своей работе Д. Прайс

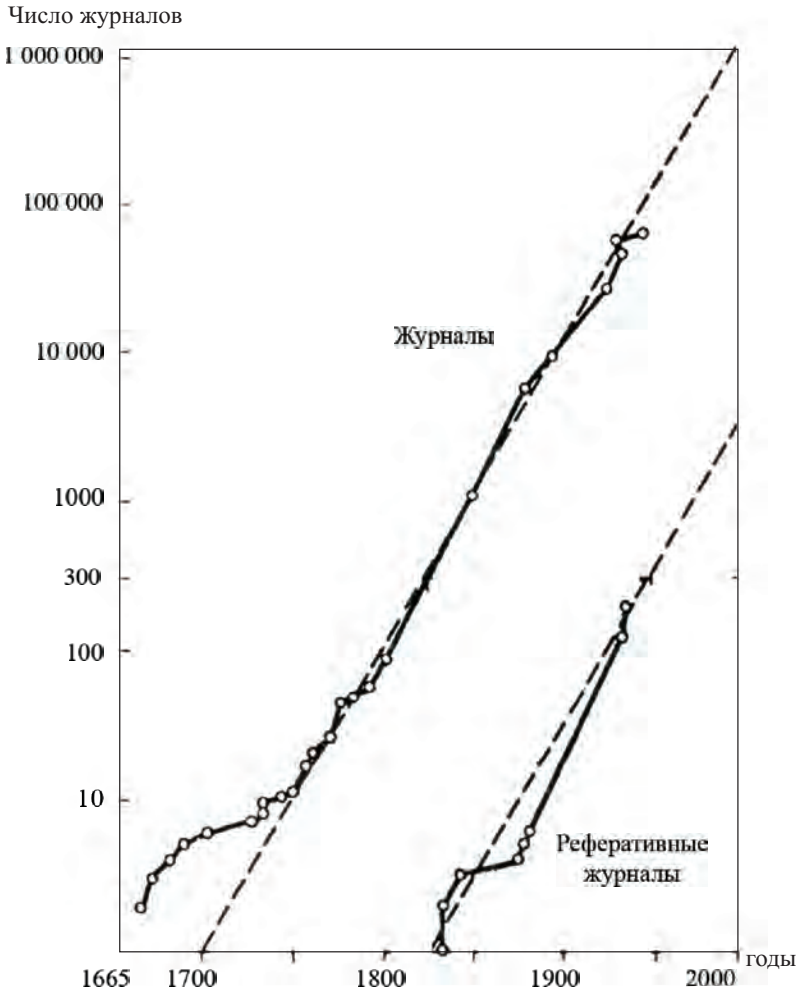

Рис. 3.1.1. Рост общего числа научных и реферативных журналов

отмечает, что реферативные журналы, начинают издаваться в тот момент, когда общее число журналов (по наименованиям) достигает значения 300.

Для общения ученых помимо публикации трудов используются различные формы конференций, симпозиумов и рабочих групп

и, конечно, личные связи. В XIX в. важную роль стали играть реферативные журналы (далее − РЖ), которые выпускались и выпускаются по отраслям наук в ряде стран.

Справка. В СССР и РФ выпуском РЖ занимается Всероссийский институт научно-технической информации Российской академии наук (ВИНИТИ РАН), который является одним из крупнейших центров научной и технической информации в области точных, естественных и технических наук в России и мире. ВИНИТИ был создан Постановлением Совета Министров СССР в 1952 г. по инициативе президента АН СССР академика А.Н. Несмеянова и первоначально носил название Институт научной информации. Основные задачи ВИНИТИ РАН − выполнение научных исследований и информационное обеспечение фундаментальных и прикладных исследований, инновационных работ в области точных, естественных и технических наук. Институт создавался как советский вариант информационного механизма мировой науки, в качестве образцов использовались такие РЖ, как *Mathematical Reviews, Chemical Abstracts*. В состав отраслевых редакционных коллегий РЖ входили более 140 академиков и 75 членов-корреспондентов РАН. Членами редколлегии РЖ "Математика" были выдающиеся математики академики А. Н. Колмогоров (1903−1987), П. С. Александров (1896−1982), П.С. Новиков (1901−1975) [История ВИНИТИ]. В течение 30 лет, с 1956 по 1986 гг. институт возглавлял профессор А. И. Михайлов (1905 −1988). Со временем его руководства (в эти годы в институте работало 20 000 сотрудников) связывают становление государственной системы научно-технической информации и завоевание ВИНИТИ мирового признания [Википедия]. В 60-80-е гг. XX в. теме "наукометрия" серьезное внимание уделяли в своих работах пионеры отечественной информатики В. В. Налимов, А. И. Михайлов, А. И. Черный, Р. С. Гиляревский [Михайлов и др., 1968]. В 1974 г. в ВИНИТИ были предприняты попытки создания отечественного указателя научного цитирования (УНЦ). Более подробно о "наукометрии" в СССР можно прочитать в работе [Грановский, 2000].

Наиболее простым способом осуществления измерений с помощью цитат является определение индекса цитирования – количества цитирований, полученных данным документом или совокупностью документов за тот или иной период времени из

конкретной совокупности цитирующих документов. При применении данного индекса к статьям, опубликованным в том или ином конкретном журнале, он может быть уточнен путем расчета импакт-фактора, среднего количества цитирований, полученных статьями, опубликованными в некотором журнале за указанный период времени. Данный параметр позволяет сравнивать "импакт" журналов, опубликовавших разное количество статей. Г. Пинский и Ф. Нарин, в работе [Pinski, Narin, 1976] предложили дальнейшее рафинирование индекса цитирования с учетом объема статьи, престижности цитирующего журнала и других характеристик.

История создания индексов (или указателей) научного цитирования начинается с 70-х гг. XIX в., когда практически одновременно появляются индекс юридических документов *Shepard's Citations* в 1873 г. и индекс научных публикаций по медицине *Index Medicus* в 1879 г. (просуществовал вплоть до 2004 г., всего было издано 45 выпусков). Революционный шаг в области библиотечного дела (переход от бумажных носителей к цифровым) был сделан в США в 1960 г., когда Ю. Гарфилд организовал Институт научной информации (*Institute for Scientific Information, ISI*). Эта организация занималась вопросами составления библиографических баз данных (ББД) научных публикаций, их индексированием и определением индекса цитируемости, импакт-фактора и других статистических показателей научных работ. Основным продуктом компании с 1961 г. является *Science Citation Index* (*SCI*), первоначально охватывавший данные приблизительно из 600 журналов и увеличивший это количество к 2010 г. до 16 520.

В это же время были изобретены два метода идентификации связи между документами: библиографическое сочетание (далее БиС) [Kessler, 1963] и коцитирование (далее КЦ) [Маршакова, 1973], [Small, 1973]. Теория и практическое применение методов

БиС и КЦ были проанализированы в работах [Weinberg, 1974] и [Bellardo, 1980] соответственно. Индекс цитирования и БиС были характерными методами анализа цитирования в 1960-е гг., однако в 1970-е большинство исследований сосредоточилось вокруг анализа на основе КЦ, который представляет особый интерес как средство картографирования научных специальностей [Small, Griffith, 1974].

В 1973 г. началась публикация "*Отчета о цитировании журналов*" *(Journal Citation Report, JCR*) в виде дополнительного тома к индексу цитирования. В этом отчете приводятся импакт-факторы всех журналов, индексируемых ББД *WoS*, он также содержит классификацию журналов по цитированию и импакт-факторам, а также два ранжированных списка для каждого журнала: журналы, которые цитируют данный журнал наиболее интенсивно, и журналы, которые в данном журнале наиболее часто цитируются [Garfield, 1973]**.** Также ежегодно публикуется список наиболее цитируемых ученых, на основе которого, в частности, составляется "Академический рейтинг университетов мира". В 1992 г. *ISI* был преобразован и с 2006 г. функционирует в составе *Thomson Reuters***.**

Заметим, что до середины 1960-х гг. обработка библиометрической информации осуществлялась вручную. Поскольку это был достаточно трудоемкий и нудный процесс, большинство экспериментов, естественно, были весьма ограниченными по объему. Ранние исследования цитирования часто основывались на списках ссылок из статей, опубликованных в небольшом количестве журналов. Появление вычислительной техники значительно улучшило ситуацию, стало возможным продуцирование индексов, содержащих данные цитирования из тысяч документов. Начались широкомасштабная разработка и применение новых методов анализа цитирования, основанных на библиографических

базах данных, осуществляющих на регулярной основе индексацию цитирований своего массива документов.

В 2005 г. в работе [Hirsch, 2005] был представлен *h*-индекс (индекс Хирша) в качестве адекватной оценки научной продуктивности исследователя, альтернативной простым характеристикам, таким как общее число публикаций или общее число цитирований. Этот библиометрический показатель, основанный на количестве публикаций и количестве цитирований этих публикаций, может использоваться в качестве количественной оценки продуктивности ученого, группы ученых, университета или страны в целом.

В настоящее время речь идет об индексировании цитирования для различных уровней агрегирования. В качестве единиц анализа могут выступать отдельные статьи или книги, журналы, авторы, академические департаменты, университеты, города, штаты, страны и даже научные инструменты, например телескопы [Abt, 1980]. Если предположить, что цитирование является показателем важности, можно использовать такой анализ для определения наиболее важных ученых, публикаций, департаментов и т. п. в той или иной дисциплине или субдисциплине. Данное предположение – всего лишь одно из многих заслуживающих более тщательного рассмотрения, если нужно понять (интерпретировать) результаты анализа цитирования.

Справка. Индекс цитирования научных статей − реферативная база данных научных публикаций, индексирующая ссылки, указанные в пристатейных списках этих публикаций, и предоставляющая количественные показатели этих ссылок (такие как суммарный объем цитирования, *h*-индекс и др.)

В середине XX в. факт цитирования стал базовым строительным блоком для совокупности индикаторов науки, представляющих ее с новой точки зрения. Однако такое представление по-

средством цитирования зависит от того, как осуществляется цитирование. Очевидно, что это определяется также распределением ссылок в исследуемом массиве научных публикаций. В свою очередь, создание индексов цитирования ведет к трансформации ссылки в соответствующее цитирование. Тем самым создается новое качество: “Все цитаты равны” [Smith, 1981], рассмотренное выше. Анализ цитирования в данном отношении напоминает стандартизацию [Latour, Woolgar, 1986]: “Ничто так не способствует созданию универсальности, как удаление локальных признаков объектов…”.

3.2. Что измеряется с помощью ссылок и цитирований?

Что измеряет цитирование? Один из вариантов ответа на этот вопрос предложен в монографии Х. Моеда [Moed, 2005]: «Цитирование представляет собой проявление базовых процессов, которые могут быть изучены с точки зрения различных дисциплин. Для того чтобы понять, на что указывают цитирования, и соотнести значения счетчиков цитирования с такими библиометрическими концепциями, как “эффективность исследования”(*Research performance*), “квалификация ученого” (*Scholarly quality*), “влияние” (*Influence*) или “важность” (*Impact*), нужно вникнуть в сущность этих процессов. Теоретическое понимание этих процессов вносит вклад в тот круг проблем, который часто называют “теорией цитирования”».

Ниже приведены пять подходов, на основании которых используются, интерпретируются или получают теоретическое обоснование показатели, основанные на цитировании: *физический*, *социологический*, *психологический*, *исторический* и *информационный* (или *коммуникационный*). Один подход может включать несколько различных установок “парадигм”. Следует отметить, что в одном исследовании может быть высказано более одной точки зрения.

3.2.1. *Физический подход*. Историк науки Д. Прайс являлся представителем такого подхода к количественному исследованию научной деятельности. Он утверждал, что имеющиеся показатели научной деятельности должны быть взаимосвязаны на основании простых законов, что позволит создать почву для их феноменологической интерпретации. “С определенной осторожностью можно предположить, что нам требуется общественно-научный эквивалент ньютоновского хода, когда были взяты столь неопределенно используемые термины вроде силы, работы и энергии […] и приведены в порядок с помощью простых уравнений” [Price, 1980]. Примером физического подхода к библиометрическим исследованиям также является работа [Van Raan, 1998].

3.2.2. *Социологический подход* дает ответ на вопрос о научной эффективности работ исходя из того, как ученые действуют, что они утверждают, что и как цитируют в своих публикациях. Данный подход характеризует восприятие действительности тем или иным ученым и относит заявления ученых, их публикации и практику ссылок к *социальным актам*. В рамках социологии науки укоренился четкий взгляд на науку как на вид общественной деятельности. В качестве иллюстрации можно сравнить теоретические работы [Merton, 1968] и [Callon, Law, Rip, 1986]. Если К. Мертон понимает ученого, в первую очередь, как “незаинтересованного искателя научной истины”, М. Каллон с соавторами утверждают, что ученый − это, в первую очередь, “предприниматель”, который работает в научно-исследовательском институте − предприятии по производству знаний.

3.2.3. *Психологический подход* сосредоточен на вопросах изучения психологии ссылок. Среди обсуждаемых ниже авторов стоит отметить работу Б. Кронина [Cronin, 1984], в которой особое внимание уделяется потенциальным возможностям психологического подхода к исследованию цитирования и ссылок. Типичным

примером исследований в рамках данного подхода является анализ мотивации цитирующего. Такой анализ часто имеет в основе материалы анкетирования ученых с целью получения их оценок по заранее определенному списку возможных мотивов. Пример исследования мотивации цитирующего содержится в работе [Brooks, 1986].

3.2.4. *Исторический подход*. В рамках исторических исследований может быть сделан акцент на когнитивные масштабы научной деятельности и историю развития научных идей, а также на личности, внесшие особый вклад. Таких примеров достаточно много. Например, работы по историографии Ю. Гарфилда явились мощным инструментом прослеживания развития научных идей с использованием анализа цитирования, например [Garfield, Pudovkin, Istomin, 2003]. Многолетний анализ структуры и развития исследований специальностей также может быть включен в этот тип исторических исследований, это, например, работа Г., в которой проведен анализ коцитирования [Small, 1977]. В свою очередь Д. Прайс в работах [Price, 1961], [Прайс, 1966] делает акцент на социальных, экономических и институциональных условиях, в которых проводятся исследования, как на уровне отдельных ученых или исследовательских подразделений, так и на уровне глобальной научной системы в целом.

3.2.5. *Коммуникационный подход*. К. Боргман определяет научную коммуникацию следующим образом: “Под научной коммуникацией мы понимаем исследование того, как ученые любой области (например, физических, биологических, гуманитарных наук) используют технологии и распространяют информацию по формальным и неформальным каналам. Исследование научной коммуникации включает развитие научной информации, связи между областями научных исследований и дисциплинами, информационные потребности и использование групп индиви-

дуальных пользователей, связи между формальными и неформальными методами коммуникации" [Borgman, 1990].

3.2.6. *Точки зрения*. В монографии [Moed, 2005] анализируются точки зрения отдельных ученых на проблему взаимосвязи между ссылками и цитированиями. Мы процитируем результаты проведенного Х. Моедом анализа теоретических позиций ряда ученых, исследовавших показатели, имеющие в основе цитирование.

Идея Ю. Гарфилда состоит в том, что работы, цитируемые в том или ином документе, символизируют конкретное содержание, например, метод, конкретную концепцию или гипотезу. Цитирующий документ демонстрирует отношение к цитируемой работе [Garfield, 1964]. С точки зрения Ю. Гарфилда списки литературы регулируются стандартами оптимальной научной экспозиции и практикой, рассматривающей хорошую библиографию как признак учености. Это делает их более целесообразными для целей индексации, чем, например, названия документов. Автор работы [Salton, 1963] также рассматривает ссылки как показатель содержания документа, который может быть использован как документальный дескриптор.

Поскольку авторы ссылаются на предыдущие материалы для поддержки, иллюстрации или развития тех или иных идей, акт цитирования является выражением важности материала. Вероятно, наиболее объективным критерием отражения важности материала для текущего исследования является общее количество таких выражений. Количество цитирований данного материала в некотором журнале в целом является в равной мере объективным и важным критерием качества журнала [Garfield, 1979].

Г. Смолл развивает идею Ю. Гарфилда о ссылке как предметном символе и вносит вклад в теоретические основы использования индекса цитирования. Он изучает вопрос о том, какого рода

“предметы” выделяются посредством ссылок на весьма высокоцитируемые документы. В традиционной науке цитируемые работы являются “первоисточниками”, которые авторы привлекают для придания дополнительного смысла своему тексту. Если посмотреть на ссылку под этим углом зрения, она представляет собой процесс маркировки. Формулировка, на которую указывает номер ссылки, маркирует или характеризует цитируемый документ. Цитируя тот или иной документ, автор создает его значение [Small, 1978], в результате цитируемый документ становится символом некоторой концепции, которая включает выводы, сделанные по результатам экспериментов, методологии, типы данных, метафизические представления, теоретические постулаты или уравнения. Такие документы, как правило, часто цитируются и называются “концептуальными символами”. Гипотеза Г. Смолла состоит в том, что “…ученый привносит с собой репертуар таких коллективных концепций и их соответствующих документальных символов”.

В одном из тезисов Р. Мертона утверждается, что движущей силой исследователя является стремление к славе и признанию. Однако научная система, в частности, система открытой публикации, организована таким образом, что ее институциональная цель − расширение научных знаний − и личные выгоды взаимосвязаны. Как и другие институты, наука имеет свою систему стимулирования за эффективность работы. Это стимулирование имеет характер, главным образом, приобретения авторитета, поскольку стремление к знаниям определяется, в первую очередь, как незаинтересованный поиск истины, и лишь во вторую очередь как способ заработать на жизнь [Merton, 1957].

Для большинства ученых главным поощрением за вклад в продвижение знаний является признание со стороны коллег. Ради получения такого поощрения ученые открыто публикуют резуль-

таты своих исследований, чтобы они могли быть использованы и признаны коллегами. В то же время они берут на себя обязательство признания источников знаний, положенных в основу их собственной работы. Последнее обязательство часто в обобщенном виде выражается в следующей формуле: “воздать должное там, где его нужно воздать”. Ученый не имеет выбора: право частной собственности здесь устанавливается путем ее раздачи, и для того, чтобы получить признание коллег, нужно раздать свою собственность другим.

Ссылка выполняет одновременно инструментальную и символическую функцию в передаче и расширении знаний. В инструментальном смысле она говорит о работе, с которой ранее мы могли быть незнакомы, и некоторые из таких работ могут представлять для нас дополнительный интерес. В символическом смысле она регистрирует в вечном архиве права интеллектуальной собственности признанного источника с помощью независимого подтверждения претензии на знание, принятое или прямо отвергнутое, которое было сделано в указанном источнике [Merton, 1996].

Представление о цитировании как показателе влияния и инструменте оценки научного вклада прямо вытекает из следующего заявления: “Если чья-то работа не будет замечена и использована другими в системе науки, могут возникнуть сомнения в ее ценности” [Merton, 1977].

Дж. и С. Коулы развивают концепцию анализа цитирования в качестве исследовательского инструмента при проведении социологических исследований. Они исходили из того, что цитирование может быть использовано как показатель влияния [Cole, Cole, 1967], [Cole, Cole, 1971]. Данное исследование выполнено на основе теоретических работ Р. Мертона, в частности, идеи интеллектуальной собственности и символической функции ссылки.

Главной целью Дж. и С. Коулов является разработка исследовательского инструмента, который можно было бы использовать в социологическом исследовании для проверки гипотез в области социальной стратификации и смежных с ней аспектов. В некоторых областях исследования авторы наблюдали положительную корреляцию между степенью цитирования отдельных статей или авторов и независимой оценкой и пришли к выводу о том, что анализ цитирования был полезен для данного случая. Авторы считают, что главной целью в социологических исследованиях является не изучение конкретного лица, но анализ связей между переменными.

Социологическая точка зрения, которую развивали Р. Мертон и его последователи, часто именуется “нормативной” точкой зрения или парадигмой [Cronin, 1984]. Поведение при организации ссылки в целом регулируется нормами. Из этого не следует, что все ученые во всех случаях строго выполняют эти нормы; речь идет лишь о том, что среди практиков существует общее представление о важности соблюдения этих норм для продвижения науки и знаний. В литературе воздается должное тогда, когда это необходимо. Тем самым подтверждается интеллектуальный долг общества перед первооткрывателем.

Вторая социологическая точка зрения может быть обозначена как социальная интерпретация ссылок. Эта конструктивистская точка зрения исходит из того, что ученые цитируют в своих интересах, защищая свои претензии от оппонентов, убеждая других, завоевывая тем самым доминирующее положение в научном сообществе. Например, Дж. Гильберт ввел в оборот идею, что ссылка помогает в попытке убедить. Для подтверждения выводов своего исследования авторы обычно цитируют документы, которые, как они полагают, их аудитория считает “авторитетными”.

Исследователи, работающие в некоторой зрелой области, придерживаются общей уверенности в том, что та или иная опубликованная работа является важной и правильной, другие работы являются тривиальными, некоторые, возможно, являются ошибочными, а большинство – неподходящими для текущих интересов. Поэтому авторы, как правило, цитируют “важные и правильные” статьи, могут цитировать “ошибочные” для того чтобы оспорить их, и избегают цитирования “тривиальных” и “нерелевантных” работ [Gilbert, 1977].

В прямой конфронтации с нормативной точкой зрения Дж. Гильберт заявляет: «В этой связи можно утверждать, что научная “норма”, предполагающая, что нужно цитировать исследования, от которых зависит некоторая работа, скорее всего, вызвана не доминирующим стремлением подтвердить “право собственности”, а желанием ученого убедить своих коллег, использовав для этого все имеющиеся ресурсы, включая упомянутые авторитетные статьи, которые могут цитироваться для подкрепления собственных аргументов» [Gilbert, 1977].

С этой точки зрения, цитирование является критерием авторитетности той или иной статьи, или, в более общем плане, ее риторической силы, которая определяется как степень, в которой цитируемая статья вписывается в риторику цитирующего автора.

Автор монографии “*Citation Process*” Б. Кронин утверждает, что для глубокого понимания процесса цитирования необходимо обратиться к психологии науки и проанализировать взаимодействие между “институциональными нормами и личными соображениями”. «Цитирование следует воспринимать как процесс создания научной статьи… Результатом этого процесса является список ссылок … Характер и состав списков отражают личность и профессиональную среду, в которой работают авторы. Элементы в “химии” цитирования являются практически бесконечными, и

именно этот факт делает необходимым конкретный учет цитирования» [Cronin, 1984]. Автор отвергает “мертонианское” представление, которое предполагает, что в основе цитирования лежит “признание конкретного и универсально признанного множества норм”, но, в то же время, процесс цитирования, как он полагает, “не характеризуется случайностью и противоречивостью”.

Б. Мартин и Дж. Ирвин (*Ben Martin, John Irvine*) проводят различие между качеством исследования, важностью и влиянием научной публикации. Качество – свойство публикации и исследования, которое в ней описано. Под важностью понимается потенциальное влияние исследования на деятельность, осуществляемую вокруг этого исследования, как если бы в науке существовала “идеальная коммуникация”. Влияние определяется как реальное воздействие. Эти авторы в основном соглашаются с представлением Коулов о цитировании как “несовершенном” критерии влияния. Однако “…акцент на количество цитирований необходим для продвижения по службе, расширения штата сотрудников, получения грантов и приобретения известности авторами, их предыдущими работами и учреждениями, в которых они работают [Martin, Irvine, 1983].

С. Козенс (*Susan Cozzens*) указывает на различие между системой поощрения и риторической системой в науке. Первая описана в работах Р. Мертона и является воплощением этикета цитирования, который предполагает, что при использовании опубликованной новой идеи необходимо эксплицитно процитировать ее источник. Это обусловлено стратегическим использованием ссылок в качестве оружия для защиты претензий на знание и убеждения коллег в значимости их собственных исследований. С другой стороны, в работе [Cozzens, 1989] утверждается, что “…цитирование, в первую очередь, является частью текста, которому требуется поддержка”. Иными словами, риторическая

система является доминирующей. Поощрительная система с ее этикетом цитирования является лишь вторичным критерием выбора ссылок. Другими важными элементами этой системы являются прямая похвала со стороны коллег, продвижение по службе, получение грантов и премий. С. Козенс высказывает гипотезу о том, что цитирования могут быть распределены между поощрительной и риторической системами. “Ясно, что главной функцией того или иного документа является убедительная защита претензий на знание, что искусство написания научных статей состоит в приведении в определенный порядок имеющихся риторических ресурсов для достижения этой цели” [Cozzens, 1989].

Г. Цукерман в ответ на тезис Дж. Гильберта, согласно которому ссылка является средством убеждения, защищает позицию, предполагающую, что цитирование является критерием интеллектуального влияния. Она утверждает, что даже если известная работа знаменитого ученого цитируется для целей убеждения, такое цитирование может отражать когнитивное влияние, и подчеркивает, что мотивы цитирующих авторов и их последствия, отражающие влияние, различаются в аналитическом смысле. Г. Цукерман формулирует вопрос: “Если предположить, что авторы обычно цитируют важные или авторитетные статьи, стремясь быть убедительными, что же делает такие цитируемые работы важными или авторитетными?” – и сама отвечает на него: “…признание коллегами когнитивной ценности источников, которые стали влиятельными, благодаря высокой степени цитирования” [Zuckerman, 1987].

Г. Уайт (*Howard White*) изучал работы двух групп аналитиков, придерживающихся различных подходов к исследованию ссылок и цитирования. Первый подход он называет “библиографическим”, выделяя конкретные случаи и индивидуальные особенности или “пристрастия”. Второй связан с поиском абстрактных моделей высоко агрегированных данных.

Поскольку эти группы работают на разных уровнях реальности, разрыв между ними действительно не может быть преодолен. «Это то же самое, как если бы каждая из них исследовала один и тот же город, но по-своему: одна, живя в нем, другая, пролетая над ним на самолете. Вид “с земли” и “с воздуха” дает в результате описания реальности, которые в некоторой точке становятся несоизмеримыми» [White, 1990].

Г. Уайт делает акцент на описании потенциальных возможностей анализа больших файлов данных. Он утверждает, что карты коцитирования или анализ цитирования в целом представляют собой инструменты, с помощью которых создается интеллектуальная история. Он полагает, что эти инструменты представляют “историю консенсуса относительно важности авторов или работ”. Такой консенсус не является эксплицитным в том смысле, что практики в той или иной области сознательно заявили о своем согласии с этим; это, скорее, имплицитный консенсус, который автор квалифицирует как “социальный конструкт вроде общественного мнения” [White, 1990].

Г. Уайт считает, что карты, построенные на основе анализа коцитирования, являются “объективными” в том смысле, что они основываются на алгоритмах, работающих на крупных файлах данных, что они воспроизводимы и что даваемая ими картина “шире, чем та, которая может быть получена любым отдельным ученым”. В то же время автор подчеркивает, что «…они (ученые) не претендуют на то, что их методы делают субъективное суждение ненужным или лишают силы традиционные “предметные” методы интервьюирования и ознакомления с первоисточниками» [White, 1990].

А. Ван-Раан (Anthony Van Raan) говорит, что анализ поведения авторов при организации ссылки не является наиболее целесообразным способом исследования обоснованности анализа

цитирования. «Это то же самое, как если бы физик пытался определить рамки термодинамики, создавая “теорию” о поведении отдельных молекул» [Van Raan, 1998].

А. Ван-Раан утверждает, что наибольшее количество цитируемых статей в списке литературы является типично “модальными” статьями, которые практически не привлекают цитирование, а небольшая часть представляет собой высокоцитируемые статьи. Поэтому анализ поведения при цитировании имеет дело, главным образом, с большими количествами цитирований модальных статей, тогда как для анализа цитирования в отношении эффективности исследования наиболее важными являются относительно небольшое количество высокоцитируемых статей.

А. Ван-Раан предлагает “термодинамическую” теорию цитирования, для которой важны не отдельные цитирующие, а “ансамбль” цитирующих, к которому применим статистический подход в понимании функции распределения “поведенческих характеристик” цитирующих авторов. Давление, объем и температура являются главными параметрами термодинамического ансамбля, включающего огромное множество молекул. Аналогичным образом, анализ цитирования находится на “термодинамической стороне”: он касается ансамбля, включающего множество цитирующих. “Разумеется, индивидуальные характеристики цитирующих интересны, но функции распределения этих характеристик составляют часть мира, которая релевантна для библиометрического анализа” [van Raan, 1998].

П. Воутерс (*Paul Wouters*) проводит разграничение между двумя теоретическими концепциями информации. С одной стороны, в определении К. Шеннона, информация понимается как формальная сущность, из которой происходят все значения. С другой точки зрения, по определению Г. Бейтсона (*Gregory Bateson*), информацией является “любая разница, которая делает разницу”.

П. Воутерс называет эту концепцию “парадигматической” [Wouters, 1999b]. В исследовании научной коммуникации развиваются обе теоретические концепции информации. Необходимо отметить, что информационный подход заимствует элементы физического и социологического подходов. Например, концепция информации может быть смоделирована аналогично физической концепции энтропии, а отношения цитирования или отношения между авторами могут быть использованы для выявления социальных структур.

П. Воутерс внес вклад в развитие рефлексивной теории индикаторов, подчеркнув релевантность различия между ссылками и цитированием. Он полагает, что в поисках всеобъемлющей теории цитирования социологические исследования поведения при ссылке ведут в “тупик”. Эти исследования вносят вклад, скорее, в теорию ссылки, чем в теорию цитирования.

Х. Моед в монографии [Moed, 2005] резюмировал теоретические позиции ученых, которые внесли вклад в ответ на вопрос “Что измеряется с помощью ссылок и цитирований?”(табл. 3.2.1).

Таблица 3.2.1

Точки зрения на то, что измеряется ссылками и цитированием

Авторы	**Что такое ссылки**	**Что измеряет цитирование**
[Garfield, 1964], [Salton, 1963]	Ссылки являются дескрипторами содержания документа	
[Garfield, 1979]	Ссылки отражают научные и информационные потоки	Оценка полезности, важности, влияния, зависящая от частоты, с которой цитируется документ
[Small, 1978]	Цитирование порождает авторскую интерпретацию цитируемого документа, т. е. участвует в процессе создания концептуальных символов	Цитируемые документы являются обозначением концепции, включающей экспериментальные и теоретические разработки

[Merton, 1957], [Merton, 1977], [Merton, 1996], [Zuckerman, 1987]	Ссылка регистрирует право интеллектуальной собственности и независимое признание со стороны коллег	Интеллектуальное влияние
[Cole, Cole, 1967], [Cole, Cole, 1971].		Социально установленное качество
[Gilbert, 1977]	Является инструментом убеждения	Авторитетность, или в более общем смысле риторическую силу, определяемую как степень, до которой цитируемая работа соответствует риторике цитирующей работы
[Cronin, 1984]	Характер и состав списка ссылок отражает авторскую индивидуальность и профессиональную среду	Неясно, что измеряет цитирование; следует изучить соотношение норм, принятых в научной среде, и собственных соображений автора
[Martin, 1983]	Темп цитирования только частично указывает на важность публикации, на него влияют такие факторы как социальное и политическое давление (коммуникационная практика), известность цитируемого автора и места его работы	Для правильного использования и интерпретации темпа цитирования следует правильно отбирать группы, в рамках которых производится сравнение, и дополнительно использовать альтернативные индикаторы производительности
[Zuckerman, 1987]	Мотивы цитирующего автора и их последствия аналитически различны	Цитирования могут служить “посредниками” для прямых мер интеллектуального влияния, таких как суждения коллег и присуждение наград

[Cozzens, 1989]	Ссылки находятся в точке пересечения поощрения, риторики и коммуникационной среды, но первоочередную роль играют риторические факторы	Каждая мотивация генерирует некоторую долю счетчика цитирования и должна быть принята во внимание
[White, 1990]	Текстуальное родство отражает прямое признание соответствующего документа	Карты коцитирований и анализ цитирования являются мерой исторического консенсуса относительно важности авторов и работ
[Van Raan, 1998]	Ссылки до некоторой степени носят частный характер, но в больших совокупностях возможные “перекосы” нейтрализуются	Вершина распределения “термодинамического” набора из многих цитирующих указывает на “высшую” группу исследований
[Wouters, 1999b]	Ссылки являются продуктом деятельности исследователя и принадлежат цитирующему тексту. После индексации они становятся атрибутами цитируемого текста, таким образом, между ними возникает существенная разница	Цитирования – это продукт деятельности того, кто осуществляет индексацию. Валидация индексов цитирования не может основываться исключительно на поведенческом характере ссылок

На основании сказанного выше следует констатировать, что дискуссия на тему “Что измеряется ссылками и цитированием?”, по существу, остается открытой.

На вопрос о смысле цитирования лучший, на наш взгляд, вариант ответа принадлежит группе сотрудников *Thomson Reuters*, который мы приводим полностью: “Идея, стоящая за индексаци-

ей цитирований, чрезвычайно проста. Если признать, что ценность информации определяется теми, кто ее использует, что может быть лучше для оценки качества работы, чем измерение влияния, которое она производит на сообщество в целом. В этом случае наибольшее количество членов научного сообщества (т. е. тех, кто использует или цитирует исходный материал) определяет влияние или взаимодействие идей и на ее автора, и на объем наших знаний" [Adler, Ewing, Taylor, 2009].

3.3. Продуцирование цитат

Вопросы продуцирования цитат изучались в монографии [Egghe, Rousseau, 1990]. Для определения этого процесса была использована аллегорическая цепь изготовления товара: "сырье $\rightarrow$ производство $\rightarrow$ изделие". Применительно к анализу цитирования "сырьем" (*Sources*) в этой цепи являются статьи, а "изделиями" (*Items*) – ссылки или цитирования. Под термином "производство" фигурирует процесс производства информации (*Information Production Processes, IPP*).

Приведем формальное определение. *IPP* – это тройка вида (S, I, V), где $S = [0, T]$ – замкнутый интервал от 0 до T, $I = [0, A]$ – замкнутый интервал от 0 до A, V – строго возрастающая дифференцируемая функция, V: $S \rightarrow I$, такая что $V(0)=0$, $V(T)=A$. Элементы *Sources* представлены множеством S, а элементы *Items* – множеством I.

Свойства *IPP*. Во-первых, известно, что *IPP* является хорошим средством для исследования комплексности в терминах фрактальной теории. Сошлемся на монографию [Мандельброт, 2002] для обоснования того, что $1/\beta$, где β – экспонента Мандельброта, является фрактальным параметром *IPP*.

Очевидно, что *IPP* цитирования показывает (значительно) больше комплексности в смысле фрактальных параметров, чем

“ординарные” *IPP*, например библиография. Действительно, в последнем случае источниками являются журналы, а изделиями – статьи; в первом случае начинают с различных статей в качестве источников и исследуют ссылки или цитирование в качестве изделий. В случае если исследуются такие понятия, как коцитирование или библиографическое сочетание, комплексность еще более увеличивается (в интуитивном контексте: в квадратичной форме), что также отражается в значительном увеличении фрактального размера. Следовательно, читатель должен быть убежден в важности фрактальной теории цитирования. Эту идею поддерживает [Van Raan, 1991].

IPP цитирования может быть также отграничено от библиографии в том смысле, что “изделия” могут иметь более одного источника. Действительно, некоторая статья может использоваться несколько раз в качестве ссылки (цитирующие статьи (источники) в этой связи сочетаются в библиографическом отношении) и таким же образом в случае (двойственном) коцитирования. В данном случае *IPP* цитирования сопоставимо с *IPP*, когда авторы рассматриваются в качестве “источников”, а публикации – в качестве “изделия”: здесь также одно изделие может иметь несколько источников. В работе [Egghe, 1994] эти более сложные *IPP*, в которых изделия могут иметь несколько источников, исследуются с использованием “простых” *IPP* (изделий, происходящих из одного источника) и функций свертки с целью получения базовых распределений комплексного *IPP*.

Справка. Свертка функций − математическая операция двух функций f и g, порождающая третью функцию, которая обычно может рассматриваться как модифицированная версия одной из первоначальных.

С середины 1960-х гг. XX в. на основе ББД *ISI* начался процесс обращения и ссылок в цитаты и, далее, в библиометрические индикаторы, которые, по предположениям, отражают реальность

науки и определяют какой-либо ее феномен. Можно сказать, что в 70-80-е гг. появились первые научные индикаторы. Например, стало ясно, что шанс быть процитированным в значительной степени зависит от научной специальности, ее масштабов и культуры цитирования. В этой связи были определены несколько типов "частот нормализованного цитирования". Индекс цитат был организован на различных уровнях агрегирования, от индивидуального исследователя до стран. Перед тем как приступить к анализу цитат, необходимо рассмотреть вопросы, лежащие в основе процесса продуцирования цитат как такового.

Ранее мы говорили о том, что ссылки и цитирования следует разграничивать и что ссылка рождает цитату. Необходимо помнить, что невозможно обработать каждую ссылку, сделанную в каждой научной статье в мире. Семиозис цитирования является точной операцией – ссылки и цитирования должны соответствовать друг другу, здесь важна не идентичность, а достоверная инверсия. Известно, что процесс индексирования неизбежно содержит "ошибки", отсюда любой анализ цитирования, связанный с обработкой больших объемов данных цитирования, содержит значительное количество "ошибок", которые сложно идентифицировать. Вследствие этого весьма актуальна проблема качества данных, в рамках которой изучается выбор ссылки и целостность инверсии.

3.3.1. *Качество ссылки.* Автор может сослаться на несуществующий текст, просто указав неправильный номер страницы или год либо допустив ошибку в спеллинге. В таких случаях составитель индекса цитирования может создать цитату как атрибут несуществующего текста. Акт создания цитаты рефлексивен по отношению к приведению ссылки. В работе [Moed, Vriens, 1989] сообщается, что примерно для каждых десяти цитат, одно цитирование содержит те или иные ошибки, большая часть которых относится к привычкам цитирующих авторов "копировать ссыл-

ки из других статей". Другой важной проблемой является спеллинг названий и имен авторов в журнале.

3.3.2. *Выбор ссылок* предполагает принятие решения о том, какого типа текстуальные ссылки следует обрабатывать как ссылки, а также о том, какие журналы следует использовать как источники. По мнению авторов [Hicks, Potter 1991], *SCI* не дает правильного представления о текстах, которые повлияли на ученых: "…каждый текст следует рассматривать своего рода непризнанной ссылкой", однако "…используется лишь тонкий срез потенциальных ссылок". И далее: "…цитирование образует лишь тонкий, но яркий слой, зажатый между пустой породой. И именно этот в значительной степени ограниченный, нерепрезентативный, лежащий на самой поверхности слой *ISI* фетишизирует и выдает за весьма желанный и ходовой товар".

С точки зрения проводимого исследования эту "геологическую" метафору необходимо интерпретировать несколько иным образом. Фактически авторы говорят не столько о "геологии цитирования" как таковой, сколько о "геологии его сырья" – ссылок. Тот факт, что *ISI* извлекает только часть источника потенциальной ссылки, не простая случайность, а следствие ограничений, налагаемых экономическими особенностями продуцирования цитат. В настоящее время данная метафора особенно актуальна, поскольку "добыча" (*Data mining*) цитат является очень дорогостоящим делом, требующим масштабных трудозатрат обработки значительных объемов научной информации.

Соблюдение *целостности инверсии* в процессе продуцирования цитат может создавать дополнительные проблемы, большинство из которых возникает вследствие неопределенности, присущей этому процессу. В работах [Smith, 1981] и [Egghe, Rousseau, 1990] сформулированы проблемы, связанные с инверсией, например, регистрация только первых авторов цитируемых текстов, употребление идентичных имен для различных организаций и т. п.

Очевидно, что для конвертации в цитату ссылка должна существовать в легко распознаваемом формате. Массовое продуцирование цитат зависит от стандартизации. Не случайно Д. Прайс, Ю. Гарфилд и их коллеги часто выступали за изменение политики научных журналов в отношении ссылок. В частности, эти ученые пытались убедить редакторов в растущей важности формата ссылки: “Теперь, когда определение индекса цитирования стало важной и неотъемлемой частью автоматизированных систем, с помощью которых мы получаем доступ как к архивным публикациям, так и к литературе, образующей фронт научных исследований, может потребоваться пересмотр редакторской практики в отношении цитирования” [Price, 1969]. Это в интересах не только исследователей, но и составителей индексов. Чем более стандартизированными являются ссылки, тем менее дорогостоящим станет продуцирование цитат.

Еще одним аспектом выбора сырья для продуцирования цитат является формирование массива научной периодики. Конечно, для обеспечения полноты поиска ББД должна содержать некоторую репрезентативную выборку литературы, однако на деле справедливо обратное. Распределение характеристик в библиометрии часто является асимметричным в соответствии с законом Бредфорда [Бредихин, Кузнецов, 2012], согласно которому практически все документы “высокого” качества находятся в ограниченном количестве ресурсов, т. е. журналов по естественным наукам.

3.4. Статистические характеристики

Одной из первых работ (если не первой!) по рассматриваемой теме следует считать статью историка науки Д. Прайса [Price, 1965], которая посвящена обоснованию тезиса: “…система библиографических ссылок позволяет вскрыть природу научных исследований в целом”. Автор предпринимает попытку воспроизвести систему научных работ, связывая каждую опубликован-

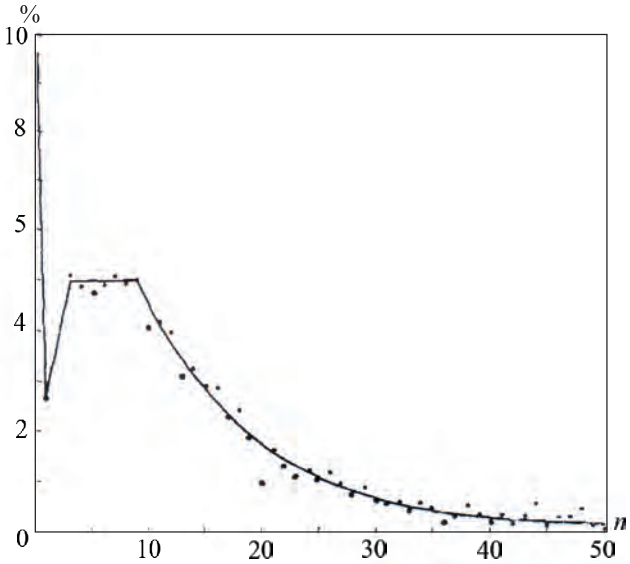

Рис. 3.4.1. Распределение научных работ
по числу библиографических ссылок

ную работу со всеми другими работами, имеющими к ней непосредственное отношение. Учитываются характерные связи, устанавливаемые цитированием одной работы в других работах. Представим общие характеристики этой системы в виде ряда эмпирических фактов, конструируемых на основе интерпретации исходных данных. (Здесь и далее в этом разделе статистические данные и рисунки заимствованы из работы [Price, 1965], которые, в свою очередь, были опубликованы Ю. Гарфилдом в его работах 1961 г.). Несмотря на то что эти данные, взятые из недавнего прошлого, могли устареть, они хорошо иллюстрируют определения, оценки и методы анализа, используемого Д. Прайсом.

Первая характеристика – общее число ссылок. В среднем на каждую научную работу приходится 15 ссылок. Распределение научных работ по числу содержащихся в них ссылок показано на рис. 3.4.1.

Установлено, что примерно 10 % статей не содержит ссылок вообще; вместе с тем 50 % ссылок относится к 85 % работ, которые представляют собой работы “нормального” типа и содержат примерно 25 ссылок каждая. Распределение в этой области представляет собой хорошо выраженное плато. Действительно, приблизительно 5 % статей попадает в любую группу, которая содержит от 3 до 10 ссылок. На другой конец оси абсцисс попадают обзорные статьи, содержащие многочисленные ссылки. Примерно 25 % всех ссылок приходится на 5 % статей (общего числа), содержащих 45 и более ссылок каждая, а в среднем 74 ссылки на статью. Наконец, 12 % ссылок в этой категории падает на “тяжелый вес”, составляющий 1 % (общего числа статей); статьи этого типа содержат 84 и более ссылок каждая; среднее число ссылок в этой категории равно 170 на статью. Заметим, что число статей с n ссылками в категории “тяжелого хвоста” убывает по закону $1/n^2$.

В течение уже нескольких веков объем издаваемой научной литературы ежегодно увеличивается по экспоненциальному закону [Прайс, 1966] и, по-видимому, будет продолжать возрастать с такой же скоростью. На основании собственных оценок Д. Прайс делает первый вывод: “усредняя за достаточно длительный период по всей мировой научной периодике, мы обнаружим, что каждая научная работа, независимо от того, когда она опубликована, будет процитирована примерно один раз в год”.

Вторая характеристика – распределение числа ссылок на статью или число цитирований (рис. 3.4.2). В среднем, за произвольно взятый год 35 % всех опубликованных статей не цитируются вовсе, а 49 % статей цитируются один раз (n=1). Далее распределение имеет следующий вид: 9 % статей цитируется два раза; 3 % – три; 2 % – четыре; 1 % – пять; оставшиеся статьи цитируются

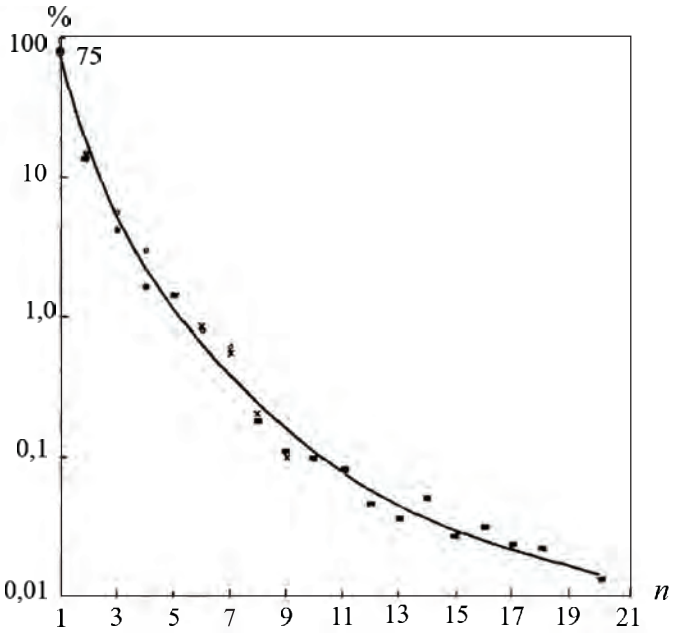

Рис. 3.4.2. Распределение числа цитирований на одну статью

по шесть и более раз. Итак, для больших n число цитирований статей убывает по закону ~ n^3.

Далее Д. Прайс формулирует две гипотезы: 1) ежегодно приблизительно 10 % научных статей “умирает”, поскольку никто их больше не цитирует; 2) вероятность того, что любая “живущая” статья будет как минимум один раз процитирована в каком-либо году, составляет 60 %.

Допуская, что данные гипотезы верны, Д. Прайс указывает направление исследований: «…имеется настоятельная необходимость выяснить, какова вероятность того, что статья, часто цитируемая в настоящий момент, будет достаточно часто цитироваться и в будущем” и мечтает “…дальнейшая работа в этой области может привести к открытию, позволяющему выявлять “классические” работы, которые окажутся столь отчетливо различимы, что

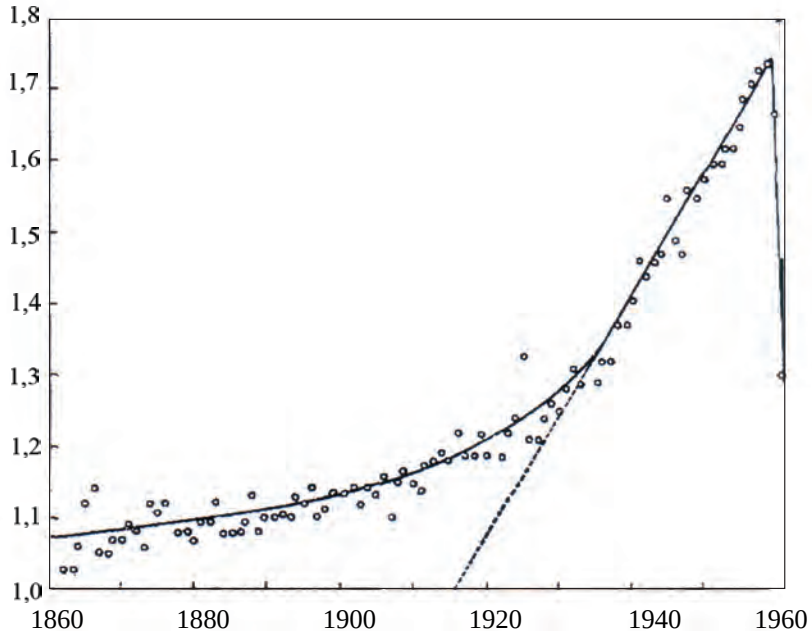

Рис. 3.4.3. Распределение цитат по времени

их можно будет автоматически выделять по областям науки с помощью некой процедуры …таким образом, появится возможность издавать мировой журнал “реально нужных” статей». Д. Прайс высказывает сожаление: “…мы еще очень мало знаем о связи – если она вообще существует, между числом, которое указывает, сколько раз была процитирована данная статья, и числом библиографических ссылок, которые эта статья содержит”, что также можно рассматривать как указание на нерешенную задачу.

Третья характеристика – “фактор оперативности” (*Immediacy factor*), дающая числовые оценки часто цитируемым работам, на основе которых можно проводить их сравнение. Основная тенденция, заключающаяся в том, что наиболее цитируемые работы являются и наиболее свежими, показана на рис. 3.4.3, где количе-

ство цитирований на публикацию рассматривается как функция от возраста работы. Видно, что приблизительно 70 % всех цитируемых работ относится к нормальной кривой роста, дающей удвоение каждые 13,5 лет, и что приблизительно 30 % следует отнести к пикам “оперативности”. Автор делает вывод: “…можно сказать, что упомянутые 70 % работ отражают случайное распределение цитирований всех научных работ, когда-либо опубликованных, а год публикации не имеет существенного значения. Однако 30 % цитированных работ представляют собой в высшей степени избранные ссылки на научную периодику, половина которых приходится на работы с “возрастом” от одного года до шести лет”. Если данная пропорция сохраняется для всего массива научной периодики, то указанные выше 30 % работ можно отнести к “фронту исследований”. Рассмотрим отношение чисел работ, цитированных в 1961 г., к числам работ, опубликованных с 1860 по 1961 гг. Функция распределения представлена на рис. 3.4.3.

На рис. 3.4.3 приведено отношение количества всех цитирований 1961 г. к количеству цитируемых работ, опубликованных в конкретном году, за период с 1860 по 1960 гг. Отношение является мерой кратности цитирований; на рис. 3.4.3 видно, что кратность резко уменьшается с годами, эта мера может рассматриваться также как степень доступности цитируемых работ. Наблюдается резкий спад отношения в зависимости от времени, что дает основание разделить научную периодику на две категории с существенно различающимися временами жизни – классическую и эфемерную. В работе [Barton, Kebler, 1960], основанной, по мнению Д. Прайса, “…на достаточно скудных данных”, обнаружено в частности, что время жизни, а следовательно, относительные пропорции классической и эфемерной литературы, значительно меняются от одной области науки к другой. Например, математика, геология и ботаника являются в высшей степени

классическими; инженерная деятельность в области химии, механики и металлургии, а также физика являются в высшей степени эфемерными; химия и физиология представляют собой более или менее однородную смесь классики и эфемерности.

Четвертая характеристика относится к возрасту ссылок, она определена в работе [Прайс, 1971] и называется индексом Прайса. Научная публикация, помимо сообщения научной информации, является выражением существующего в данный момент в науке положения ученого или группы ученых. Именно поэтому библиографические ссылки, так же как состав и порядок соавторов, выступают в качестве указателей на социальные связи. Удобным показателем социальной связи оказывается также общее количество библиографических ссылок в статье и их возраст. В работе “О сетях цитирования” автор, используя данные *SCI*, показал, что существует две популяции ссылок, частично перекрывающие друг друга. С одной стороны, это ссылки на “архивную” литературу, распределенные достаточно равномерно, кроме того их количество медленно уменьшается по мере старения литературы. С другой стороны, это ссылки на литературу “оперативного воздействия”, сравнительно современную и относящуюся к “фронту исследований”. Важной характеристикой автор считает отношение количества “оперативных” ссылок к количеству “архивных” ссылок. Сущность индекса Д. Прайс формулирует следующим образом: “… вполне разумным будет взять в качестве индекса отношение количества ссылок на литературу не более чем пятилетней давности к общему количеству ссылок” (выбор пятилетнего периода ученым также обоснован).

Справка. Информация (от лат. *Informatio* − осведомление) − есть выбор одного (или нескольких) сигналов, параметров, вариантов, альтернатив и т. п. из многих возможных, и этот выбор должен быть запомнен. В теории информации разработаны универсальные

математические (статистические) методы измерения информации, которые совершенно не зависят от способов ее передачи, типов материальных носителей и формы сигналов в каналах связи, а также от конкретного содержания передаваемых сообщений. С теоретико-информационной точки зрения это некое идеальное сообщение, уменьшающее или полностью исключающее неопределенность в выборе одной из нескольких возможных альтернатив [ЭЭиФН, 2009]. В упрощенном виде − это значимые сведения о чем-либо, когда форма их представления также является информацией.

Завершим этот параграф цитатой из работы [Прайс, 1966]: "…цитирование образует сеть, связывающую все работы в единый комплекс. Каждая статья возникает на фундаменте других статей и сама, в свою очередь, становится одним из отправных моментов для следующей. Указание на источник − наиболее яркое проявление этого научного способа кирпичной кладки".

3.5. Элементарные расчеты на основе цитирования

На основе "сырых" данных о цитировании можно построить простые индикаторы. В настоящее время их достаточно, но ни один из них не обладает достаточной полнотой. Однако все они обладают общими чертами, утверждает П. Воутерс в работе [Wouters, 1999b]: "Во-первых, в основе всех индикаторов лежит меняющееся сочетание параметров – ссылок и цитирования, способ комбинирования которых определяет важные результирующие характеристики. Во-вторых, индикаторы направлены на выявление "другой" реальности, отличной от той, которая была представлена конкурирующими индикаторами. В-третьих, все индикаторы строятся на взаимной основе".

3.5.1. Простейшим библиометрическим индикатором является *индекс Прайса*.

$$PI = \left(N_1 / N_2 \right) \times 100 ,$$

где N_1 – количество ссылок на работы, которые были опубликованы менее чем за шесть лет до цитирующей публикации; N_2 – общее количество ссылок.

Иными словами, индекс Прайса, который представляет собой количество ссылок за последние пять лет в виде процента от общего количества ссылок, отражает некий “критерий новизны” литературы, цитируемой в данной статье, журнале или специальности [Moed, 1989]. Индекс Прайса может быть рассчитан для данного года [Price, 1970]. Несмотря на простоту этого индекса, его интерпретация представляется весьма расплывчатой.

3.5.2. *Импакт-фактор журнала*. Этот индикатор, введенный в оборот Ю. Гарфилдом, вычисляется по формуле

$$IF = C / N,$$

где C – количество цитирований, которое получил тот или иной журнал на работы, опубликованные за определенный период времени; N – общее количество публикаций в журнале за тот же период. Для значений *IF*, опубликованных *WoS*, выбран период два года. Импакт-фактор был задуман как индикатор, обеспечивающий возможность сравнения журналов по вероятности их цитирования, и является примером нормализации журналов по частоте. Конечно, частота цитирования является функцией многих переменных, помимо научных достоинств журнала; например, она зависит от количества статей, опубликованных в журнале, и от их объема.

Значения некоторых элементарных индикаторов вычисляются на регулярной основе владельцами ББД на своих информационных массивах. В РФ в качестве источников библиографической информации, в том числе данных о цитировании, достаточно широко используются ББД: *WoS*, *Scopus*, *Medline* и РИНЦ.

3.6. Модель ветвления цитирования

Начиная с 60-х годов XX в., библиометрия проявляет интерес к сравнительным исследованиям популярности научных статей (см., например, [Garfield, 1979]), однако область этого интереса ограничена подсчетом цитирований. В 2003 г. М. Симкин и В. Ройчодхури опубликовали статью под названием “Прочти, потом цитируй!” [Simkin, Roychowdhury, 2003], в которой говорится о том, что некоторые авторы научных публикаций цитируют работы путем копирования ссылок из других работ, т. е. цитируют работы, которые они не читали. Авторы предложили метод оценки доли цитирующих статью людей, действительно прочитавших ее. Важно, что это можно определить без тестирования ученых, исключительно на основе информации, имеющейся в базе данных цитирования *WoS*.

Изучая проблему распределения цитирований научных публикаций, М. Симкин и В. Ройчодхури разработали модель “случайно цитирующих ученых” или “ветвления цитирования” [Simkin, Roychowdhury, 2007a], которая описывает распределения ссылок на научные работы по годам и авторам. Несмотря на простоту, модель оказалась способной учесть ряд свойств эмпирически наблюдаемых распределений цитирования, в том числе эффект накопленного преимущества (иначе, эффект Матфея), при котором частота цитирования конкретной статьи пропорциональна числу цитирований, которые она уже получила. С ее помощью можно также объяснить эффект “отложенного спроса” на научные идеи, т. е. резкий рост ссылок на некоторые научные публикации после долгих лет забвения. Основным аналитическим аппаратом этой модели является теория ветвящихся процессов.

При создании научной публикации ученый, как правило, читает последние выпуски научных журналов и выбирает из них ссылки для цитирования. В основе этой модели лежат два пред-

положения: а) автор ссылается на свежую статью, которую он только что прочел, для встраивания своей работы в современные тенденции; б) автор ссылается на старые статьи, цитированные в свежих статьях, для определения места своей работы в контексте прежних достижений. Авторы модели считают, что разумная оценка времени работы ученого над конкретной статьей соответствует одному году. Поэтому следует принять, что “недавний” в этой модели означает предыдущий год. Для достижения математической разрешимости модели введем дискретизацию времени с единицей в один год.

Ежегодно публикуются N статей. В среднем на опубликованную статью приходятся N_{ref} ссылок. Ежегодно доля ссылок α приходится на случайно выбранные статьи предыдущего года; оценка цитирования по реальным данным дает $\alpha \approx 0{,}1$ [Price, 1965]. Остальные цитирования случайным образом копируются из списка ссылок, использованных в статьях за предыдущий год.

Когда N велико, эта модель приводит к распределению Пуассона для цитирований за первый год. Вероятность получения n цитирований равна

$$p(n) = \frac{\lambda_0^n}{n!} \mathrm{e}^{-\lambda_0}, \qquad (3.6.1)$$

где λ_0 – среднее ожидаемое число цитирований – вычисляется по формуле

$$\lambda_0 = \alpha N_{ref} \qquad (3.6.2)$$

Число цитирований второго года, обусловленных каждым цитированием первого года (как и цитирования третьего года, обусловленные цитированием второго года и т. д.), также отвечают распределению Пуассона, но на этот раз со средним значением

$$\lambda = (1 - \alpha) \qquad (3.6.3)$$

В этой модели представлен ветвящийся процесс с цитированием первого года, соответствующим “детям”, цитированием второго года, соответствующим “внукам”, и т. д.

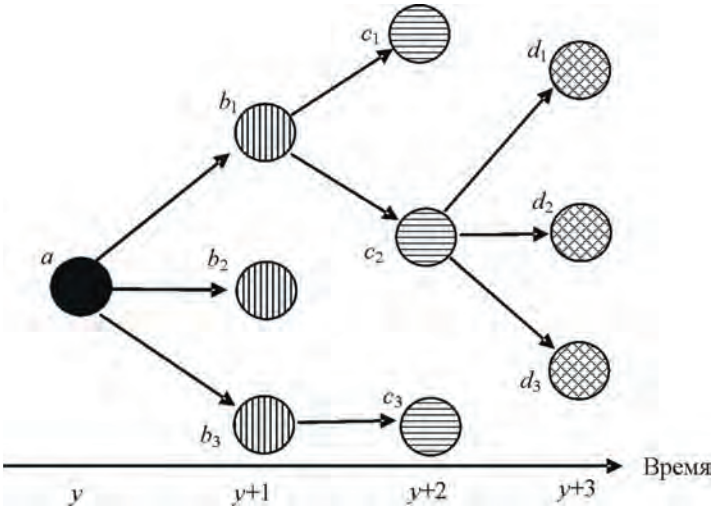

Рис. 3.6.1. Ветвление цитирования

На рис. 3.6.1 показан ветвящийся процесс цитирования, в котором в течение первого года после публикации статья a цитировалась в трех других статьях b_1, b_2, b_3, написанных учеными, которые прочли a. В течение второго года одно из этих цитирований было скопировано в двух статьях c_1 и c_2, одно цитирование было скопировано в одной статье c_3, а одно вообще не копировалось. Это приводит к трем цитированиям второго года. В течение третьего года цитирования c_1, c_3 вообще не копировали, а c_2 было скопировано в трех статьях d_1, d_2, d_3.

Подставляя уравнение (3.6.1) в уравнение производящей функции [Harris, 1963]

$$f(z) = \sum_{n=0}^{\infty} p(n) z^n \qquad (3.6.4)$$

получаем производящую функцию для цитирования первого года

$$f_0(z) = \mathrm{e}^{(z-1)\lambda_0} \qquad (3.6.5)$$

Аналогично, производящая функция для последующих лет имеет вид

$$f(z) = \mathrm{e}^{(z-1)\lambda} \qquad (3.6.6)$$

Процесс легче анализировать, приняв $\lambda = \lambda_0$ или $\lambda_0 / \lambda = N_{ref}\alpha / (1-\alpha) = 1$, поскольку при этом получается простой ветвящийся процесс, в котором все поколения определяются одной и той же вероятностью появления потомков.

На основании этой модели можно построить функции распределения цитирований на статьи, опубликованные в том же году, и статьи, цитированные в том же году.

Справка [Мат. энц.]. Ветвящийся процесс (далее ВП) – это случайный процесс, описывающий широкий круг явлений, связанных с размножением и превращением каких-либо объектов (например, частиц в физике, молекул в химии и т. п.). Основным математическим предположением, выделяющим класс ВП, является предположение независимости размножения частиц друг от друга. Однородный во времени ВП $m(t)$ с однотипными частицами определяется как марковский процесс со счетным числом состояний 0, 1, 2,…, переходные вероятности которого $P_{ij}(t)$ удовлетворяют дополнительному условию ветвления

$$P_{ij}(t) = \sum_{j_1+\cdots+j_i=j} P_{1j_1}(t)P_{1j_2}(t)\ldots P_{1j_i}(t)\,.$$

Состояния 0, 1, 2,…, в ВП интерпретируются как числа частиц. Вероятность $P_{ij}(t)$ равна вероятности

$$P\{\mu(t+t_0) = j \mid \mu(t_0) = i\}$$

того, что i частиц за время t превращаются в j частиц. Основным аналитическим аппаратом ВП являются производящие функции:

$$F(t;\ s) = \sum_{n=0}^{\infty} P\{\mu(t) = n \mid \mu(0) = 1\} s^n\,.$$

В работе [Simkin, Roychowdhury, 2007b] высказано предположение о том, что по числу опечаток в ссылках можно определить число ученых, прочитавших работу. Поскольку опечатки при цитировании встречаются весьма редко, то только “выдающиеся” работы, т. е. работы, получившие значительное число цитирований, дают достаточный объем информации для проведения такого исследования. М. Симкин и В. Ройчодхури изучили опечатки, допущенные в цитировании выбранной для примера статьи[1], которая на момент изучения имела 4300 цитирований. Из этого числа 196 цитирований содержали опечатки, из них только 45 различные.

Далее, тех авторов, которые прочли оригинальную статью, а затем процитировали ее в своей работе, будем называть “читателями”, а тех авторов, которые не читали оригинал, а только процитировали его, будем называть “цитателями”. Отношение числа “читателей” к числу “цитирующих” R можно оценить как отношение числа разных опечаток D к полному числу опечаток T.

$$R \approx D / T \qquad (3.6.7)$$

Подставляя реальные значения $D = 45$, $T = 196$ в (3.6.7) получаем $R = 0{,}23$. Тщательный анализ дает близкий ответ.

По статистике опечаток определим среднее число раз n_p распространения типичной опечатки:

$$n_p = \frac{T - D}{D} \qquad (3.6.8)$$

Число случаев распространения опечатки отвечает числу копирований из статьи, в которой впервые появилась опечатка, или из одной из последующих статей, в которой копировали из нее

[1] Kosterlitz, J. M.; Thouless, D. J. (1973). Ordering, metastability and phase transitions in two-dimensional systems // J. of Physics C: Solid State Physics 6: 1181–1203.

(или копировали из скопировавшей стати и т. д.). Цитирование с опечаткой ничем не отличается от правильного цитирования в том, что касается копирования. Это означает, что выбранные случайным образом цитирования в среднем скопированы (включая скопированные со скопированных и т. д.) n_p раз. Цитирования с прочтением ничем не отличаются от цитирований без прочтения в случае копирования. Поэтому каждое цитирование с прочтением в среднем было скопировано n_p раз. Таким образом, доля цитирований с прочтением равна:

$$R = \frac{1}{1+n_p} \tag{3.6.9}$$

После подстановки уравнения (3.6.8) в уравнение (3.6.9) получаем уравнение (3.6.7).

Однако следует отметить, что среднее число случаев распространения опечатки не равно числу копирований цитирования, а равно числу его *правильных* копирований. Обозначим среднее число копирований цитирования (включая скопированное со скопированного и т. д.) для конкретного цитирования через n_c. Его можно определить из n_p следующим образом. Величина n_c состоит из двух частей: n_p (правильно скопированные цитирования) и цитирования с опечаткой. Если вероятность сделать опечатку равна M, число правильно скопированных цитирований равно n_p, то общее число скопированных цитирований равно $n_p/(1-M)$, и число цитирований с опечаткой равно $n_pM/(1-M)$. Поскольку каждое цитирование с опечаткой было само скопировано n_c раз, получаем следующее уравнение для n_c:

$$n_c = n_p + n_p \times \frac{M}{1-M} \times (1+n_c) \tag{3.6.10}$$

Уравнение (3.6.10) имеет решение:

$$n_c = \frac{n_p}{1 - M - n_p \times M} \qquad (3.6.11)$$

После подстановки уравнения (3.6.8) в уравнение (3.6.11) получим:

$$n_c = \frac{T - D}{D - MT} \qquad (3.6.12)$$

Отсюда получаем:

$$R = \frac{1}{1 + n_c} = \frac{D}{T} \times \frac{1 - (MT)/D}{1 - M} \qquad (3.6.13)$$

Вероятность введения опечатки можно оценить как $M = D/N$, где N – общее число цитирований. После подстановки в уравнение (3.6.13) получим:

$$R = \frac{D}{T} \times \frac{N - T}{N - D}. \qquad (3.6.14)$$

Подставляя $D = 45$, $T = 196$ и $N = 4300$ в уравнение (3.6.14), получим $R \approx 0{,}22$, что очень близко к первоначальной оценке, полученной с помощью уравнения (3.6.7).

В своей монографии [Moed, 2005] Х. Моед приводит высказывание Ю. Гарфилда, касающееся теории М. Симкина и В. Ройчодхури: «При составлении индекса цитирования применяется ряд неявных процедур, унифицирующих варианты того, как авторы организуют ссылки. Поэтому, то, что доступно в отчетах *SCI*, не обязательно совпадает с тем, что указал автор. Игнорирование таких процедур привело этих авторов к “вопиющей ошибке”, в результате чего появилась теория об авторах, которые цитируют не читая». Х. Моед замечает, что спорным является само утверждение, что если автор копирует ссылку, то он не читал работу, на которую ссылается. Однако применяемая модель сама по себе интересна. Авторы присоединяются к этому мнению.

Глава 4. Начала анализа*

В основе вещей лежит число.
Пифагор

4.1. Общие сведения

Если документ d_i цитирует документ d_j, то эту взаимосвязь можно представить в виде стрелки, ведущей из d_i в d_j (рис. 4.1.1). Таким образом, на основе множества документов D можно построить направленный ациклический граф, который называется графом цитирования или сетью цитирования. Последний термин обычно используется, когда граф является связным.

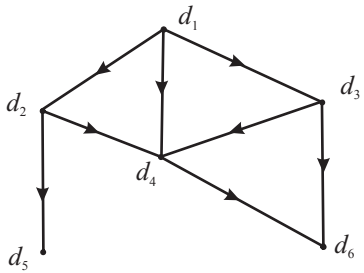

Рис. 4.1.1. Сеть цитирования

Справка. Ориентированный граф (орграф) – (мульти) граф, ребрам которого присвоено направление. Направленные ребра именуются также *дугами*, а в некоторых источниках [Ope, 2008] и просто ребрами. Направленный ациклический граф – случай ориентированного графа, в котором отсутствуют *направленные циклы*, т. е. пути, начинающиеся и кончающиеся в одной и той же вершине. Направленный ациклический граф является обобщением дерева [Википедия].

В матрице инцидентности графа (называемой матрицей цитирования) на пересечении строки, помеченной цитирующим документом d_i, и столбца, помеченного цитируемым документом d_j, ставится единица. Табл. 4.1.1 представляет матрицу, соответству-

*В данной главе частично использован материал третьей главы монографии [Egghe, Rousseau, 1990], дополненный результатами современных исследований по рассматриваемым темам.

ющую графу, приведенному на рис. 4.1.1. Строки и столбцы можно поменять местами, так чтобы столбцы были помечены цитируемыми документами, а соответствующая инцидентная матрица получается транспонированием первоначальной матрицы.

Таблица 4.1.1

Матрица инцидентности

Документ	d_1	d_2	d_3	d_4	d_5	d_6
d_1	0	1	1	1	0	0
d_2	0	0	0	1	1	0
d_3	0	0	0	1	0	1
d_4	0	0	0	0	0	1
d_5	0	0	0	0	0	0
d_6	0	0	0	0	0	0

Справка. Связный граф – граф, содержащий ровно одну компоненту связности. Это означает, что между любой парой вершин этого графа существует как минимум один путь. Ориентированный граф называется сильносвязным, если в нем существует (ориентированный) путь из любой вершины в любую другую, или, что эквивалентно, граф содержит ровно одну сильносвязную компоненту. Ориентированный граф называется слабосвязным, если является связным неориентированный граф, полученный из него заменой ориентированных ребер неориентированными [Википедия].

Граф цитирования можно представить с учетом хронологии. На рис. 4.1.2 круги соответствуют документам, а документы расположены по годам в соответствии с датой публикации. Вертикальная ось означает время, годы расположены в порядке возрастания сверху вниз.

Из такого графа, даже без знания контекста, можно сделать некоторые выводы [Cawkell, 1974]. Документ d_2 имеет значительное влияние на все последующие работы, поскольку многие цитируют его. Документы d_{13} и d_{14}, вероятно, схожи между собой по содержанию, так как цитируют одни и те же работы (d_{10}, d_{11}, d_{12}). Как будет разъяснено далее, документы d_{13} и d_{14} образуют

библиографическое сочетание. Чем больше у документов общих ссылок, тем они более похожи. До 1988 г. документы, представленные на рис. 4.1.2, образовывали две несвязанные группы. В 1988 г. эти группы стали связными за счет документа d_{12}, в котором были процитированы d_7 и d_8, а в следующем году эти группы стали еще более связными, поскольку, например, d_8 и d_{12} были одновременно процитированы из d_{14}, а d_{12} и d_6 были одновременно процитированы из d_{15}. Подчеркнем, что взаимосвязь двух групп проявилась в документе d_{12}.

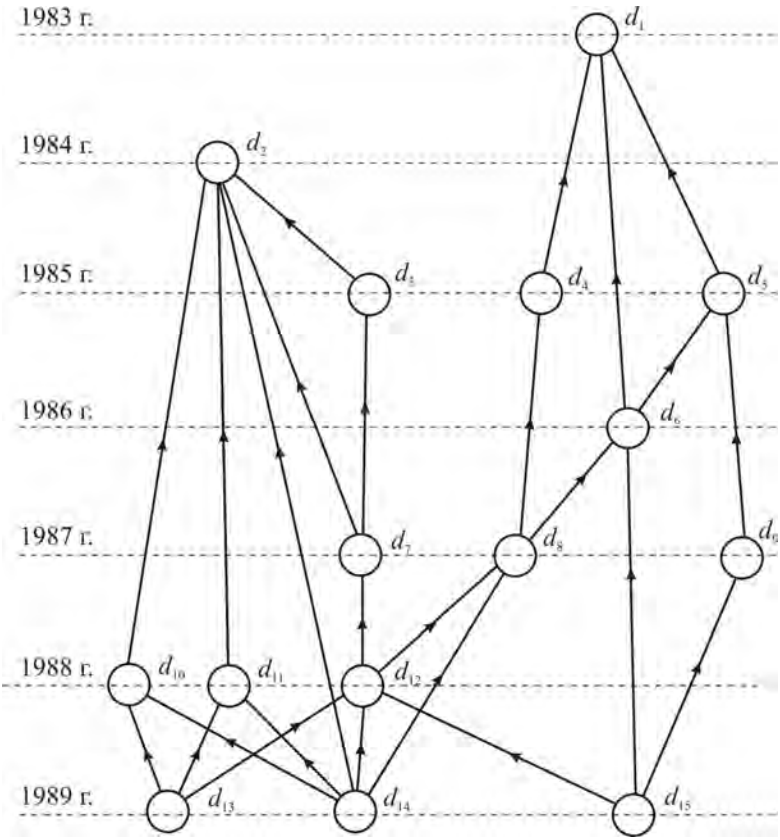

Рис. 4.1.2. Сеть цитирования, упорядоченная по годам

4.2. Графы цитирования

Утверждение 1. [Kochen, 1974]. *Пусть d – некоторый документ, $C(d)$ − множество ссылок в d. Пусть также $C^{-1}(d)$ – множество всех документов, от которых d получил цитирования. Если d_0 – некоторый фиксированный документ, то*

$$d_0 \in \bigcap_{d \in C(d_0)} C^{-1}(d). \qquad (4.2.1)$$

Утверждение (4.2.1) следует из обозначений. Оно констатирует, что когда мы строим множество $C(d_0)$ и берем любой документ d из этого множества, то d_0 принадлежит множеству всех документов, которые цитируют d, т. е. $C^{-1}(d)$. Утверждение 1 проиллюстрировано на рис. 4.2.1.

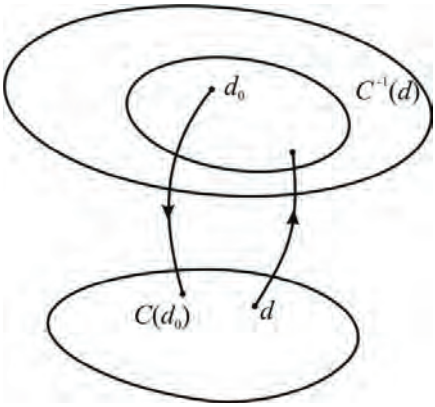

Рис. 4.2.1 Иллюстрация утверждения 1

Следствие (Л. Эгге). *Для каждого документа d_0 верно:*

$$\bigcap_{d \in C(d_0)} C^{-1}(d)$$

− множество всех документов d_1, таких, что список ссылок d_1 включает список ссылок d_0.

Доказательство. Нужно доказать, что два множества совпадают. Пусть d_1 принадлежит пересечению $\bigcap_{d \in C(d_0)} C^{-1}(d)$, тогда d_1

принадлежит и $C^{-1}(d)$. И так для каждого элемента d из $C(d_0)$. Это означает, что каждый документ d из списка ссылок d_0 принадлежит и списку ссылок d_1.

Пусть теперь список ссылок d_1 включает список ссылок d_0. Тогда $C(d_0) \subset C(d_1)$, следовательно

$$\bigcap_{d \in C(d_1)} C^{-1}(d) \subset \bigcap_{d' \in C(d_0)} C^{-1}(d').$$

Согласно утверждению 1 $d_1 \in \bigcap_{d \in C(d_1)} C^{-1}(d)$, т. е. верно, что $d_1 \in \bigcap_{d \in C(d_0)} C^{-1}(d)$. Следствие доказано.

Теорема 2. *Если граф цитирования для непустого конечного множества документов D мощности n слабо связан, тогда для каждого документа d_0 из множества D верно:*

$$D = \bigcup_{j=0}^{N-1} C_j , \qquad (4.2.2)$$

где

$$C_j = \bigcup_{d_j \in C_{j-1}} \left[C\left(d_j \right) \cup C^{-1}\left(d_j \right) \right],$$

$$j > 0 \quad \text{и} \quad C_0 = \left\{ d_0 \right\}.$$

Доказательство. Возьмем любое $d_0 \in D$, где $C(d_0)$ –множество ссылок в d_0, а $C^{-1}(d_0)$ – множество документов, цитирующих d_0. Тогда C_1 – множество всех документов, цитирующих d_0 или цитируемых в d_0. Согласно требованию слабой связности C_1 – не пусто, если D не совпадает с d_0, в этом случае теорема тривиальна. Перейдем к C_2. Согласно утверждению 1 $d_0 \in C_2$.

Теперь будем доказывать методом от противного. Допустим, что некоторое $d \in D$ не принадлежит $\bigcup_{j=0}^{N-1} C_j$. Получаем противо-

речие: в силу связности существует конечный путь из d в d_0, тогда существует $j \le N$, такое, что $d \in C_j$.

Теорема 2 определяет алгоритм для получения всех документов данной коллекции, в случае если коллекция однородна, т. е. граф цитирования слабо связан. Более того, если D – большой компьютерный файл, то алгоритм дает процедуру исследования ядра тематики (допустим, что d_0 – основной документ по тематике). Метод известен как "*cycling*".

Теорема 3. *Среднее количество ссылок в документе, умноженное на количество рассматриваемых документов, равно среднему количеству цитирований на фиксированный документ, умноженному на общее количество различных ссылок.*

Доказательство. Пусть C – матрица цитирований для рассматриваемой коллекции документов, т. е., если $c_{ij} = 1$, то документ d_i содержит ссылку r_j. Столбцы обозначены документами, на которые была хотя бы одна ссылка из документов коллекции. Для коллекции из n документов среднее количество ссылок на документ представляется в виде суммы

$$\frac{1}{n}\sum_{i=1}^{n}\left(\sum_{j=1}^{p} c_{ij}\right),$$

где p – количество различных ссылок; $\sum_{j=1}^{p} c_{ij}$ – количество ссылок в документе d_i. Тогда среднее количество ссылок в документе, умноженное на количество документов, можно представить как общее количество единиц в матрице цитирования:

$$\sum_{i=1}^{n}\sum_{j=1}^{p} c_{ij}.$$

С другой стороны, среднее количество цитирований на документ можно представить в виде

$$\frac{1}{p}\sum_{j=1}^{p}\left(\sum_{i=1}^{n} c_{ij}\right),$$

где $\sum_{i=1}^{n} c_{ij}$ – количество цитирований на r_j (j-ю ссылку). Тогда среднее количество цитирований, умноженное на количество различных ссылок, будет равно общему количеству единиц в матрице цитирования. Теорем доказана.

Пример**.** Предположим, имеется матрица размером 20×120 и известно, что среднее количество ссылок в документе равно 10. Согласно теореме 3, среднее количество цитирований на документ равно 10×20/120=1,67.

4.3. Матричное представление

Цель данного пункта – показать, как такие часто используемые понятия, как “количество соавторов” или “количество публикаций за один год” можно представить формально. Рассмотрим матрицу авторства W, формально описывающую взаимосвязи авторов со своими публикациями за определенный период времени. Следуя работе [Krauze, McGinnis, 1979], предположим, что для данной научной дисциплины известна матрица $W=[w_{ij}]$ размерностью $m \times n$, элементы которой определены как

$$w_{ij} = \begin{cases} 1, \text{если автор } i \text{ способствовал появлению документа } j, \\ 0, \text{в противном случае,} \end{cases}$$

где $i = 1, 2, \ldots, m$; $j = 1, 2, \ldots, n$.

Поскольку матрица имеет размеры (m,n), будем считать, что есть m авторов и n документов. Пусть U – вектор-столбец, состоящий из одних единиц размерности m либо n в зависимости от операции умножения, в которой он участвует. Для любой матрицы

X обозначим через X^{T} транспонированную матрицу. Для матрицы U размерностью $1\times k$ или $k\times 1$ знак транспонирования будем опускать. Элемент i, j матрицы X обозначим через $(X)_{ij}$. В случае вектора, например, WU, его i-я компонента обозначается как $(WU)_i$.

Учитывая введенные определения и принятые ограничения и используя теорию матриц, дадим формальные определения некоторым известным терминам.

а) *Количество документов у автора*. Количество документов, в которые сделал вклад автор i, – это i-я компонента вектора WU, т. е. это $(WU)_i$, что эквивалентно

$$\sum_{j=1}^{n} w_{ij} = (WU)_i = (WW^{\mathrm{T}})_{ii}. \qquad (4.3.1)$$

б) *Количество соавторов*. Количество соавторов у документа j равно $(W^{\mathrm{T}} U)_j$, что эквивалентно

$$\sum_{i=1}^{m} w_{ij} = (W^{\mathrm{T}}U)_j = (W^{\mathrm{T}}W)_{jj}. \qquad (4.3.2)$$

в) *Сотрудничество авторов*. Количество совместных работ для авторов i и j: $(WW^{\mathrm{T}})_{ij}$, $i \neq j$, что эквивалентно

$$\sum_{k=1}^{n} w_{ki} w_{kj} = (WW^{\mathrm{T}})_{ij}. \qquad (4.3.3)$$

При $i = j$ $(WW^{\mathrm{T}})_{ij}$ это множество работ, в которые автор i сделал вклад.

г) *Количество общих авторов у документов*. Количество авторов, сделавших вклад в документы i и j: $(W^{\mathrm{T}}W)_{ij}$, $i \neq j$, что эквивалентно,

$$\sum_{k=1}^{m} w_{ki} w_{kj} = (W^{\mathrm{T}}W)_{ij}. \qquad (4.3.4)$$

Для $i = j$ – количество авторов документа i.

Определим матрицу цитирования по аналогии с матрицей авторства. Предположим, что имеется m документов: d_1, …, d_m, из которых сделаны ссылки на n документов: r_1, …, r_n. Рассмотрим матрицу $C=[c_{ij}]$ размерностью $m \times n$, определенную как

$$c_{ij} = \begin{cases} 1, & \text{если из документа } d_i \text{ имеется ссылка на документ } r_j \\ 0, & \text{в противном случае,} \end{cases}$$

где $i = 1, 2, \ldots, m$; $j = 1, 2, \ldots, n$.

д) *Количество ссылок в документе* d_i:

$$\sum_{j=1}^{n} c_{ij} = (CU)_i = (CC^{\mathrm{T}})_{ii}. \qquad (4.3.5)$$

е) *Количество цитирований, полученных документом* r_j:

$$\sum_{i=1}^{m} c_{ij} = (C^{\mathrm{T}}U)_j = (C^{\mathrm{T}}C)_{jj}. \qquad (4.3.6)$$

ж) *Количество общих ссылок в документах* d_i *и* d_j (коэффициент библиографического сочетания) при $i \neq j$:

$$\sum_{k=1}^{n} c_{ik}c_{jk} = (CC^{\mathrm{T}})_{ij}. \qquad (4.3.7)$$

з) *Количество общих цитирований документов* r_i *и* r_j (коэффициент коцитирования) при $i \neq j$:

$$\sum_{k=1}^{m} c_{ki}c_{kj} = (C^{\mathrm{T}}C)_{ij}. \qquad (4.3.8)$$

4.4. Векторная модель

Для формального представления текстовой информации была разработана векторная модель семантики (*Vector space model, VSM*) [Salton, Wong, Yang, 1975], главная идея которой состоит в представлении каждого документа коллекции в качестве вектора. Это дает возможность сравнивать объекты путем сравнения представляющих их векторов в какой-либо метрике и судить о сход-

стве документов. *VSM* является основой для решения многих задач информационного поиска, как то: поиск документа по запросу, ранжирование документов, классификация документов, кластеризация документов. Это реализуется следующим образом.

Документ в *VSM* рассматривается как неупорядоченное множество “термов”. Термами в информационном поиске называют слова, ключевые слова или фразы. Различными способами можно определить вес терма в документе — “важность” слова для идентификации данного текста. Например, можно подсчитать количество употреблений терма в документе, так называемую частоту терма: чем чаще слово встречается в документе, тем больший у него вес. Если терм не встречается в документе, то его вес в этом документе равен нулю.

Все термы, которые встречаются в документах обрабатываемой коллекции, можно упорядочить. Если для некоторого документа выписать по порядку веса́ всех термов, включая отсутствующие в этом документе, получится вектор, который и будет представлением данного документа в векторном пространстве. Размерность этого вектора, как и размерность пространства, равна количеству различных термов во всей коллекции и одинакова для всех документов.

Допустим, имеется коллекция документов d_j и запросов q:

$$d_j = \left(w_{1,j}, w_{2,j}, \ldots, w_{n,j} \right),$$
$$q = \left(w_{1,q}, w_{2,q}, \ldots, w_{n,q} \right).$$

Здесь $w_{i,j}$ – вес терма i в документе d_j; n – количество уникальных термов в коллекции документов. Располагая таким представлением для всех документов, можно, например, находить расстояние между точками пространства и решать задачу подобия документов: чем ближе расположены точки, тем более похожи соответствующие документы или тем более запрос соответствует документу.

Для определения векторной модели необходимо указать, каким именно образом будет вычисляться вес терма в документе. Существует несколько стандартных способов задания функции взвешивания:

− булевский вес, равный 1, если терм встречается в документе, и 0 в обратном случае;

− *tf* (*term frequency*, частота терма): вес определяется как функция количества вхождений терма в документе;

− *tf-idf* (*term frequency − inverse document frequency*, частота терма − обратная частота документа): вес определяется как произведение функции количества вхождений терма в документ и функции от величины, обратной количеству документов коллекции, в которых встречается данный терм.

Пусть D множество всех документов, t - терм. Определим f(t, d) – число вхождений терма t в документ d, где $d \in D$.

TF (*term frequency* − частота слова, локальный параметр). Простой способ определения tf(t, d) = f(t, d). Для учета длины документа вводится нормализация частоты вхождения терма, которая, например, может выглядеть так:

$$\mathrm{tf}(t,d) = \frac{\mathrm{f}(t,d)}{\max\{\,\mathrm{f}(w,d) : w \in d\,\}}$$

IDF (*inverse document frequency* − обратная частота документа, глобальный параметр): инверсия частоты, с которой некоторое слово встречается в документах коллекции. Учет IDF уменьшает вес широкоупотребительных слов. Для каждого уникального слова в пределах конкретной коллекции документов существует только одно значение IDF.

$$\mathrm{idf}(t,D) = \log \frac{|D|}{|\{\, d \in D : t \in d \,\}|}$$

где |D| – количество документов в корпусе; $|\{d \in D: t \in d\}|$ – количество документов, в которых встречается терм *t*. Для предотвращения деления на ноль к выражению в знаменателе принято прибавлять единицу. Выбор основания логарифма в данной формуле не имеет значения, поскольку изменение основания приводит к изменению веса каждого слова на постоянный множитель, что не влияет на соотношение весов.

Таким образом, мера *tf-idf* является произведением двух сомножителей:

$$\text{tf-idf}(t,d,D) = \text{tf}(t,d) \times \text{idf}(t,D)$$

Большой вес в *tf-idf* получат слова с высокой частотой в пределах конкретного документа и с низкой частотой употреблений в других документах.

Угол отклонения вектора документа от вектора запроса отражает близость документа к запросу: чем меньше угол, тем более вероятно, что документ соответствует запросу. На практике достаточно легко вычислить косинус угла между векторами [Солтон, 1979] (рис. 4.4.1).

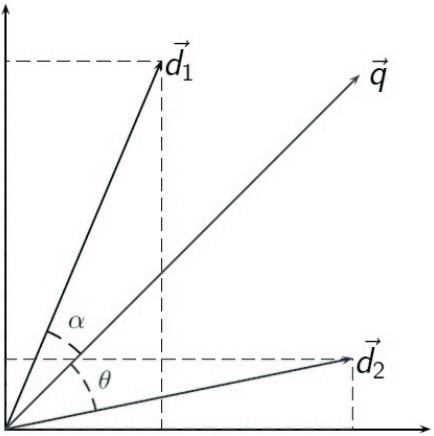

Рис. 4.4.1. Документы и запрос, представленные в виде векторов

Угол θ можно вычислить, например, следующим образом

$$\cos\theta = \frac{\mathbf{d}_2 \cdot \mathbf{q}}{\left\| \mathbf{d}_2 \right\| \left\| \mathbf{q} \right\|}$$

Здесь $\mathbf{d_2}\cdot\mathbf{q}$ – скалярное произведение векторов, $\|\mathbf{d_2}\|$ − норма вектора $\mathbf{d_2}$, $\|\mathbf{q}\|$– норма вектора $\mathbf{q}$,

$$\left\| \mathbf{q} \right\| = \sqrt{\sum_{i=1}^{n} q_i^2} \ .$$

Все векторы данной модели неотрицательны, поэтому нулевое значение косинуса соответствует перпендикулярным векторам, что означает полное несоответствие, а значение 1 – совпадающим векторам.

В классической модели, предложенной в работе [Salton, Wong, Yang, 1975], веса термов вектора документов определяются как произведение локальных (tf) и глобальных (idf) параметров.

Косинусная мера подобия документа d_j и запроса q вычисляется в виде

$$\mathrm{sim}\left(d_j, q \right) = \frac{\mathbf{d}_j \mathbf{q}}{\left\| \mathbf{d}_j \right\| \left\| \mathbf{q} \right\|} = \frac{\sum_{i=1}^{n} w_{i,j} w_{i,q}}{\sqrt{\sum_{i=1}^{n} w_{i,j}^2} \sqrt{\sum_{i=1}^{n} w_{i,q}^2}} \ .$$

В работе [Salton, MacGill, 1983] аналогичным образом определяется $\mathrm{sim}(d_i, d_j)$ − мера подобия двух документов d_i и d_j:

$$\mathrm{sim}_s(d_i, d_j) = \frac{\sum_{k=1}^{n} \left(d_{i,k} \times d_{j,k} \right)}{\sqrt{\left(\sum_{k=1}^{n} d_{i,k}^2 \times \sum_{k=1}^{n} d_{j,k}^2 \right)}} = \frac{\mathbf{d}_i \mathbf{d}_j}{\left| \mathbf{d}_i \right| \left| \mathbf{d}_j \right|},$$

где $|\mathbf{d_i}|$ − длина вектора $\mathbf{d_i}$. Будем называть данную меру индексом Солтона. В этих же терминах можно определить коэффициент подобия Жаккара:

$$\mathrm{sim}_J(d_i, d_j) = \frac{\sum_{k=1}^{n}\left(d_{i,k} \times d_{j,k} \right)}{\sum_{k=1}^{n} d_{i,k} + \sum_{k=1}^{n} d_{j,k} - \sum_{k=1}^{n}\left(d_{i,k} \times d_{j,k} \right)}.$$

Справка. Индекс Жаккара (*Jaccard index*), известный также как коэффициент подобия Жаккара (*Jaccard similarity coefficient*), − это статистический параметр, используемый для сравнения сходства и различия наборов образцов. Пусть имеются наборы образцов *A* и *B*. Индекс Жаккара определяется как результат деления мощности множества, состоящего из пересечения наборов, на мощность множества, состоящего из объединения наборов:

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|}.$$

Расстояние Жаккара − это мера различия между наборами образцов, противоположная коэффициенту сходства, ее можно получить, если из единицы вычесть коэффициент сходства, или, что эквивалентно, путем деления мощности множества, получаемого путем вычитания мощности пересечения из мощности объединения, на мощность объединения:

$$J_\sigma(A,B) = 1 - J(A,B) = \frac{|A \cup B| - |A \cap B|}{|A \cup B|}.$$

Преимущества представленной модели по сравнению с булевой моделью заключаются в следующем: она базируется на линейной алгебре; веса не являются бинарными; позволяет вычислять непрерывную меру подобия между запросами и документами; имеется возможность ранжировать документы согласно вероятной важности.

Основным недостатком данной модели является сложность представления больших документов, а также то, что ключевые слова для поиска должны точно совпадать с термами; в случае поиска

подцепочек в словах классификация может дать ложные результаты; невозможно оценить семантическую эквивалентность документов; информация о порядке термов в документе утеривается; термы статистически независимы; веса интуитивны, а не формальны.

Мера *tf-idf* часто используется для представления документов коллекции в виде числовых векторов, отражающих важность использования каждого терма в каждом документе. VSM-модель дает возможность сравнивать тексты, сравнивая представляющие их векторы в какой-либо метрике (евклидово расстояние, косинусная мера, манхэттенское расстояние, расстояние Чебышева и др.), т. е. производя кластерный анализ.

4.5. Библиографическое сочетание

Существует два метода проведения формальных исследований взаимосвязи документов, которые называются “библиографическим сочетанием” и “коцитированием”. Эти методы базируются на разных принципах выделения взаимосвязи между двумя документами.

Метод “библиографического сочетания” (*Coupling*, далее – БиС) предложен Кесслером в работе [Kessler, 1963a]. В основу этого метода положен принцип выделения взаимосвязи между двумя публикациями на том основании, что в них цитируется один и тот же документ, причем интенсивность их взаимосвязи определяется числом библиографических ссылок, являющихся общими для обеих публикаций. Согласно данному методу две работы прочно и навсегда связаны, и эта связь не зависит от новых поступлений в массиве публикаций [Маршакова, 1981].

Говорят, что два документа d_A и d_B находятся в состоянии БиС, если существует хотя бы один документ, например d_C, на который имеется ссылка в документах d_A и d_B. Очевидно, в изучаемом массиве документов может существовать не один доку-

мент типа d_C. Поэтому для определения “силы связи” (Coupling strength) документов d_A и d_B вводится коэффициент БиС (далее − КБС), который определяется количеством общих ссылок в документах d_A и d_B. На рис. 4.5.1 показан БиС документов d_A и d_B с КБС, равным двум; здесь документы d_A и d_B выступают в роли цитирующих, а документы d_C и d_D − в роли цитируемых.

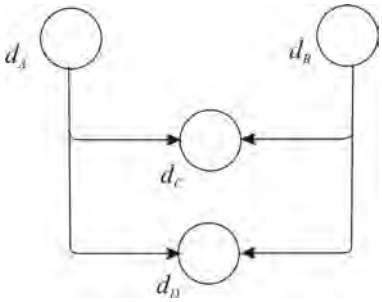

Рис. 4.5.1. Библиографическое сочетание документов d_A и d_B

Предлагая метод БиС, М. Кесслер исходил из гипотезы, что библиография научной публикации дает автору возможность показать интеллектуальную среду, в которой он работает; и если две статьи содержат сходную библиографию, то между ними есть скрытое родство [Kessler, 1965].

Технический аппарат метода БиС может быть представлен следующим образом:

а) одна и та же библиографическая ссылка, используемая в двух документах, называется единицей сочетания между этими документами;

б) несколько документов образуют группу $G_A(P_0)$, если каждый документ группы имеет по меньшей мере одну единицу сочетания с исследуемой публикацией P_0;

в) коэффициент БиС между P_0 и любым элементом группы G_A измеряется числом единиц связи n между ними.

В работе [Kessler, 1963] определены два критерия библиографического сочетания.

Критерий 1. Документы образуют группу $G_A(P_0)$, если каждый документ группы имеет хотя бы одну единицу сочетания с документом P_0, т. е. в группу $G_A(P_0)$ входят документы, содержащие хотя бы одну общую ссылку с документом P_0. Значение коэффициента БиС между документом P_0 и произвольным документом X из группы $G_A(P_0)$ − это количество единиц сочетания для пары (P_0, X). Через $G_A(P_0; n)$ обозначим подмножество элементов группы $G_A(P_0)$, которые имеют коэффициент сочетания с P_0, равный n (рис. 4.5.2). Знак "$\rightarrow$" указывает направление цитирования ("кто кого цитирует"); $G_A(P_0) = \{P_0, X, Y, Z\}$; $G_A(P_0;\ 2) = \{X, Y\}$. В отличие от этого определения будем считать, что документ P_0 библиографически сочетается сам с собой с коэффициентом, равным количеству ссылок в P_0. Если в документе P_0 нет ссылок, то библиографическое сочетание отсутствует, т. е. P_0 сам с собой не сочетается.

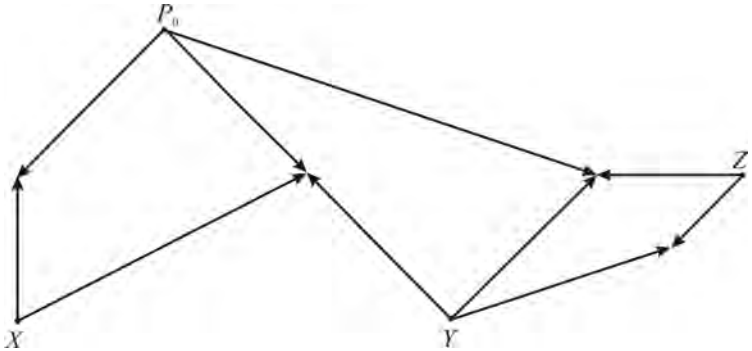

Рис. 4.5.2. Группа $G_A(P_0)$

Критерий 2. Документы образуют группу G_B, если каждый элемент группы имеет хотя бы одну единицу сочетания с каждым другим элементом этой группы.

На рис. 4.5.2 подмножество $\{P_0, X, Y\}$ образует такую группу G_B. Документ Z не может быть включен в эту группу, так как он не имеет ни одной общей ссылки с документом X.

В работе [Kessler, 1963] приведены результаты следующего эксперимента. Рассматривались 265 публикаций из 97 тома журнала *Physical Review*. Каждая публикация рассматривалась в качестве P_0, т. е. было порождено 265 групп G_A. Количество публикаций в группе варьировалось от единицы (нет сочетаний) до двенадцати. В результате было продемонстрировано существование G_A-феномена в хорошо отредактированном множестве публикаций и устоявшейся области науки.

Приведем несколько утверждений, касающихся групп G_A, $G_A(n)$, G_B. Доказательство этих утверждений представлено в работе [Egghe, Rousseau, 1990].

Утверждение $F1$. Если документ имеет непустой список ссылок, то множество, состоящее из одного этого документа, образует G_B и $P_0 \in G_A(P_0)$ для любого P_0.

Утверждение $F2$. Если $P_1 \in G_A(P_0)$, то $P_0 \in G_A(P_1)$.

Утверждение $F3$. Если $P_1 \in G_A(P_0)$ и $P_2 \in G_A(P_0)$, то в общем случае можно только утверждать, что $P_1 \in G_A(P_2)$, если P_0 имеет только одну ссылку.

Утверждение $F4$. $\bigcup_{n \in \mathbb{N}_0} G_A(P_0; n) = G_A(P_0)$.

Утверждение $F5$. Если $m \neq n$ и $m, n \in \mathbb{N}_0$, то

$G_A(P_0;\ n) \cap G_A(P_0;\ m) = \varnothing$.

Утверждение $F6$. Если $P_1 \in G_B$ и $P_2 \in G_B$, то $P_1 \in G_A(P_2)$ и $P_2 \in G_A(P_1)$.

*Утверждение F*7. Пусть Ω − множество рассматриваемых цитирующих документов. Если $\forall\ P_0 \neq P_1 \in \Omega$ верно условие $P_1 \notin G_A(P_0)$, то либо в документе P_1 нет ссылок, либо P_1 имеет только уникальные ссылки (на которые не ссылается ни один документ из множества Ω). При этом выражение $\forall P_0 \in \Omega$: $P_1 \notin \mathrm{G}_\mathrm{A}(P_0)$ подразумевает, что $P_1 \notin \mathrm{G}_\mathrm{A}(P_1)$, т. е. P_1 не имеет ссылок (для случая, когда P_1 считается связанным сам с собой при наличии ссылок).

*Утверждение F*8. Если $P_0 \in \mathrm{G}_B \subset \Omega$, то верны следующие соотношения между множествами:

$$\bigcap_{P_1 \in \Omega} G_A(P_1) \subset \bigcap_{P_1 \in G_B} G_A(P_1) \subset G_A(P_0) \subset \bigcup_{P_1 \in G_B} G_A(P_1) \subset \bigcup_{P_1 \in \Omega} G_A(P_1) \subset \Omega,$$

$$G_B \subset \bigcap_{P_1 \in G_B} G_A(P_1);\quad G_B \supset \varnothing \subset \bigcap_{P_1 \in \Omega} G_A(P_1).$$

*Утверждение F*9. Если $|\Omega| = m$, то $\forall\ P_0 \in \Omega$, $\forall i > m$ верно $G_A(P_0; i) = \varnothing$.

*Утверждение F*10. $\left| G_A(P_0) \right| = \sum_{n=1}^{|\Omega|} \left| G_A(P_0; n) \right|$.

*Утверждение F*11. В множестве документов, каждый из которых содержит, по меньшей мере, одну ссылку, отношение “библиографически сочетается с” обладает свойствами рефлексивности, симметричности и не обладает свойством транзитивности.

Справка. Бинарным отношением называется подмножество декартова произведения двух множеств. В частности, бинарным отношением на множестве называется непустое множество упорядоченных пар элементов этого множества. Бинарные отношения могут обладать различными свойствами, такими как

− рефлексивность: $\forall x \in M\ (xRx)$;

− антирефлексивность: $\forall x \in M \neg (xRx)$;

− симметричность: $\forall x,\ y \in M\ (xRy \Rightarrow yRx)$;

– антисимметричность: $\forall x,\ y \in M\,(xRy \wedge yRx \Rightarrow x = y)$;

– транзитивность: $\forall x,\ y,\ z \in M\,(xRy \wedge yRz \Rightarrow xRz)$;

– асимметричность: $\forall x,\ y \in M\,\big(\,(xRy \Rightarrow \neg(yRx)\,\big)$.

Асимметричность эквивалентна одновременной антирефлексивности и антисимметричности отношения.

По мнению М. Кесслера, метод БиС в первую очередь может быть использован для определения возможной принадлежности рассматриваемого документа к какой-либо существующей группе. Если известно, что публикация P_0 относится к области интересов исследователя, то система поиска должна предоставлять также и $G_A(P_0)$, т. е. все документы, библиографически сочетающиеся с P_0. В качестве средства поиска метод БиС обладает следующими свойствами:

а) БиС не зависит от слов и языка; весь процесс осуществляется в терминах чисел;

б) для выявления связанных документов не требуется экспертная оценка;

в) группа документов, ассоциирующихся с данным документом, расширяется как в прошлое, так и в будущее; если цитирование документа продолжается, то группа БиС документов растет;

г) метод БиС отражает текущее состояние используемой базы данных по цитированию.

Критика метода БиС содержится в работе [Martyn, 1964]. Согласно этой работе, единица сочетания не является мерой взаимосвязи, так как факт наличия общих ссылок не является гарантией того, что ссылки относятся к одной и той же части информации. Критический обзор метода БиС представлен также в работе [Weinberg, 1974], автор которой указывает, что метод БиС будет лучше всего работать по отношению к обзорным публикациям, поскольку они содержат достаточное количество ссылок на “старые” работы.

4.6. Коцитирование

Второй метод анализа взаимосвязи документов − метод проспективной связи, или коцитирования (*Co-citation*, далее − КЦ) − разработан в 1973 г. одновременно в СССР и США И. В. Маршаковой [Маршакова, 1973] и Г. Смоллом [Small, 1973]. В основе данного метода лежит принцип выделения взаимосвязи между двумя публикациями на основе цитирования их одними и теми же документами.

Говорят, что два документа d_A и d_B связаны отношением КЦ, если оба документа указаны в списке ссылок как минимум одного "третьего" документа. Очевидно, что непосредственно после публикации d_A и d_B ничего нельзя сказать о том, находятся они в состоянии коцитирования или нет, картина проясняется по мере появления ссылок на эти работы. С появлением новых работ в какой-либо области науки связи между ретроспективными работами, изучаемые данным методом, могут измениться, поэтому такая связь документов была названа проспективной (в отличие от ретроспективной БиС по Кесслеру). Проспективная связь публикаций максимально зависит от развития науки и наиболее интересна с точки зрения управления [Маршакова, 1981].

В методе КЦ "сила связи" документов d_A и d_B − коэффициент коцитирования (Co-citation frequency, далее − ККЦ) определяется как число документов, которые одновременно содержат ссылки на документы d_A и d_B. Это определение поясняет рис. 4.6.1, на котором пара документов d_A и d_B находится в состоянии КЦ с коэффициентом три; здесь документы d_A и d_B выступают в роли цитируемых, а документы d_C, d_D и d_E − в роли цитирующих.

Технический аппарат метода КЦ может быть представлен следующим образом [Маршакова, 1981]:

а) документы d_A и d_B считаются связанными отношением КЦ, если существует хотя бы один документ, ссылающийся на документы d_A и d_B;

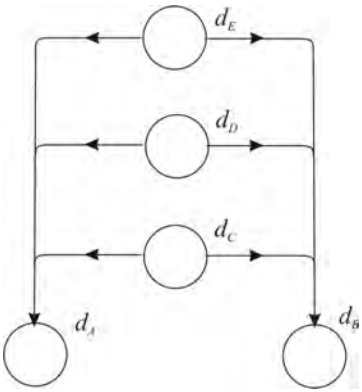

Рис 4.6.1. Коцитирование документов d_A и d_B

б) ККЦ между документами d_A и d_B измеряется числом работ, одновременно цитирующих d_A и d_B с учетом порога, зависящего от величины математического ожидания числа общих ссылок на документы d_A и d_B, полученного в предположении независимости цитирования документов;

в) документы, связанные отношением КЦ, образуют кластер, если они обнаруживают “сильные” связи друг с другом внутри кластера и “слабые” связи с документами, не входящими в этот кластер.

Используя обозначения теории множеств, можно определить коэффициент КЦ пары документов d_A и d_B следующим образом. Пусть Ω_A – множество документов, цитирующих документ d_A, Ω_B – множество документов, цитирующих документ d_B.

Тогда $\Omega_A \cap \Omega_B = F$ – множество документов, цитирующих одновременно d_A и d_B. Мощность множества F (обозначается как $|F|$) определяет коэффициент коцитирования документов d_A и d_B. Разумеется, чтобы быть точными, следует указывать рассматриваемый период, так как имеется существенное различие, рассматривается один год или десять лет.

Относительный ККЦ определяется следующим образом:

$$\frac{\left|\Omega_A \bigcap \Omega_B\right|}{\left|\Omega_A \bigcup \Omega_B\right|}. \qquad (4.6.1)$$

Заметим, что подобным образом можно определить КБС. В разделе 4.3 ККЦ определяется в матричном виде (см. уравнение (4.3.8)).

Как и в случае библиографического сочетания, определяются два критерия.

Критерий B_1. Множество документов образует группу КЦ $M_A(P_0)$, если каждый документ группы процитирован хотя бы один раз вместе с публикацией P_0. Через $M_A(P_0; n)$ обозначим группу публикаций, которые хотя бы n раз процитированы вместе с P_0.

Критерий B_2. Множество документов образует группу КЦ M_B, если каждый член группы процитирован хотя бы один раз вместе с каждым членом группы.

Теперь можно сформулировать и доказать утверждения $F1$ – $F11$ для отношения КЦ. Предлагается выполнить читателю в качестве упражнения.

Анализ КЦ можно использовать для конкретизации деталей взаимосвязи между ключевыми идеями в конкретной научной области. На основе результатов анализа КЦ представляется возможным смоделировать интеллектуальную структуру изучаемой научной области.

Приведем пример из работы [Small, 1973], в которой по результатам анализа цитирования построена сеть цитирований для некоторого массива публикаций в области физики элементарных частиц, которые одновременно принадлежат как минимум двум областям: "*Theory of broken chiral symmetry*" и "*Current algebra*" (данные из *SCI*, 1971). Для получения общей картины анализа сравним данные, приведенные на рис. 4.6.2 и в табл. 4.6.1, табл. 4.6.2.

Таблица 4.6.1

Массив публикации

Документ	Автор(ы), название работы, журнал, год
d_1	Bjorken J.D. Applications of the Chiral U(6)×U(6) Algebra of Current Densities. Phys. Rev., 148, 1467, 1966.
d_2	Gasiowicz S., Geffen D.A. Effective Lagrangians and Field Algebras with Chiral Symmetry. Rev. of Modern Physics, 11, 531, 1969.
d_3	Gell-Mann M. Symmetries of Baryons and Mesons. Phys. Rev., 125, 1067, 1962.
d_4	Gell-Mann M. The Symmetry Group of Vector and Axial Vector Currents. Physics,1, 63, 1964.
d_5	Gell-Mann M., Oakes R., J., et al. Behavior of Current Divergences Under $SU_3 \times SU_3$. Phys. Rev., 175, 2195, 1968.
d_6	Glashow S.L., Weinberg S. Breaking Chiral Symmetry. Phys. Rev. Letters, 20, 224, 1968.
d_7	Lovelace C. A Novel Application of Regge Trajectories. Phys. Letters B, 264, 1968.
d_8	Veneziano G. Construction of a Crossing-Symmetric, Regge-Behaved Amplitude for Linearly Rising Trajectories. Nuovo Cimento, 57, 190, 1968.
d_9	Weinberg S. Pion Scattering Lengths. Phys. Rev. Letters, 17, 616, 1966.
d_{10}	Wilson K.G. Non-Lagrangian Models of Current Algebra. Phys. Rev., 179, 1499, 1969.

Таблица 4.6.2

Информация о цитировании

Документ	d_2	d_3	d_4	d_5	d_6	d_7	d_8	d_9	d_{10}
d_1	1, 6	9, 0	7, 1	13, 3	5, 1	2, 0	2, 0	2, 2	11, 4
d_2		6, 9	9, 2	11, 5	10, 8	1, 3	1, 0	9, 6	1, 17
d_3			19, 4	32, 0	18, 1	3, 0	0, 0	6, 3	7, 1
d_4				20, 0	15, 0	4, 0	0, 0	9, 1	6, 1
d_5					50, 5	7, 0	7, 0	19, 1	18, 4
d_6						1, 1	0, 0	10, 2	7, 0
d_7							68, 1	21, 1	1, 4
d_8								17, 0	1, 0
d_9									0, 1

Примечание. Каждая клетка (*i, j*) содержит две цифры, указывающие значения ККЦ и КБС между публикациями *i* и *j* соответственно (данные *SCI*, 1971).

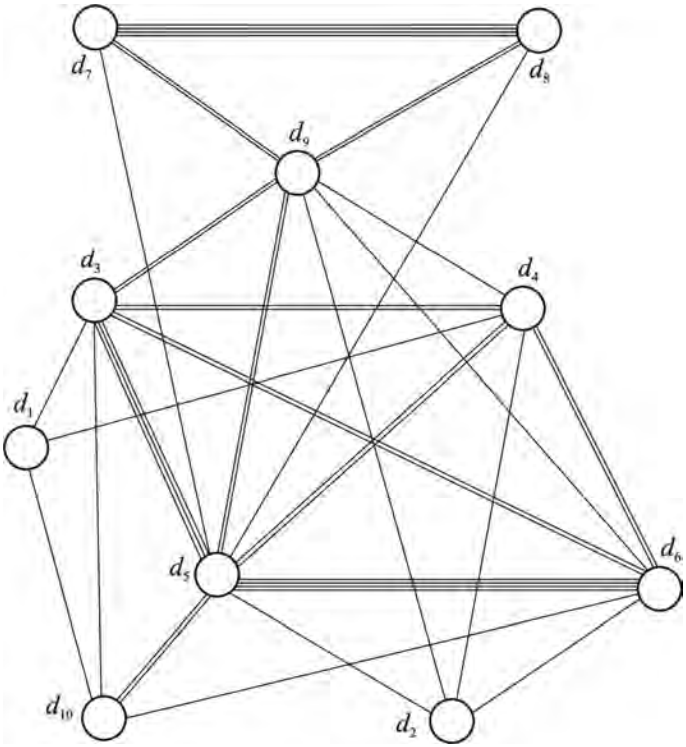

Рис. 4.6.2. Сеть коцитирования для массива публикаций из табл. 4.6.1

На этом рисунке четыре линии между вершинами − ККЦ ≥ 49; три линии – 25≤ ККЦ ≤ 48; две линии – 13 ≤ ККЦ ≤ 24; одна линия − 7 ≤ ККЦ ≤ 12; (остальные случаи не показаны).

Графы, построенные на основе результатов анализа КЦ, существенно отличаются от графов, построенных на данных о БиС, однако оба метода пригодны для выявления подобий в массивах научной документации и интерпретации одновременно высказанных идей в междисциплинарных исследованиях. Результаты

КЦ могут также использоваться при поиске информации, например для получения списка новых документов для каждой коцитируемой пары. Метод КЦ может быть применен для построения кластера или ядра публикаций для определенной предметной области исследований. Эта идея нашла свое отражение при создании “Атласа науки” (см. *ISI”s Atlas of Science,* [Garfield, 1981] и [Garfield, 1987]). В работе [Sharabchiev, 1989] приведены результаты сравнения кластеров, полученных на основании БиС и КЦ.

Критика метода БиС, приведенная в работе [Martyn, 1964], может быть отнесена и к методу КЦ, поскольку тот факт, что две работы находятся в состоянии коцитирования, не означает, что они содержат одинаковые блоки информации. Автор работ [Edge, 1977] и [Edge, 1979] выражает мнение, что численные результаты, полученные на основе анализа КЦ, могут иметь только ограниченное использование. Критику метода КЦ можно найти также в работе [Sullivan, et al., 1977].

4.7. Модель Маршаковой

Рассмотрим математическую модель взаимосвязи научных документов, представленную в работе [Маршакова, 1981]. В основе этой модели лежит изучение связей между документами по общим цитирующим работам. Такая связь в работе И. В. Маршаковой была названа проспективной. Далее будем считать, что выражения “документы находятся в состоянии проспективной связи” и “документы находятся в состоянии коцитирования” означают одно и то же.

Пусть M − множество (массив) исследуемых документов:

$$M = \{y_1, y_2, \ldots y_m\}. \qquad (4.7.1)$$

В данном множестве выделим два подмножества: $B \subset M$ цитируемых документов и $P \subset M$ цитирующих документов. Очевидно, что один и тот же документ $y \in M$ может одновременно

входить в подмножества B и P. На парах $< y_i, y_j > \in M \times M$ зададим отношение $y_i \, f \, y_j$, которое означает, что документ y_j цитирует документ y_i, так что

$$\forall\, y_i \,(y_i \in B)\; \exists\, y_j \text{ такое, что } (y_i \, f \, y_j);$$

$$\forall\, y_j \,(y_j \in P)\; \exists\, y_i \text{ такое, что } (y_i \, f \, y_j).$$

Далее будем называть цитируемые документы базовыми, а цитирующие документы – проспективными. Тогда B будет подмножеством базовых документов, а P − подмножеством проспективных документов.

Пусть b − число элементов множества B; p − число элементов множества P. Имеем

$$b = |B|; \qquad (4.7.2)$$

$$p = |P|. \qquad (4.7.3)$$

Очевидно, что b − число базовых (цитируемых) документов в множестве M, а p – число проспективных документов в том же множестве M.

Обозначим через $p(y)$ число проспективных документов базового документа y, т. е. $p(y)$ – количество работ, цитирующих документ y, $P(y)$ − множество проспективных документов базового документа. Если $p(y) = 0$, то документ y не входит в подмножество B. Обозначим через $b(y)$ число базовых работ проспективного документа y, т. е. $b(y)$ – это количество работ, цитируемых документом y. Если $b(y) = 0$, то документ y не входит в подмножество P.

Отношение f можно задать матрицей M_1:

$$M_1 = \begin{bmatrix} m_{11} \cdots m_{1m} \\ \cdots m_{ij} \cdots \\ m_{m1} \cdots m_{mm} \end{bmatrix}. \qquad (4.7.4)$$

Элемент матрицы m_{ij} может принимать одно из двух значений: $m_{ij} = 0$ означает, что соотношение $y_i \, f \, y_j$ не выполняется; $m_{ij} = 1$ означает выполнение соотношения $y_i \, f \, y_j$.

Сумма элементов i-й строки равна числу проспективных работ документа y_i, т. е. $p(y_i)$:

$$p(y_i) = \sum_{j=1}^{m} m_{ij} . \qquad (4.7.5)$$

Число проспективных работ документа y_i будем называть объемом базового документа $y_i \in B$. Аналогично сумма элементов j-го столбца равна количеству базовых работ проспективного документа y_j, т. е. $b(y_j)$:

$$b(y_j) = \sum_{i=1}^{m} m_{ij} . \qquad (4.7.6)$$

Число $b(y_j)$ будем называть весом проспективного документа $y_j \in P$.

Таким образом, каждому документу $y_i \in B$ можно сопоставить его объем $y_i \to p(y_i)$, а каждому документу $y_j \in P$ можно сопоставить его вес $y_j \to b(y_j)$.

Введем следующие пороговые значения для объема базового и веса проспективного документов. Пусть

$$p(y_i) \geq q ; \qquad (4.7.7)$$

$$b(y_j) \geq \rho . \qquad (4.7.8)$$

Ограничение (4.7.7) означает, что для дальнейшего исследования отбираются те базовые документы, объем которых не меньше q; а так как объем базового документа – не что иное, как число проспективных работ, то условие (4.7.8) означает, что каждая проспективная работа, отвечающая ограничению (4.7.7), должна иметь не менее ρ базовых документов.

На подмножестве $B \subset M$ цитируемых документов определим отношение τ между двумя документами y_i и y_j следующим образом: условие $y_i \ \tau \ y_j$ выполняется тогда и только тогда, когда $P(y_i) \cap P(y_j) \neq \varnothing$. Ясно, что отношение τ является отношением

толерантности (рефлексивность и симметричность). Иными словами, два базовых документа y_i и y_j толерантны, если они имеют общие проспективные работы.

Сопоставим паре $(y_i, y_j) \in B \times B$ величину $\xi_{ij,}$ которая представляет собой число общих ссылок на документы y_i и y_j:

$$\xi_{ij} = \xi(y_i, y_j) = \left| P(y_i) \bigcap P(y_j) \right|. \qquad (4.7.9)$$

Очевидно, что $\xi(y_i, y_j) = \xi(y_j, y_i)$ и

$$\xi_{ii} = \xi(y_i, y_i) = p(y_i). \qquad (4.7.10)$$

Сформируем симметричную n-мерную матрицу

$$M_2 = \begin{bmatrix} \xi_{11} \cdots \xi_{1n} \\ \cdots \xi_{ij} \cdots \\ \xi_{n1} \cdots \xi_{nn} \end{bmatrix}. \qquad (4.7.11)$$

Величина ξ_{ij} не может быть выбрана в качестве единственной меры интенсивности связи (силы связи) двух документов, поскольку число общих ссылок для двух документов зависит от суммарного числа ссылок на документы y_i и y_j. Иначе говоря, сила связи двух документов зависит от объемов $P(y_i)$ и $P(y_j)$ документов y_i и y_j. Поясним это на примере.

Пример. Пусть общее число цитирующих документов в массиве M равно 1000. Предположим, что на документ y_i ссылаются авторы 100 работ и столько же авторов ссылаются на документ y_j. В этом случае появление ссылки, например на документ y_i, можно рассматривать как случайную величину с вероятностью $p_i = 100/1000 = 0{,}1$. При условии, что ссылки на документы y_i и y_j являются независимыми событиями, вероятность их совместного появления $p_{ij} = p_i \times p_j$. В рассматриваемом примере $p_{ij} = 0{,}01$. Отсюда определим математическое ожидание числа общих ссылок на документы y_i и y_j по формуле $E_{ij} = p_{ij} \times p$, где p – число цити-

рующих документов, в данном примере − 1000. Таким образом, наличие десяти общих ссылок вовсе не свидетельствует о тесной связи документов.

И. В. Маршакова предлагает оценивать силу связи двух документов y_i и y_j, опираясь на такое число общих цитирований, которое существенно превосходит математическое ожидание, полученное в предположении независимости цитирования документов y_i и y_j. Основой для выбора таких порогов может служить биномиальное распределение, при малых значениях p_{ij} соответствующее распределению Пуассона, где появление события n раз и более зависит только от математического ожидания. Если, например, полагать существенными только такие значения числа общих цитирований, которые можно считать случайностью не более, чем в пяти случаях из ста, то для каждого дискретного значения $\delta = \{1,2,...,9\}$ появления общих цитирований можно определить соответствующее математическое ожидание $E_{ij.}$

Таблица 4.7.1

Математическое ожидание числа общих цитирований

δ	1	2	3	4	5	6	7	8	9
E_{ij}	≤ 0,36	≤ 0,82	≤ 1,44	≤ 1,94	≤ 2,60	≤ 3,28	≤ 3,98	≤ 4,70	≤ 5,42

И наоборот, зная, что $E_{ij} = \left[p(y_i) \times p(y_j) \right] \Big/ p$, где $p = |P|$, по табл. 4.7.1 можно определить соответствующее значение δ, которое примем в качестве поправочного коэффициента при оценке силы связи двух документов.

Выражение для определения силы связи двух базовых документов y_i и y_j в окончательном виде записывается следующим образом:

$$\xi'_{ij} = \xi_{ij} - \delta. \qquad (4.7.12)$$

Введем понятие связанной группы документов. Пусть $G \subseteq B$ – некоторое множество базовых документов. Тогда

$$\forall y_i \in G \qquad \sum_j \xi'(y_i, y_j) >> \sum_k \xi'(y_i, y_k), \quad (4.7.13)$$

где $y_j \in G$, $y_k \in B \setminus G$. Условие (4.7.13) можно считать определением того, что множество G образует связную группу документов. Предлагается принять это определение в качестве основы процедуры классификации документального потока по общим цитирующим (проспективным) работам.

4.8. Круговая модель мультицитирования

Ранее был определен коэффициент коцитирования двух документов d_X и d_Y как мощность пересечения множеств $\Omega_X \cap \Omega_Y$, где Ω_X и Ω_Y – множества документов, цитирующих документы d_X и d_Y соответственно. Аналогичным образом можно рассматривать пересечение трех множеств документов (три-цитирование) и n множеств (n-цитирование). Эти определения закладываются в понятие уровней агрегации.

Центральная идея круговой модели [Small, 1974] состоит в следующем: документ может быть представлен в виде круга, площадь которого пропорциональна количеству его цитирований, т. е. коэффициент цитирования полагается равным $k\pi r^2$, где k – константа пропорциональности; r – радиус окружности. Таким образом, значение коэффициента цитирования определяет значение r. Такая модель влечет за собой следующую интерпретацию коцитирования и три-цитирования. Если два документа находятся в состоянии коцитирования, то ККЦ будет равен k, умноженному на площадь пересечения двух соответствующих кругов. Если документы находятся в состоянии три-цитирования, то число совместных цитирований будет равно k, умноженному на площадь

пересечения трех кругов. Поскольку для трех документов радиусы полностью определяются числом цитирований, а расстояние между центрами – частотой коцитирования, определяется площадь пересечения, и она может быть использована в качестве оценки значения счетчика три-цитирований.

В работе [Small, 1974] исследовались цитирования шести публикаций 1972 г., относящихся к теме "*Particular physics*" (табл. 4.8.1). Здесь и далее рассматриваемые публикации и документы будем обозначать по их индексу, т. е. d_X и X указывают на один документ и (или) публикацию.

Таблица 4.8.1

Частота цитирования шести документов (данные 1974 г.)

Публикация	**Код**	**Число цитирований**
Amati D., Stanghellini A., et al. Nuovo Cimento, 26, 896-954, 1962	A	88
Benecke J., Chou T. T., et al. Phys. Rev., 188, 2159-2169, 1969	B	127
Caneschi L., Pignotti A.. Phys. Rev. Lett., 22, 1219-1223, 1969	C	39
DeTar C. E., Jones C. E., et al. Phys. Rev. Lett., 26, 675-676, 1971	D	76
Feynman R. P. Phys. Rev. Lett., 23, 14-15, 1969	F	187
Mueller A. H. Phys. Rev. D. 12, 2963-2968, 1970	M	142

Поскольку все документы часто цитировались в 1972 г. и коцитированы друг с другом, можно вычислить все 15 (число всех сторон и диагоналей шестиугольника) расстояний между этими шестью документами, используя приведенные ниже равенства и частоту коцитирования между каждыми двумя документами d_X и d_Y, как описано выше.

Количество коцитирований d_X и d_Y, равно площади пересечения кругов цитирования, а следовательно равно площади сектора

XZFS плюс площадь сектора *YZES* минус площадь четырехугольника *XZYS* и определяется уравнением

$$\frac{\Delta(ZXS)}{2} r_X^2 + \frac{\Delta(ZYS)}{2} r_Y^2 - d_{XY} \times r_X \times \sin\left(\frac{\Delta(ZXS)}{2} \right). \quad (4.8.1)$$

Здесь и далее $\Delta(XYZ)$ − угол *XYZ*.

Более того, имеем

$$\frac{\Delta(ZXS)}{2} = \arccos\left(\frac{r_X^2 - r_Y^2 + d^2{}_{XY}}{2 r_X d_{XY}} \right) \text{ рад}; \quad (4.8.2)$$

$$\frac{\Delta(ZYS)}{2} = \arccos\left(\frac{r_Y^2 - r_X^2 + d^2{}_{XY}}{2 r_Y d_{XY}} \right) \text{ рад}; \quad (4.8.3)$$

$$\sin\left(\frac{\Delta(ZXS)}{2} \right) = \left[1 - \left(\frac{r_X^2 - r_Y^2 + d^2{}_{XY}}{2 r_X d_{XY}} \right)^2 \right]^{1/2} \quad (4.8.4)$$

(d_{XY} − расстояние между d_X и d_Y). Поскольку r_X и r_Y определяются из счетчиков цитирования, значение d_{XY} можно определить из счетчиков коцитирования и приведенных выше равенств. На практике требуется итеративная компьютерная процедура, так как d_{XY} не может быть выражено в явном виде.

Площадь пересечения кругов (заштрихована на рис. 4.8.1), если она существует, соответствует коцитированию.

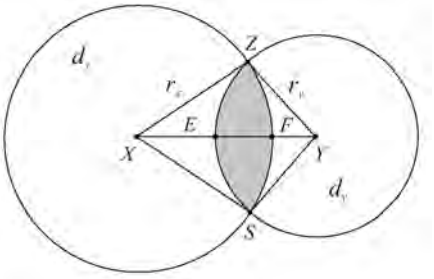

Рис. 4.8.1. Круговая модель коцитирования

Матрица расстояний для документов из табл. 4.8.1 приведена в табл. 4.8.2.

Таблица 4.8.2

Таблица расстояний для документов из таблицы 4.8.1

Код	**B**	**C**	**D**	**F**	**M**
A	8,15	6,51	7,34	8,81	8,08
B		5,09	7,72	4,40	6,50
C			4,69	6,38	6,05
D				8,58	5,35
F					8,19

Нижняя часть матрицы (под диагональю) не заполнена, поскольку расстояние симметрично.

Справка (обратная тригонометрическая функция arccos). Дана функция $y = \cos x$, которая на всей области определения является кусочно-монотонной. Рассмотрим интервал, на котором она строго убывает и принимает все свои значения только один раз – $[0, \pi]$. На этом интервале существует обратная функция $y = \arccos x$, график которой симметричен графику $y = \cos x$ на отрезке $[0, \pi]$ относительно прямой $y = x$. Арккосинусом числа m называется такое значение угла x, для которого $\cos x = m$, $0 \le x \le \pi$, $|m| \le 1$.

По аналогии с коцитированием определим термин “три-цитирование” для некоторого массива изучаемых документов. Говорят, что три документа d_A, d_B и d_C связаны отношением три-цитирования, если все три документа d_A, d_B и d_C указаны в списке ссылок как минимум одного документа из этого массива. Коэффициент три-цитирования определяется как число документов массива, которые одновременно содержат ссылки на документы d_A, d_B и d_C.

Теперь можно вычислить коэффициенты три-цитирования для документов из табл. 4.8.1 на основании круговой модели и сравнить их с фактическими данными. Для этого необходимо рассмо-

треть все комбинации из шести по три, $C_6^3 = 20$, т. е. 20 групп по три документа в каждой. Пример: для двух групп документов BFM и ACD, показанных на рис. 4.8.2, площадь пересечения кругов соответствует коэффициенту три-цитирования этих документов.

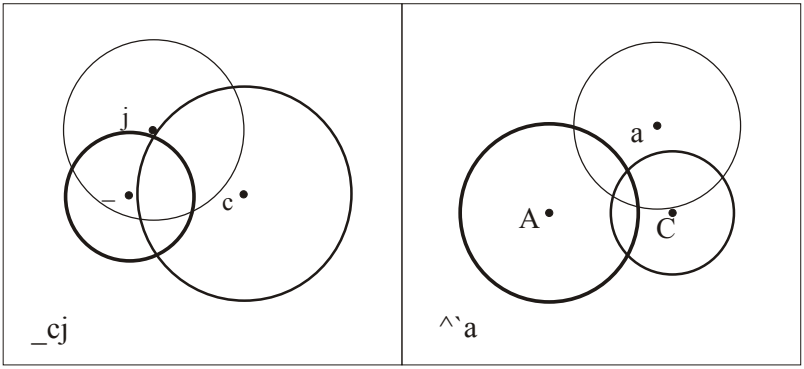

Рис. 4.8.2. Группы документов BFM и ACD

В табл. 4.8.3 приведены реальные значения счетчиков три-цитирования, результаты моделирования (предсказанные значения) и данные об их сравнении. Результат проверки по критерию хи-квадрат (*Chi-square test*) с 20 степенями свободы показал, что суммарное значение χ^2 составляет 5,549, т. е. уровень доверия равен 90 % (критическое значение равно 28,41) [Королюк и др., 1985, с. 124, 547]. Полученные результаты дают основание сделать вывод, что применение круговой модели позволяет достаточно точно оценить значение счетчиков три-цитирования.

Таблица 4.8.3

Результаты моделирования и сравнения

№	Три-цитирование	Точное значение	Предсказание	Значение χ^2
1	BFM	41	39,81	0,036
2	BCF	24	22,60	0,087
3	BDF	19	16,25	0,465

4	ABF	18	19,57	0,126
5	DFM	18	18,62	0,021
6	CFM	17	16,31	0,029
7	BCM	16	16,94	0,052
8	BDM	15	17,07	0,251
9	CDF	14	11,97	0,344
10	AFM	13	12,13	0,062
11	BCD	13	11,27	0,266
12	CDM	11	13,65	0,514
13	ABM	9	11,37	0,494
14	ADM	8	9,72	0,304
15	ACF	7	6,96	0,000
16	ABC	7	6,60	0,024
17	ACM	6	5,86	0,003
18	ADF	5	6,31	0,272
19	ABD	3	4,94	0,762
20	ACD	2	4,56	1,437

На основании данных из табл. 4.8.2 и метода многомерного масштабирования [Kruskal, 1964] можно построить круговую модель три-цитирования (рис. 4.8.3).

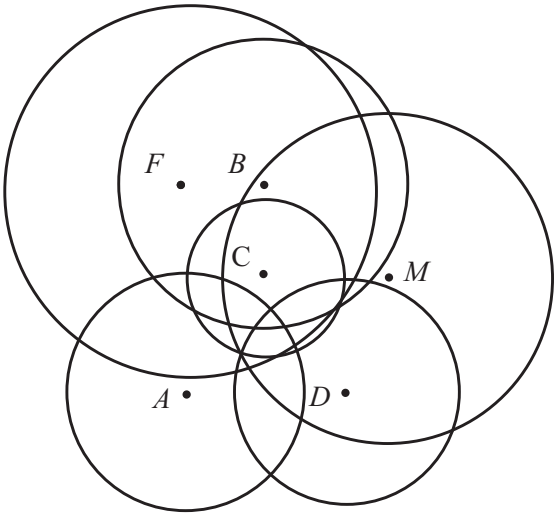

Рис. 4.8.3. Круговая модель три-цитирования

Справка. Критерий согласия Пирсона, или *критерий χ^2 (Хи-квадрат)* − наиболее часто употребляемый критерий для проверки гипотезы о законе распределения. Во многих практических задачах точный закон распределения неизвестен, т. е. является гипотезой, которая требует статистической проверки.

Обозначим через X исследуемую случайную величину. Пусть требуется проверить гипотезу H_0 о том, что эта случайная величина подчиняется закону распределения $F(x)$. Для проверки гипотезы проведем выборку, состоящую из n независимых наблюдений над случайной величиной X. По выборке можно построить эмпирическое распределение $F^*(x)$ исследуемой случайной величины. Сравнение эмпирического распределения $F^*(x)$ и теоретического (или, точнее, гипотетического − т. е. соответствующего гипотезе H_0) распределения $F(x)$ производится с помощью специального правила − критерия согласия. Одним из таких критериев и является критерий Пирсона.

Для проверки критерия вводится статистика

$$\chi^2 = N\sum\frac{\left(P_i^{emp} - P_i^{H_0}\right)^2}{P_i^{H_0}},$$

где $P_i^{H_0} = F(x_i) - F(x_{i-1})$ − предполагаемая вероятность попадания в i-й интервал; $P_i^{emp} = \frac{n_i}{N}$ − соответствующее эмпирическое значение; n_i − число элементов выборки из i-го интервала, N − полный объем выборки. Также используется расчет критерия по частоте, тогда получаем

$$\chi^2 = \sum\frac{\left(V_i - NP_i^{H_0}\right)^2}{NP_i^{H_0}},$$

где V_i − частота попадания значений в интервал. Эта величина, в свою очередь, является случайной (в силу случайности X) и должна подчиняться распределению χ^2 [Википедия].

Справка. Квантили распределения хи-квадрат − числовые характеристики, широко используемые в задачах математической статистики, таких как построение доверительных интервалов, проверка статистических гипотез и непараметрическое оценивание. Пусть F_n − функция распределения $\chi^2(n)$ с n степенями свободы и $\alpha \in [0,1]$.

Тогда α-квантилью этого распределения называется число $\chi^2_{\alpha,n}$, такое что $F_n\left(\chi^2_{\alpha,n}\right) = \alpha$. В Википедии приведена таблица квантилей $\chi^2_{\alpha,n}$, полученная с помощью пакета MATLAB. Зная значения параметров *n* и α, по этой таблице можно получить значение $\chi^2_{\alpha,n}$. В рассматриваемом случае α = 0,9, *n* = 20, $\chi^2_{0.9,20} = 28{,}4120$ [Википедия].

4.9. Мера *Coupling Angle*

Ранее было показано, что на основе матрицы цитирования *A* для каждой пары документов *i*, *j* можно получить коэффициенты библиографического сочетания (и коцитирования) и составить новую матрицу библиографического сочетания (и коцитирования). В работе [Sen, Gan, 1983] предложено использовать нормированную меру, названную *Coupling Angle* (*CA*), которая определяется как косинус угла двух булевских векторов $\mathbf{d_i}$ и $\mathbf{d_j}$ из скалярного произведения

$$CA(\mathbf{d_i}\mathbf{d_j}) = \frac{\mathbf{d_i}\mathbf{d_j}}{\left|\mathbf{d_i}\right|\left|\mathbf{d_j}\right|},$$

где $|\mathbf{d}|$ − длина вектора $\mathbf{d}$.

CA будет принимать значение 1, если два булевских вектора параллельны, и значение 0, если эти векторы перпендикулярны. Тогда коэффициент библиографического сочетания двух публикаций можно представить как угол между соответствующими булевскими векторами – строками матрицы цитирования, что поможет найти подходящие пороги на основе их геометрической интерпретации. Публикации, представленные перпендикулярными векторами, независимы, а статьи, представленные параллельными векторами, относятся к одной тематике. Тогда два документа можно считать относящимися к связанным тематикам, если угол между векторами, представляющими эти документы, не

превосходит заданного угла $0^o \le \varphi < 90^o$. Например, угол $\varphi = 60^o$ соответствует порогу $CA = 0{,}5$, а угол $\varphi = 75^o$ – порогу $CA \approx 0{,}25$.

В работе[Glänzel, Czerwon, 1996] показано, что в описанном выше случае CA совпадает с мерой Солтона (r_{ij}) в обозначениях, принятых в работе [Salton, MacGill, 1983]. Итак, имеем $CA(\mathbf{d}_i\,\mathbf{d}_j) = r_{ij}$, и мера Солтона интерпретируется геометрически как угол между двумя булевыми векторами. Матрицу, представляющую библиографические связи и их коэффициенты, можно получить из булевой матрицы A.

Произведение матриц $B = AA^{\mathrm{T}}$ дает матрицу размером $n \times n$ библиографических сочетаний, B_{ij} – коэффициент сочетания публикаций i, j. Наконец, матрица

$$R = \mathrm{Diag}(B)^{-1/2} \times B \times \mathrm{Diag}(B)^{-1/2} = [r_{ij}] = [CA(\mathbf{d}_i\,\mathbf{d}_j)] \quad (4.9.1)$$

согласно Солтону, является матрицей подобия для пар документов i и j. Здесь Diag(B) означает матрицу, главная диагональ которой совпадает с B, а значения остальных элементов равны нулю. Матрицу, элементы которой r_{ij} являются мерами Солтона, будем называть косино-нормализованной.

Существует популярное определение меры прочности БиС по Солтону как отношения количества общих ссылок в документах к среднему геометрическому значению количества ссылок в каждом документе.

Библиографическое сочетание является симметричным соотношением без всяких ограничений. Более того, r_{ii} принимает значение единица как отображение прочности библиографической связи любого документа с собой. Это имеет смысл, поскольку такое соотношение следует считать максимальным. Сказанное выше означает, что мера Солтона подходит для оценки “прочности” библиографической связи документов.

Матрицу, элементы которой представляют коэффициенты коцитирования также можно получить из булевой матрицы A. Именно, произведение матриц $C = A^{T}A$ дает матрицу, а C_{ij} – коэффициент коцитирования публикаций i, j. В соответствии с (4.9.1) легко получить косино-нормализованную матрицу коцитирования.

В работе [Glänzel, Czerwon, 1996] приведена методика использования библиографического сочетания для определения основных документов, указывающих горячие и передовые тематики исследований. Показана возможность применения методики библиографического сочетания одновременно для оценки исследований и поиска информации.

В работе [Ahlgren, et al., 2003] показано, что нормализация до геометрического среднего более эффективна для случая, когда распределение не является нормальным, чем нормализация до арифметического среднего. В работе [Salton, McGill, 1983] показано, что использование косинусных мер позволяет обнаруживать больше особенностей в данных, чем использование исходных мер.

В работе [Braam, et al., 1988] исследуется влияние порогов цитирования и коцитирования и степени зависимости структуры кластера от выражения, используемого для подсчета коэффициента коцитирования ККЦ. Сравниваются две меры подобия: индекс Жаккара и индекс Солтона.

Индекс Жаккара в применении к цитированиям определяется следующим образом:

$$S_J(i, j) = \frac{coc(i, j)}{cit(i) + cit(j) - coc(i, j)} \quad ,$$

где $S_J(i,j)$ – относительная частота коцитирования между документами i и j (*Relative co-citation frequency*); $cit(k)$ – количество

цитирований, полученных документом *k*; *coc*(*i, j*) – количество коцитирований, полученных документами *i* и *j*.

Индекс Солтона − косино-нормированная частота коцитирования − определяется следующим образом:

$$S_S(i,j) = \frac{coc(i,j)}{\left(cit(i) \times cit(j) \right)^{\frac{1}{2}}} \quad .$$

Значения обоих индексов располагаются на отрезке [0,1]. Если оба документа цитируются хотя бы один раз и не коцитируются, то меры равны нулю. Оба индекса равны единице, если *cit*(*i*) = *cit*(*j*) = *coc*(*i, j*).

Несмотря на то что согласно работе [Braam, et al., 1988] меры различны, часто нормированная мера по Солтону в два раза выше, чем по Жаккару: $S_S(i,j) = 2\, S_J(i,j)$.

4.10. Анализ истории коцитирования

Рассматривая историю цитирования документа можно заметить, что со временем количество цитирований уменьшается, соответственно, уменьшается и ККЦ. В этом пункте рассмотрим динамику изменения ККЦ за фиксированный промежуток времени. Исходные данные о цитировании взяты из *WoS/SCI*. Для определения степени связи между документами вводится коэффициент *RC* (*Relative Co-Citation*), принимающий значения на интервале [0, 1] и вычисляемый по формуле (4.6.1). Работа [Egghe, Rousseau, 1990] содержит четыре примера, которые демонстрируют случаи получения предельных значений *RC*. Приведем один пример с целью демонстрации методики проведения подобных исследований.

Пример. Анализируется история цитирования и коцитирования двух публикаций за период 1966–1982 гг. Первая публикация − *Kawarabayashi K., Suzuki M. Partially Conserved Axial-vector*

Current and the Decays of Vector Mesons. Phys. Rev. Lett. 1966. V. 16. P. 255–257. Вторая публикация – *Riazuddin, Fayyazuddin. Algebra of Current Components and Decay Widths of ρ and K* Mesons*. *Phys. Rev*. 1966. V. 147. P. 1071–1073. Обозначим через Y – год; K и R – число цитирований первой и второй публикаций соответственно; C – коэффициент КЦ первой и второй публикаций. Значения коэффициентов K и R выберем из *WoS/SCI*, коэффициент RC вычисляется согласно формуле 4.6.1. Построим табл. 4.10.1, демонстрирующую историю цитирования и коцитирования обеих работ в период 1966–1982 гг. и динамику изменения коэффициента RC.

Таблица 4.10.1

Динамика цитирования и коцитирования

Y	***K***	***R***	***C***	***RC***
1966	21	6	3	0,13
1967	51	31	25	0,44
1968	82	67	58	0,64
1969	54	41	36	0,61
1970	39	38	31	0,67
1971	31	24	22	0,67
1972	18	14	11	0,52
1973	23	18	18	0,78
1974	18	11	11	0,61
1975	11	4	3	0,25
1976	6	5	5	0,83
1977	7	7	6	0,75
1978	6	5	5	0,83
1979	4	6	4	0,67
1980	4	4	4	1,00
1981	7	4	4	0,57
1982	3	3	3	1,00

Нетрудно заметить, что коэффициент *RC* за редким исключением имеет тенденцию к возрастанию и его значение стремится к единице. В этом примере проиллюстрирован достаточно редкий случай. Обычно (в обозначениях раздела 4.6)

$$\max\left\{\left(\left|\Omega_A \cap \Omega_B\right|\right)/\left|\Omega_A\right|,\ \left(\left|\Omega_A \cap \Omega_B\right|\right)/\left|\Omega_B\right|\right\}$$

стремится к единице. Это случается, например, когда содержание одного документа совпадает с частью другого документа, либо когда одна публикация принадлежит маститому автору, а другая − неизвестному. В этом случае имеет место практика цитирования менее известного автора наряду с цитированием более известного, но не наоборот. Разумеется, в общем случае сходимость к единице отсутствует. В заключение заметим, что большинство пар документов никогда не коцитируются!

Граф коцитирования может использоваться для анализа структуры научных областей. В графе коцитирования вершины означают документы, и две вершины соединяются ребром, если ККЦ превосходит некоторый порог. Во многих случаях использование порога приводит к несвязному графу, в котором документы разбросаны по компонентам. В работах некоторых исследователей (см., например, [Small, Griffith, 1974]), эти компоненты интерпретируются как научные специальности или разделы специальностей.

Для исследования такого деления на компоненты и, следовательно, их интерпретации в работе [Shaw, 1985] проверялась так называемая гипотеза случайности графа (*Random graph hypothesis*). Предполагалось, что ребра графа коцитирования (при определенном пороге) случайным образом выбираются из множества всех возможных ребер. Это предположение составляет нулевую гипотезу, определяющую граф, для которого не важна кластерная структура.

Результаты тестов показывают, что по мере увеличения значения порога важные ассоциации между парами нарушаются, соот-

ветствующие документы оказываются в разных компонентах, и деление становится статистически необоснованным. Аналогично, когда порог уменьшается, образуются незначащие ассоциированные пары, несвязанные документы попадают в одну компоненту, и вновь деление становится статистически необоснованным. Из этого следует, что могут существовать две критические точки для порога коцитирования, определяющие границы статистической обоснованности. В рамках этих границ компоненты могут иметь интерпретацию. Вне границ результаты являются атрибутами кластерной техники, но не указывают на скрытую структуру данных.

Возможности анализа на основе коцитирования изучаются по сей день. Например, в работе [Chen, Lien, 2011] коцитирование рассматривается как способ выявления интеллектуальной структуры областей знаний.

4.11. Анализ контекста цитирования

Кратко обсудим работы по анализу контекста цитирования. Заметим, что первая группа из классификации мотиваций цитирования (концептуальное или техническое, см. 2.1), предложенной в работе [Moravcsik, Murugesan, 1975], обращается именно к контексту.

В работах [Garfield, 1970] и [Small, 1978] высказывается мнение, что цитируемые документы становятся в некотором смысле символом идеи, которую они содержат. В этом ключе организация ссылок представляет собой процесс маркировки. Более того, это одно из объяснений полезности индексов цитирования.

Для исследования вопроса превращения цитируемых документов в “стандартные символы” в работе [Small, 1978] рассматриваются высокоцитируемые документы в области химии. Показано, что индивидуальный контекст цитирования может рассматриваться как экземпляр представления символа. Далее необходи-

мо определить, до какой степени этот символический контекст разделяется разными авторами. Для этого определяется понятие "единообразие использования" как процент контекста, разделяемого большинством цитирующих авторов.

В работе [Small, 1978] приведено наблюдение, касающееся частоты, с которой работы вовлечены в так называемое избыточное (*Redundant*) цитирование. Избыточное цитирование возникает в одной из следующих ситуаций: как указание на одновременное и независимое открытие либо при наличии нескольких хороших источников, в которых приводится одна и та же концепция или процедура.

Итак, высокоцитируемые работы могут рассматриваться как экземпляр (*Exemplar*) или "контекстный символ". Документы занимают различные ниши в пространстве познания авторов, являясь представителями идей, ассоциирующихся с этими документами. Если относиться к ссылкам как к части символической системы языка, то анализ цитирования и коцитирования можно соотнести с литературой по психологии, изучающей ассоциативный отзыв, распознавание слов и искусственный интеллект. Поскольку в одном тексте может быть несколько ссылок и идей, получается сложная система, которая в определенной степени разделяется другими авторами и постоянно меняется [Small, 1987]. Специальные схемы, включенные в *ISI's Atlas of Science*, являются результатом анализа контекста [Small, 1986].

В работе [Gipp, Beel, 2009] представлен подход к выявлению взаимосвязанности документов на основании метода коцитирования, который автор называет *Citation Proximity Analysis* (далее − *CPA*). В отличие от традиционного использования метода КЦ, когда выявление взаимосвязанных публикаций проводится на основе списка ссылок документа, в данном случае взаимосвязанность документов определяется на основе изучения полного

текста документа и уточнения, из какой части документа сделаны ссылки на соответствующие работы. Два документа считаются семантически взаимосвязанными, если в них рассматриваются одинаковые исследовательские задачи.

На основании изучения полного содержания документа вычисляется индекс близости *Citation Proximity Index* (далее − *CPI*). Вычисление проводится в три этапа.

Этап 1. Выполняется разбор содержимого документа и анализ цитирований включая их расположение в документе с помощью модифицированной версии пакета *parsCit* [parsCit] и специально разработанного программного обеспечения.

Этап 2. Цитирования соотносятся с соответствующими ссылками в списке литературы. Утверждается, что в результате выполнения этапов 1, 2 возможная ошибка составляет не более 3 %.

Этап 3. Далее исследуется близость между цитируемыми документами. Чем ближе их цитирование располагается в цитирующем документе, тем они более семантически взаимосвязаны. Например, если два документа цитируются в рамках одного предложения, то это свидетельствует о наивысшей степень близости − *CPI* = 1; если в одном параграфе, то *CPI* = 1/2; если в одной главе, то *CPI* = 1/4; если в одном журнале (документе), то *CPI* = 1/8 и т. д.

Конечно, для уточнения шкалы весов *CPI* необходимы дополнительные исследования, вероятно, следует различать область исследования и тип документа (технический отчет, спецификация патента и т. д.). Кроме того, результаты анализа могут быть улучшены, если рассматривать цитирования одного и того же документа из различных работ и одинаковые вхождения цитирования документов в исходные тексты. Вычисление *CPI* требует значительных усилий, так как необходимо исследовать полный текст и поставить в соответствие каждой ссылке библиографическую единицу, что на практике не всегда возможно, прежде всего из-за

различных стилей ссылок, невозможности оптически распознавать некоторые тексты, а также из-за недоступности полных текстов некоторых документов. Автор метода провел эмпирические сравнения результатов анализа на базе методов коцитирования и *CPA* и на основе этих сравнений утверждает, что метод *CPA* “существенно лучше” определяет взаимосвязанность документов.

Кроме определения взаимосвязанных работ метод *CPA* позволяет: 1) проводить более точный автоматический анализ тематики документов, чем простая принадлежность к определенной области знаний; 2) выявить противоречивые утверждения в различных работах (с использованием дополнительных алгоритмов анализа текста); 3) выявить тип документа, например, что это передовая публикация (*State-of-the-art*).

В работе [Gipp, 2011] представлен метод выявления плагиата *Citation based Plagiarism Detection* (далее − *CbPD*), основанный на анализе цитирования. Автор отмечает, что ранее известные методы выявления плагиата, основанные только на анализе текстов документов, игнорировали цитирования. Эти методы выявляют копирование фрагментов текста и неспособны выявить перефразирование или перевод. Утверждается, что метод *CbPD*, обнаруживающий такие ситуации, позволяет детектировать плагиат на более высоком уровне.

Метод *CbPD* основан на анализе порядка и шаблонов цитирования. В работе [Gipp, 2011] приведена блок-схема алгоритма анализа шаблонов, включающая способ определения плагиата на основе изучения текстов документов. Кроме того, работа содержит сравнение возможностей различных методов детектирования плагиата.

Глава 5. Кластерный анализ

Прежде чем решать задачу, подумай
что делать с ее решением
Д. Пойа

Настоящая глава посвящена кластерному анализу (*Cluster analysis*) – аналитическому инструменту, с помощью которого можно выявлять структурные единицы научных областей на основе библиометрических данных. Кластерным анализом (далее − КА) называют задачу разбиения заданной выборки объектов (ситуаций) на подмножества, называемые кластерами, так чтобы каждый кластер состоял из схожих объектов, а объекты разных кластеров существенно различались. По мнению большинства исследователей, впервые термин “кластерный анализ” был предложен математиком Р. Трионом [Tryon, 1939]. КА представляет собой процедуру обработки информации о некоторой выборке объектов с последующим разбиением объектов на сравнительно однородные группы – кластеры. Главная цель КА − нахождение групп схожих объектов в выборке [Мандель, 1988].

Общие вопросы кластеризации представлены в обзоре [Jain, et al., 1999], где КА представлен как неконтролируемая процедура классификации, рассматриваемая в качестве одного из инструментов анализа данных. Процедуры анализа можно разделить на два типа: объясняющая и подтверждающая (формирование гипотезы или принятие решения) – в зависимости от наличия соответствующей модели исходных данных. Для обоих типов характерна группировка объектов на основе а) верности постулируемой модели или б) естественного группирования, базирующегося на понятии подобия. КА – процедура второго типа. Следует отличать неконтролируемую классификацию (кластеризацию) от контролируемой классификации, при которой имеется набор по-

меченных образцов для обучения и получения описаний классов, что впоследствии используется для соотнесения с классами непомеченных образцов. В случае кластеризации необходимо объединить непомеченные образцы в интерпретируемые группы.

Проблемой КА является недостаточное количество предварительной информации об исследуемых данных (например, отсутствие статистической модели), поэтому следует делать как можно меньше предположений относительно исходной информации. В этих условиях методология КА наилучшим образом подходит для выявления взаимосвязей между данными и позволяет определить (возможно, предварительно) их структуру.

Обычно решение задачи кластеризации включает следующие этапы:

1) подготовка данных (*Pattern representation*), включая извлечение параметров, характеризующих образцы данных и (или) выбор параметров);

2) определение меры близости (*Definition of a pattern proximity measure*), подходящей для исследуемых данных;

3) кластеризация или группирование;

4) обобщение данных (*Data abstraction*), если необходимо;

5) оценка результата (*Assessment of output*), если необходимо.

Подготовка данных включает определение числа кластеров и количества образцов, количества, типа и масштаба параметров. Выбор параметров подразумевает выбор эффективных параметров из числа доступных. Близость образцов данных обычно измеряется с помощью функций сходства или расстояния. Простой функцией дистанции является, например, евклидово расстояние. Процедура группирования в КА может быть выполнена различными способами. Группирование может быть жестким (*Hard*), детерминированным, когда каждый образец попадает точно в одну группу, и мягким (*Soft, Fuzzy*), когда образец может принадле-

жать нескольким кластерам, каждому образцу соответствует множество уровней членства в кластерах. Иерархические алгоритмы кластеризации образуют вложенные последовательности делений, построенные на основе критерия слияния или разделения кластеров. Разделительная кластеризация определяет деление, которое оптимизирует критерий кластеризации. Существуют и другие техники кластеризации, основанные на теории вероятностей и теории графов. Процесс обобщения данных состоит в получении простого и компактного представления множества данных. Простота рассматривается с точки зрения автоматического анализа или интуитивного понимания. В контексте кластеризации имеется в виду компактное описание каждого кластера.

Как оценивается результат алгоритма кластеризации? Какую кластеризацию считать хорошей, какую − плохой? Каждый алгоритм кластеризации производит кластеры, независимо от того, можно ли исходные данные разделить на кластеры. Если кластеры уже имеются, то какие алгоритмы производят кластеры “лучше”, чем другие? Оценка результата кластеризации имеет несколько граней. Во-первых, оценка самих данных. Данные, которые не содержат кластеров, не должны подвергаться кластеризации. Это исследование тенденции (*Cluster tendency*). Во-вторых, исследование достоверности кластеров (*Cluster validity*) – это уже оценка результатов кластеризации. Обычно используются определенные критерии оптимальности, однако эти критерии субъективны. Существует подобие “золотого стандарта”. Кластерная структура считается достоверной, если ее нельзя получить случайным образом, и она не является искусственно полученной на основе кластеризации. При использовании статистической техники кластеризации достоверность достигается точным использованием статистических методов и тестирующих гипотез. Различают три типа определения достоверности: 1) внешний тип

(*External*), когда полученная структура сравнивается с изначально предполагаемой; 2) внутренний (*Internal*), когда изучается, является ли структура по сути подходящей для исследуемых данных; 3) относительный (*Relative*) тест, сравнивающий две возможные структуры и оценивающий их достоинства.

При наличии большого количества разнообразных алгоритмов кластеризации трудно определить, алгоритм какого типа соответствует поставленной задаче. В работе [Dubes, Jain, 1976] приведено сравнение алгоритмов кластеризации на основании критериев, предложенных в [Fisher, Van Ness, 1971]. Эти критерии основываются на 1) способе формирования кластеров; 2) структуре данных; 3) чувствительности кластерной техники к изменениям, не касающимся структуры данных. Однако не существует критического анализа алгоритмов кластеризации, рассматривающих следующие важные вопросы: как нормализовать данные; какая мера подобия соответствует данному случаю; как использовать имеющиеся знания в рассматриваемой области; как эффективно провести кластеризацию очень большого объема исходных данных. Также не существует универсальной техники кластеризации, способной охватить большое разнообразие структур, представляющих множества данных в многомерном пространстве. Часто кластерные алгоритмы содержат неявные предположения о форме или взаимосвязи кластеров. Достаточно часто используются процедуры, предполагающие двумерное пространство, несмотря на то что большинство практических задач являются многомерными. Однако зачастую трудно дать интерпретацию данных в рамках многомерного пространства. Кроме того, данные редко соответствуют “идеальной” структуре (например, сферической или линейной). Поэтому продолжают появляться новые алгоритмы кластеризации, и каждый новый алгоритм работает несколько “лучше” на конкретном распределении образцов.

Вследствие этого исследователю важно не только разобраться в технике, которую предполагается использовать, но и знать детали процесса сбора данных и иметь представление об исследуемой области. Чем больше таких знаний, тем быстрее можно выполнить оценку структуры кластеров.

Процедура кластеризации может включать соответствующие ограничения на исходные данные. Например, может быть сделано предположение, что данные извлечены из смеси нескольких неизвестных распределений, и ставится задача выявления компонент и параметров распределений. Концепция учета плотности (*Density clustering*) кластеризации (когда кластер состоит из плотно расположенных точек пространства) и декомпозиции пространства параметров также включаются в традиционную кластерную методологию, предоставляя технику для разделения перекрывающихся кластеров.

Несмотря на чрезвычайно широкое применение понятия “кластер”, КА как формальная процедура понимается с некоторым “напряжением”, которое возникает вследствие того, что приходится пользоваться нестандартизованной, приводящей к путанице, терминологией.

Диапазон применения КА чрезвычайно широк. Современные тенденции КА и классификация методов кластеризации представлены, например, в работе [Бериков, Лбов, 2008].

5.1. Задача кластеризации

Формальная постановка задачи КА выглядит следующим образом [Википедия]. Пусть X − множество объектов, Y − множество номеров (имен, меток) кластеров. Задана функция расстояния между объектами $\rho(x, x')$. Имеется конечная обучающая выборка объектов $X^m = \{x_1,...,x_m\} \subset X$. Требуется разбить выборку на непересекающиеся подмножества, называемые кластерами, так

чтобы каждый кластер состоял из объектов, близких по метрике ρ, а объекты различных кластеров существенно различались. При этом каждому объекту $x_i \in X^m$ приписывается номер кластера y_i.

Алгоритм кластеризации − это функция a: $X \to Y$, которая любому объекту $x \in X$ ставит в соответствие номер кластера $y \in Y$. В некоторых случаях множество Y известно заранее, однако чаще ставится задача определить оптимальное число кластеров с точки зрения того или иного критерия качества кластеризации. Кластеризация (обучение без учителя) отличается от классификации (обучения с учителем) тем, что метки исходных объектов y_i изначально не заданы и даже может быть неизвестно само множество Y.

При постановке задачи кластеризации, как правило, предполагается достижение следующих результатов: изучение данных путем выявления их кластерной структуры; сжатие данных, если исходная выборка избыточно большая; обнаружение новизны (*Novelty detection*), например выделение нетипичных объектов, которые не удается присоединить ни к одному из кластеров.

Типы входных данных:

а) признаковое описание объектов − каждый объект описывается набором характеристик, называемых параметрами, которые могут быть числовыми или нечисловыми;

б) матрица расстояний между объектами − каждый объект описывается расстояниями до всех остальных объектов метрического пространства;

в) матрица сходства между объектами − учитывается степень сходства объекта с другими объектами выборки в метрическом пространстве.

В монографии [Жамбю, 1988] приведено описание двух фундаментальных требований, предъявляемых к данным, − однородность и полнота. Однородность требует, чтобы все изучаемые

сущности были одной природы. Требование полноты состоит в том, чтобы множества образцов I и параметров J представляли полную опись проявлений рассматриваемого явления.

Прежде чем перейти к обсуждению основных методологических этапов проведения КА в библиометрии, необходимо предупредить читателя, что решение задачи кластеризации принципиально неоднозначно по нескольким причинам (см., например, [Ким и др., 1989]).

1. Цель КА − поиск существующих структур. В то же время его действие состоит в привнесении структуры в анализируемые данные, т. е. методы кластеризации необходимы для обнаружения в данных структуры, которую нелегко найти при визуальном обследовании или с помощью экспертов.

2. Поскольку методы КА разрабатывались для многих научных дисциплин, они сохраняют их специфику. Это важно отметить, поскольку каждая дисциплина предъявляет свои требования к отбору данных, форме их представления и предполагаемой структуре классификации. То, что полезно для одной научной дисциплины, может оказаться ненужным для другой.

3. Многие методы КА представляют собой довольно простые процедуры, как правило, не имеющие достаточного статистического основания. Иными словами, большинство методов КА представляют собой эвристические методы, которые базируются на опыте разработчика и являются не более чем правдоподобными алгоритмами создания кластерных объектов.

4. Не существует однозначно наилучшего критерия качества кластеризации. Известен целый ряд эвристических критериев, а также ряд алгоритмов, не имеющих четко выраженного критерия, однако осуществляющих достаточно разумную кластеризацию “по построению”. Все они могут давать различные результаты. Следовательно, для определения качества кластеризации

требуется эксперт предметной области, который мог бы оценить целесообразность выделения кластеров.

5. Число кластеров, как правило, заранее не известно и устанавливается в соответствии с некоторым субъективным критерием.

6. Результат кластеризации существенно зависит от метрики, выбор которой, как правило, также субъективен и определяется экспертом. Однако следует отметить, что существует ряд рекомендаций по выбору мер близости для различных задач.

7. Различные кластерные методы могут порождать и порождают различные решения для одних и тех же данных. Кластерный метод размещает объекты по группам, состав которых может радикально различаться, если применяются различные методы кластеризации. Ключом к использованию КА является умение отличать “реальные” группировки от навязанных методом кластеризации данных. Например, начало использования в библиометрии метода КА следует датировать 1963 годом, когда М. Кесслер ввел термин “библиографическое сочетание” (см. 4.5) для определения взаимосвязи между документами. Второй, наиболее распространенный метод КА − метод коцитирования, который был разработан в 1973 г. И. В. Маршаковой и Г. Смоллом (см. 4.6). Результатом применения данных методов стало создание кластеров, которые в случае библиографического сочетания образуют цитирующие документы, а в случае коцитирования − цитируемые.

Тривиальные примеры построения кластеров на основе различных библиографических подходов приведены в работе [Boyack, Klavans, 2010]. Представленные в ней примеры иллюстрируют различие полученных результатов. Допустим, требуется разделить множество работ, опубликованных за какой-либо период времени, на группы, кластеры достаточно схожих между собой публикаций. На рис. 5.1.1 рассматриваемое множество из

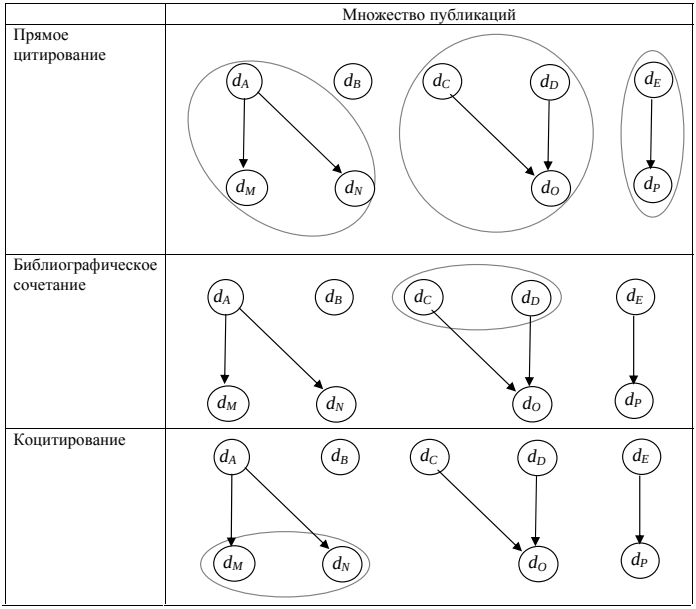

Рис. 5.1.1. Различные подходы к кластеризации множества документов

девяти публикаций обозначено прямоугольниками. Показаны варианты кластеризации на основе прямого цитирования, библиографического сочетания и коцитирования без учета внешних ссылок.

5.2. Стандартизованная матрица данных

Результаты “сырых” измерений можно представлять в виде матрицы, строки которой соответствуют наблюдаемым объектам, а столбцы – параметрам, описывающим состояние объекта. Такую матрицу будем называть матрицей исходных данных (МД). Обозначим число объектов через N, число параметров – через n. Тогда МД имеет вид $Z = [z_{ij}]$, где z_{ij} указывает на значение, которое в результате проведения измерения принимает j-й параметр на i-м объекте. Вектор $\mathbf{z_j}$, обозначающий j-й столбец матрицы Z,

имеет смысл набора значений j-го параметра на всех N объектах. Соответственно вектор $\mathbf{z_i}$, обозначающий i-ю строку матрицы Z, показывает, какие значения принимают все n параметров на i-м объекте.

Заметим, что МД существенно изменяется при изменении шкал, в которых проводится измерение. Поскольку в силу этого различные столбцы МД (параметры) трудно сопоставить, еще до проведения анализа МД следует преобразовать к “стандартному” виду, при котором средние значения всех параметров равны нулю, а дисперсии – одному и тому же числу, например единице. Такое преобразование можно понимать как приведение всех параметров к единой шкале.

Обозначим стандартизованную матрицу через $X = [x_{ij}]$, Переход от матрицы Z к матрице X осуществляется следующим образом. Для простоты положим $N = n$. Вычисляем сначала среднее значение $\overline{z}_j$ каждого параметра:

$$\overline{z}_j = \frac{1}{n}\sum_{i=1}^{n} z_{ij}\,,\ j = 1, 2,\ldots, n$$

затем величину

$$\overline{\sigma}_j{}^2 = \frac{1}{n}\sum_{i=1}^{n}(z_{ij} - \overline{z}_j)^2\,,\ j = 1, 2,\ldots, n$$

называемую выборочной дисперсией, и, наконец, элементы матрицы X:

$$x_{ij} = (z_{ij} - \overline{z}_j) / \overline{\sigma}_j \ \ (i = 1, 2,\ldots, n;\ j = 1, 2,\ldots, n).$$

После проведения этих преобразований можно утверждать, что все элементы матрицы X обладают следующими свойствами:

$$\overline{x}_j = \frac{1}{n}\sum_{i=1}^{n} x_{ij} = 0\,,\ j=1, 2,\ldots, n,$$

$$\frac{1}{n}\sum_{i=1}^{n} x_{ij}^2 = 1, \quad j = 1, 2, \ldots, n.$$

Таким образом, все параметры (столбцы матрицы) имеют нулевое среднее значение, а дисперсия каждого параметра одинакова и равна единице. Эти свойства матрицы X позволяют говорить о ней как о стандартизованной матрице данных [Браверман, Мучник, 1983].

С геометрической точки зрения матрицу X можно интерпретировать двояко. С одной стороны, можно рассматривать n-мерное пространство, оси которого соответствуют отдельным параметрам, а каждую строку матрицы X интерпретировать как вектор в этом пространстве. Такое пространство называют пространством параметров, а матрица X может быть представлена как совокупность n векторов в пространстве параметров. С другой стороны, можно рассматривать N-мерное пространство, оси которого соответствуют отдельным объектам. Тогда каждый столбец матрицы X представляет собой вектор в этом пространстве, а матрица X – совокупность n таких векторов. Данное пространство называют пространством объектов, потому что в нем в силу приведенных выше преобразований все векторы $\mathbf{x_j}$ имеют одинаковую длину, равную $\sqrt{N}$. Таким образом, вопрос о взаимосвязи параметров сводится к вопросу об угле между соответствующими векторами в пространстве объектов.

5.3. Нормализация матриц совместной встречаемости

В работе [Van Eck, Waltman, 2009] проведен анализ четырех мер совместной встречаемости (*Co-occurence*) каких-либо явлений или событий с целью определения “подходящего” способа нормализации исходных данных: коэффициента ассоциативности

(*Association strength*), “косинусной” меры (*Cosine*), индекса включения (*Inclusion index*) и индекса Жаккара (*Jaccard index*).

Пусть O − матрица встречаемости (*Occurrence matrix*) размерности $m \times n$. Например, такой матрицей является матрица цитирования документов или матрица встречаемости термов (общих слов) в документах. Обычно m – это количество документов, на которых строится мера совместной встречаемости. Будем полагать, что O – бинарная матрица, каждый элемент равен нулю или единице; $o_{ki} = 1$, если объект i встречается в документе k; $C = [c_{ij}]$ − матрица совместной встречаемости, например матрица коцитирования. Тогда для всех i и j имеем

$$c_{ij} = \sum_{k=1}^{m} o_{ki} o_{kj} ,$$

т. е., $C=O^{\mathrm{T}}O$.

Существует два типа мер подобия: косвенные (*Indirect*) и прямые (*Direct*), иногда называемые глобальными и локальными [Ahlgren, et al., 2003], [Jarneving, 2008]. Косвенные меры основаны на профилях совместной встречаемости, т. е. на векторах, содержащих коэффициенты совместной встречаемости со всеми другими документами (векторная модель), и обычно используются для анализа коцитирования.

5.3.1. *Прямые меры*. Прямые меры используются для определения подобия документов. С этой целью число совместной встречаемости объектов нормируется а) относительно встречаемости или б) совместной встречаемости каждого объекта. Будем рассматривать прямые меры подобия.

Пусть s_i – общее количество встречаемости объекта i или совместной встречаемости для объекта i. В случае а)

$$s_i = c_{ii} = \sum_{k=1}^{m} o_{ki} ,$$

в случае б)

$$s_i = \sum_{j=1,\, j\neq i}^{n} c_{ij} \ .$$

Приведем формальные определения мер подобия.

Определение 1. Прямая мера подобия – это функция $S(c_{ij}, s_i, s_j)$, обладающая тремя свойствами:

1) область определения функции:

$$Ds = \left\{ (c_{ij}, s_i, s_j) \in \mathbb{R}^3 \mid \ 0 \le c_{ij} \le \min(s_i, s_j) \text{ и } s_i, s_j > 0 \right\},$$

б) область значений – подмножество $\mathbb{R}$;

в) $S(c_{ij}, s_i, s_j)$ – симметричная мера относительно s_i, s_j, т. е. $S(c_{ij}, s_i, s_j) = S(c_{ij}, s_j, s_i)$ для всех $(c_{ij}, s_i, s_j) \in Ds$.

Определение 2 (монотонность). Две прямые меры подобия $S_1(c_{ij}, s_i, s_j)$ и $S_2(c_{ij}, s_i, s_j)$ называются монотонно связанными тогда и только тогда, когда

$$S_1(c_{ij}, s_i, s_j) < S_1(c_{ij'}, s_{i'}, s_{j'}) \Leftrightarrow S_2(c_{ij}, s_i, s_j) < S_2(c_{ij'}, s_{i'}, s_{j'})$$

для всех (c_{ij}, s_i, s_j), $(c_{ij'}, s_{i'}, s_{j'}) \in Ds$.

Свойство монотонной связанности является важным, так как многие методы анализа, рассматриваемые в библиометрии, нечувствительны к монотонным трансформациям подобия. Определим в этих терминах популярные меры подобия:

– коэффициент ассоциативности

$$S_A(c_{ij}, s_i, s_j) = \frac{c_{ij}}{s_i s_j}; \qquad (5.3.1)$$

– “косинусная” мера

$$S_C(c_{ij}, s_i, s_j) = \frac{c_{ij}}{\sqrt{s_i s_j}}; \qquad (5.3.2)$$

– индекс включения

$$S_I(c_{ij}, s_i, s_j) = \frac{c_{ij}}{\min(s_i, s_j)}; \qquad (5.3.3)$$

– индекс Жаккара

$$S_J(c_{ij}, s_i, s_j) = \frac{c_{ij}}{s_i + s_j - c_{ij}}. \qquad (5.3.4)$$

Если c_{ij} имеет числовое значение, то значения индексов изменяются в диапазоне от нуля до единицы. Нетрудно показать, что между мерами выполняются следующие отношения:

$$S_A(c_{ij}, s_i, s_j) \le S_J(c_{ij}, s_i, s_j) \le S_C(c_{ij}, s_i, s_j) \le S_I(c_{ij}, s_i, s_j).$$

Индекс S_A используется, например, в работах [Van Eck, Waltman, 2007] и [Van Eck, et al., 2006], под другими именами, например *Proximity index*, − в работе [Peters, Van Raan, 1993], *Finity or activity index* − в работе [Zitt, et al., 2000].

Значение коэффициента ассоциативности пропорционально отношению между обозреваемым совместным появлением объектов i и j и ожидаемым совместным появлением объектов i и j в предположении, что появления этих объектов статистически независимы.

Индекс S_C равен отношению между обозреваемой совместной встречаемостью и геометрическим средним встречаемости объектов. Он также интерпретируется как косинус угла между i- и j-колонками матрицы O, колонки рассматриваются как векторы в m-мерном пространстве [Salton, McGill, 1983].

Пример использования индекса включения S_I приведен в работе [McCain, 1995]. Вне конкретной области эту меру называют коэффициентом Симпсона (*Simpson coefficient*).

Индекс S_J равен отношению совместной встречаемости объектов i и j и встречаемости, по крайней мере, одного объекта из двух.

Индекс используется в работах [Small, 1973], [Peters, Van Raan, 1993].

5.3.2. *Теоретико-множественные меры подобия.* Рассмотрим понятия вероятностных (*Probabilitic*) и теоретико-множественных (*Set-theoretic*) мер и покажем, что между индексом ассоциативности и остальными тремя мерами существует фундаментальное различие. Косинусная мера, индекс включения и индекс Жаккара являются теоретико-множественными мерами, а коэффициент ассоциативности – вероятностной.

Далее будем считать, что s_i означает встречаемость объекта i. Каждая колонка матрицы встречаемости O может рассматриваться как представление множества всех документов, в котором встречается некоторый объект. Естественным способом определения подобия между объектами i и j является определение подобия между множеством документов, содержащим объект i, и множеством документов, содержащим объект j. Будем считать такую меру, основанную на подобии множеств, теоретико-множественной. Заметим, что свойства теоретико-множественных мер изучались также в работах [Egghe, Michel, 2002], [Egghe, Michel, 2003] и [Egghe, Rousseau, 2006].

Рассмотрим три естественных свойства теоретико-множественного подобия.

Свойство 1. Если $c_{ij} = 0$, то $S(c_{ij}, s_i, s_j)$ принимает минимальное значение.

Свойство 2. Для любого $\alpha > 0$, $S(\alpha c_{ij}, \alpha s_i, \alpha s_j) = S(c_{ij}, s_i, s_j)$.

Свойство 3. Если $s_i' > s_i$ и $c_{ij} > 0$, то $S(c_{ij}, s_i', s_j) < S(c_{ij}, s_i, s_j)$.

Нетрудно показать, что данные свойства не зависят друг от друга.

Определение 3. Теоретико-множественная мера подобия – это прямая мера подобия $S(c_{ij}, s_i, s_j)$, обладающая свойствами 1–3.

Косинусная мера и индекс Жаккара обладают свойствами 1–3, т. е. являются теоретико-множественными. Индекс включения не обладает свойством 3; коэффициент ассоциативности не обладает свойством 2; таким образом, они не являются теоретико-множественными мерами. Ослабим свойство 3.

Свойство 4. Если $s_i' > s_i$ и $c_{ij} > 0$, то $S(c_{ij}, s_i', s_j) \leq S(c_{ij}, s_i, s_j)$.

Определение 4. Слабая теоретико-множественная мера подобия – это прямая мера подобия $S(c_{ij}, s_i, s_j)$, обладающая свойствами 1, 2, 4.

Таким образом, индекс включения является слабой теоретико-множественной мерой подобия.

Перечислим дополнительные свойства, которые обязательны для теоретико-множественных мер.

Свойство 5. Если $S(c_{ij}, s_i, s_j)$ принимает минимальное значение, то $c_{ij} = 0$.

Свойство 6. Если $c_{ij} = s_i = s_j$, то $S(c_{ij}, s_i, s_j)$ принимает максимальное значение.

Свойство 7. Если $S(c_{ij}, s_i, s_j)$ принимает максимальное значение, то $c_{ij} = s_i = s_j$.

Свойство 8. Для любого $\alpha > 0$, если $c_{ij} < s_i$ или $c_{ij} < s_j$, то $S(c_{ij} + \alpha, s_i + \alpha, s_j + \alpha) > S(c_{ij}, s_i, s_j)$.

Очевидно, что свойства 5–8 следуют из свойств 1, 2 и 3.

Утверждение 1. Все теоретико-множественные меры подобия $S(c_{ij}, s_i, s_j)$ обладают свойствами 5–8.

Рассмотрим еще два свойства.

Свойство 9. Если $s_i' s_j' > s_i s_j$ и $c_{ij} > 0$, то $S(c_{ij}, s_i', s_j') < S(c_{ij}, s_i, s_j)$. Если $s_i' s_j' = s_i s_j$, то $S(c_{ij}, s_i', s_j') = S(c_{ij}, s_i, s_j)$.

Свойство 10. Если $s_i' + s_j' > s_i + s_j$ и $c_{ij} > 0$, то $S(c_{ij}, s_i', s_j') < S(c_{ij}, s_i, s_j)$. Если $s_i' + s_j' = s_i + s_j$, то $S(c_{ij}, s_i', s_j') = S(c_{ij}, s_i, s_j)$.

Свойства 9 и 10 более строгие, чем свойство 3, которое вытекает из 9 и 10. Косинусная мера обладает свойством 9, а индекс Жаккара – свойством 10.

Утверждение 2. Все теоретико-множественные меры подобия $S(c_{ij}, s_i, s_j)$, обладающие свойством 9, монотонно связаны с косинусной мерой.

Утверждение 3. Все теоретико-множественные меры подобия $S(c_{ij}, s_i, s_j)$, обладающие свойством 10, монотонно связаны с индексом Жаккара.

Свойство 11. Если $\min(s_i', s_j') > \min(s_i, s_j)$ и $c_{ij} > 0$, то $S(c_{ij}, s_i', s_j') < S(c_{ij}, s_i, s_j)$. Если $\min(s_i', s_j') = \min(s_i, s_j)$, то $S(c_{ij}, s_i', s_j') = S(c_{ij}, s_i, s_j)$.

Утверждение 4. Все слабые теоретико-множественные меры подобия $S(c_{ij}, s_i, s_j)$, обладающие свойством 11, монотонно связаны с индексом включения.

Теперь определим понятие обобщенной меры подобия.

Определение 5. Обобщенный индекс подобия (*Generalized similarity index*) определяется равенством

$$S_G(c_{ij}, s_i, s_j; p) = 2^{1/p} \times c_{ij} / (s_i^p + s_j^p)^{1/p},$$

где p – параметр, принимающий значения на множестве $\mathbb{R}\backslash\{0\}$.

Данный индекс интересен тем, что при различных значениях p она сходится к известной теоретико-множественной мере (слабой или нет). Например, при $p \to 0$ он стремится к косинусной мере.

Утверждение 5. Для всех конечных значений параметра p обобщенный индекс подобия является теоретико-множественной мерой.

Иными словами, обобщенная мера определяет весь класс теоретико-множественных мер. Каждая мера из этого класса

соответствует определенному значению p. Если $p \to \pm \infty$, мера становится не теоретико-множественной, а слабой теоретико-множественной.

5.3.3. *Вероятностные меры подобия.* Как сказано выше, коэффициент ассоциативности не является теоретико-множественной мерой, поскольку для него не выполняется условие 2. Рассмотрим два свойства прямых мер.

Свойство 12. Если $s_1 = s_2 = \ldots = s_n$, то $S(c_{ij}, s_i, s_j) = \alpha c_{ij}$ для всех $i \neq j$ и для некоторого $\alpha > 0$.

Свойство 13. Для всех $\alpha > 0$ $S(\alpha c_{ij}, \alpha s_i, s_j) = S(c_{ij}, s_i, s_j)$.

Определение 6. Вероятностная мера подобия определяется как прямая мера $S(c_{ij}, s_i, s_j)$, для которой выполняются свойства 12, 13.

Утверждение 6. Все вероятностные меры подобия пропорциональны коэффициенту ассоциативности.

Следствие. Прямая мера подобия не может быть одновременно (слабой или не слабой) теоретико-множественной мерой подобия и вероятностной мерой подобия.

Данное следствие показывает, что имеется существенное различие между мерами этих типов, т. е. между косинусной мерой, индексом включения, индексом Жаккара с одной стороны и коэффициентом ассоциативности – с другой.

Выясним, что представляют собой свойства 12, 13. Меры подобия применяются к совместно-встречаемым данным по следующим причинам. Совместная встречаемость может рассматриваться как результат двух независимых явлений, которые мы называем эффектом подобия и эффектом размера. Эффект подобия – это эффект, когда при прочих равных более подобные объекты чаще совместно встречаются. Эффект размера – это эффект, когда при прочих равных наиболее часто встречаемый объект

чаще совместно встречается с другими объектами. Следовательно, на меру подобия влияет размер, что разрушает смысл понятия (см. также [Waltman, Van Eck, 2007]). Возникает необходимость откорректировать меру, для того чтобы ослабить влияние размера.

Свойство 12 касается поведения прямой меры подобия в том исключительном случае, когда объекты встречаются с одинаковой частотой. Тогда эффект размера влияет на все объекты одинаково, иными словами, совместная встречаемость становится адекватной мерой, т. е. прямая мера подобия не трансформирует совместную встречаемость существенным образом.

Рассмотрим на примере свойство 13. Предположим, что i – произвольный объект и количество вхождений объекта удваивается. Можно ожидать, что приблизительно так же увеличится количество совместных вхождений. Предположим, что оно действительно увеличилось в два раза и что это совместное вхождение распределяется между другими объектами так же, как вначале. Это означает, что удвоилась совместная встречаемость с каждым объектом. Надеемся, что такое удвоение не повлияло на меру подобия, так как увеличение является пропорциональным. Следовательно, увеличение количества совместной встречаемости с другими объектами зависит только от эффекта размера и не зависит от эффекта подобия. С учетом этого естественно предположить, что подобие объекта i с другими объектами не изменилось. Свойство 13 реализует данную идею. Это касается любого пропорционального увеличения количества встречаемости и совместной встречаемости. Данное свойство отмечено, например, в работе [Van Eck, Waltman, 2008].

В табл. 5.3.1 суммированы сведения о том, какими свойствами обладают вышерассмотренные меры.

Таблица 5.3.1

Меры подобия и их свойства

Свойство	S_A	S_C	S_I	S_J
1	×	×	×	×
2		×	×	×
3	×	×		×
4	×	×	×	×
5	×	×	×	×
6		×	×	×
7	×	×		×
8		×		×
9	×	×		
10				×
11			×	
12	×	×	×	
13	×			

Примечание. Если мера обладает свойством, то в соответствующей ячейке стоит знак “×”.

Проанализируем вероятностную меру подобия (см. также [Luukkonen, et al., 1992], [Lecler, Gagne, 1994], [Zitt, et al., 2000]). Пусть p_i – вероятность того, что объект i встречается в документе, выбранном случайным образом. Ясно, что $p_i = s_i / m$. Если два объекта i и j встречаются независимо друг от друга, то вероятность того, что они встречаются одновременно в случайно выбранном документе $p_{ij} = p_i p_j$. Ожидаемое количество совместной встречаемости $e_{ij} = mp_{ij} = mp_i p_j = s_i s_j / m$. Для того чтобы измерить подобие между этими объектами, целесообразно подсчитать отношение между обозреваемым количеством совместной встречаемости, с одной стороны, и ожидаемым количеством –

с другой, при условии что встречаемости объектов независимы. Подобные аргументы приведены в более общем контексте в работе [Price, 1981].

Мера c_{ij} / e_{ij} имеет прямую вероятностную интерпретацию. Если $c_{ij} / e_{ij} > 1$, объекты встречаются более часто, чем ожидается. Если $c_{ij} / e_{ij} < 1$, то объекты встречаются более редко, чем ожидалось. Очевидно, что $c_{ij} / e_{ij} = m\, S_A(c_{ij}, s_i, s_j)$, т. е. мера c_{ij} / e_{ij} пропорциональна коэффициенту ассоциативности и, следовательно, принадлежит классу вероятностных мер подобия. Поскольку все вероятностные меры пропорциональны друг другу (утверждение 6), они могут иметь ту же вероятностную интерпретацию.

5.3.4. *Обсуждение*. Эмпирический анализ мер подобия показал, что имеется значительное различие между прямыми мерами, поэтому важно правильно выбрать нужную меру для практического применения. Покажем, как это следует делать в случаях, когда мера подобия используется в целях нормализации.

Как сказано выше, различаются непрямые и прямые меры подобия. В качестве непрямых мер часто используются косинусная мера, а также *Bhattacharyya distance* и *Jensen-Shannon divergence*, имеющие хорошие теоретические свойства, описанные в работе [Van Eck, Waltman, 2008]. Популярным, особенно для анализа коцитирования, является также коэффициент корреляции Пирсона [Википедия] (см., например, [McCain, 1990], [White, Griffith, 1981]), однако данная мера не имеет достаточных теоретических обоснований [Van Eck, Waltman, 2008], так же как и критерий хи-квадрат.

Прямые и непрямые меры имеют фундаментальные различия, их использование приводит к существенно различающимся

результатам [Schneider, et al., 2009]. Авторы работы [Van Eck, Waltman, 2009] считают, что прямые меры ближе к интуитивной идее подобия. Рассмотрим два объекта, не встречающиеся совместно вообще, но имеющие аналогичные профили совместной встречаемости. Значение прямой меры подобия будет "низким", а непрямой – "высоким". Однако заметим, что непрямые меры также могут иметь преимущество перед прямыми мерами, поскольку они вычисляются на большем объеме данных и таким образом вовлекают меньше статистической неопределенности.

Целью использования прямых мер подобия в применении к совместной встречаемости является нормализация данных. Выясним, какую меру для этого выбрать. Авторы работы [Van Eck, Waltman, 2009] считают, что следует нормализовать с использованием вероятностной меры. Как сказано выше, вероятностная мера корректирует эффект размера, это следует из свойства 13. Теоретико-множественные меры данным свойством не обладают и не корректируют должным образом эффект размера. Мера в среднем увеличивается с увеличением числа вхождений. Это подтверждается эмпирически. Более того, если проводить анализ на основе этих мер и с помощью многомерного шкалирования или иерархической кластеризации, следует обращать особое внимание на возможные несоответствия в результатах.

Рассмотрим аргументы в пользу вероятностных мер. Предположим, что мы выполняем анализ совместной встречаемости слов и хотим определить подобие между двумя словами i и j. Рассмотрим гипотетические сценарии 1 и 2. Параметры сценариев приведены в табл. 5.3.2.

Таблица 5.3.2

Два гипотетических сценария и их параметры

Параметры и меры	**Сценарий 1**	**Сценарий 2**
m	1000	1000
s_i	300	20
s_j	300	20
c_{ij}	90	6
Коэффициент ассоциативности	0,001	0,015
Косинусная мера	0,300	0,300
Индекс включения	0,300	0,300
Индекс Жаккара	0,176	0,176

В сценарии 1 слова i и j совместно встречаются чаще, чем в сценарии 2. Однако в обоих случаях относительное пересечение множеств одинаково. В обоих сценариях слово i встречается в 30 % документов, в которых встречается j, и, наоборот, j встречается в 30 % слов, где встречается i. Поскольку относительный размер пересечения одинаков, рассматриваемые прямые теоретико-множественные меры порождают одинаковое значение для подобия. Это следует из свойства 2 и на первый взгляд кажется правильным. Однако это далеко от естественного положения дел, по крайней мере при нормировании.

Рассмотрим сценарий 1. Поскольку слова статистически независимы и каждое встречается в 30 % слов, ожидаемое значение совместной встречаемости – 30 % × 30 % = 9 %. Это следует из табл. 5.3.2, и можно предположить, что слова действительно независимы. В сценарии 2 каждое из слов встречается в 2 % документов. При условии независимости полагается, что они совместно встретятся приблизительно в 0,04 % всех документов. Однако они совместно встречаются в 0,6 % документов, т. е. в 15 раз чаще, чем ожидается.

Очевидно, что теоретико-множественные меры отражают не различие между сценариями 1 и 2. Они отражают относительное перекрытие множеств, а не отклонение, базирующееся на статистической независимости. Таким образом, целесообразно использовать вероятностную меру, причем любую, так как эти меры пропорциональны друг другу.

5.4. Функции сходства и расстояния

В основе метода изучения подобия объектов лежат две функции подобия – расстояние и сходство. Ниже приведены их определения согласно гл. 1 монографии [Дюран, Оделл, 1977]. Пусть множество $I=\{I_1, I_2,\ldots, I_n\}$ обозначает n объектов, принадлежащих некоторой популяции P. Предположим также, что существует некоторое множество наблюдаемых характеристик $C = (C_1, C_2,\ldots,C_p)^{\mathrm{T}}$, которыми обладает каждый объект из I (здесь т − транспонирование). Наблюдаемые количественные характеристики будем называть измерениями. Результат измерения i-й характеристики объекта I_j обозначим символом x_{ij}, а вектор $X_j=(x_{1j,}\, x_{2j,\,\ldots,}\, x_{pj})$ будет соответствовать каждой последовательности измерений j-го объекта. Таким образом, для каждого множества объектов I существует множество векторов измерений $X=\{X_1, X_2,\ldots X_n\}$, которые описывают множество I. Заметим, что множество X можно представить как n точек в p-мерном евклидовом пространстве $\mathbb{E}^p$.

Сформулируем задачу КА. *Пусть* $m<n$ – целые числа. Задача КА заключается в том, чтобы на основе данных, содержащихся в множестве X, разбить множество объектов I на m кластеров (подмножеств) $P_1, P_2,\ldots, P_m$, так чтобы каждый объект I_i принадлежал одному и только одному подмножеству разбиения и чтобы объекты, принадлежащие одному и тому же кластеру, были

аналогичными (сходными), в то время как объекты, принадлежащие разным кластерам, были разнородными (несходными).

Решением задачи КА является разбиение, удовлетворяющее некоторому критерию оптимальности. Этот критерий может представлять собой некоторый функционал, выражающий уровни “желательности” различных разбиений и называемый целевой функцией. Для решения задачи КА необходимо количественно определить понятия сходства и разнородности. Что означает выражение “два объекта I_j и I_k различны”? Задача была бы решена, если бы i-й и j-й объекты попадали в один и тот же кластер всякий раз, когда расстояние (отдаленность) между соответствующими точками X_i и X_j было бы “достаточно малым”, и, наоборот, попадали в разные кластеры, если бы расстояние между точками X_i и X_j было бы “достаточно большим”. Таким образом, необходимо рассмотреть понятие расстояния между точками X_i и X_j из E^p с абстрактных позиций.

5.4.1. *Функция расстояния.*

Определение 1. Неотрицательная вещественная функция $d(X_i, X_j)$ называется функцией расстояния (метрикой), если:

а) $d(X_i, X_j) \geq 0$ для всех X_i и X_j из E^p;

б) $d(X_i, X_j) = 0$ тогда и только тогда, когда $X_i = X_j$;

в) $d(X_i, X_j) = d(X_j, X_i)$;

г) $d(X_i, X_j) \leq d(X_i, X_k) + d(X_k, X_j)$, где X_j, X_i и X_k – любые три вектора из E^p.

Значение $d(X_i, X_j)$ для заданных X_i и X_j называется расстоянием между X_i и X_j и оно эквивалентно расстоянию между I_i и I_j в соответствии с выбранными характеристиками $(C_1, C_2,\ldots,C_p)^{\mathrm{T}}$.

В табл. 5.4.1 приведены некоторые наиболее употребительные функции расстояния: популярная евклидова метрика; метрики l_1

и l_p, простые с вычислительной точки зрения; метрика супремум-норма, которая также легко вычисляется и включает процедуру упорядочивания.

Таблица 5.4.1

Некоторые функции расстояния

Название	Формула
1. Евклидово расстояние	$d_2(X_i, X_j) = \left[\sum_{k=1}^{p} (x_{ki} - x_{kj})^2 \right]^{1/2}$
2. l_1-норма (манхэттенское расстояние)	$d_1(X_i, X_j) = \left[\sum_{k=1}^{p} \left\| x_{ki} - x_{kj} \right\| \right]$
3. Супремум-норма (чебышевское расстояние)	$d_\infty(X_i, X_j) = \sup_{k=1,2,\ldots,p} \left\{ \left\| x_{ki} - x_{kj} \right\| \right\}$
4. l_p-норма (расстояние Минковского)	$d_p(X_i, X_j) = \left[\sum_{k=1}^{p} \left\| x_{ki} - x_{kj} \right\|^p \right]^{1/p}$

Представим n измерений $X_1, X_2, \ldots, X_n$ в виде матрицы данных размером $p \times n$:

$$X = \begin{pmatrix} x_{11} & x_{12} & \ldots & x_{1n} \\ x_{21} & x_{22} & \ldots & x_{2n} \\ \ldots & \ldots & & \ldots \\ x_{p1} & x_{p2} & \ldots & x_{pn} \end{pmatrix} = \left(X_1, X_2, \ldots, X_n \right).$$

Аналогично расстояния между парами векторов $d(X_i, X_j)$ можно представить в виде симметричной матрицы расстояний:

$$D = \begin{pmatrix} 0 & d_{12} & \ldots & d_{1n} \\ d_{21} & 0 & \ldots & d_{2n} \\ \ldots & \ldots & & \ldots \\ d_{n1} & d_{n2} & \ldots & 0 \end{pmatrix}.$$

Заметим, что диагональные элементы $d_{ii} = 0$ (i=1, 2,…,n).

5.4.2. *Функция сходства*. Понятием, противоположным расстоянию $d(X_i, X_j)$, является понятие сходства между двумя объектами I_i и I_j.

Определение 2. Неотрицательная вещественная функция $s(X_i, X_j) = s_{ij}$ называется мерой сходства, если

а) $0 \le s(X_i, X_j) < 1$ для всех $X_i \ne X_j$;

б) $s(X_i, X_i) = 1$;

в) $s(X_i, X_j) = s(X_j, X_i)$.

Пары значений мер сходства можно объединить в матрицу сходства

$$S = \begin{pmatrix} 1 & s_{12} & \dots & s_{1n} \\ s_{21} & 1 & \dots & s_{2n} \\ \dots & \dots & & \dots \\ s_{n1} & s_{n2} & \dots & 1 \end{pmatrix}.$$

Назовем величину s_{ij} коэффициентом сходства. Если каждый вектор измерения X_i состоит из нулей и единиц (бинарные данные), то эту величину называют парным коэффициентом сопряженности.

Для заданных бинарных векторов X_i и X_j введем следующие обозначения: n_{IJ} – число характеристик, которые соответствуют единицам в векторах X_i и X_j; n_{ij} – число характеристик, соответствующих нулям в этих векторах; n_{iJ} – число характеристик, дающих нуль в X_i и единицу в X_j; n_{Ij} – число характеристик, дающих нуль в X_j и единицу в X_i.

Таким образом, $n_J = n_{IJ} + n_{iJ}$ равно числу единиц в X_j, а $n_j = n_{Ij} + n_{ij}$ – числу нулей в X_j. Пример коэффициента сходства: $n_{IJ} / n_{IJ} + n_{Ij} + n_{iJ.}$ Обсуждение подобных коэффициентов приведено в [Sokal, Sneath, 1963].

В статистике используется мера линейного сходства r_{ij}, называемая коэффициентом корреляции. Следует отметить, что если X_i и X_j рассматривать как координаты точек в пространстве $\mathbb{E}^p$, являющихся концами двух векторов с началом в начале координат, то согласно работе [Андерсон, 1963].

$$r_{ij} = \cos\theta = \left[\sum_{k=1}^{p} x_{ki} x_{kj} \right] \Big/ \left[\sum_{k=1}^{p} x_{ki}{}^2 \sum_{k=1}^{p} x_{kj}{}^2 \right]^{\frac{1}{2}}, \qquad (5.4.1)$$

где θ – угол между векторами. Из равенства (5.4.1) следует, что $-1 \le r_{ij} \le 1$.

Будем говорить, что объекты I_i и I_j сходны положительным образом (положительно), если значение r_{ij} “близко” к 1, отрицательно сходны, если значение r_{ij} “близко” к –1, и несходны, если значение r_{ij} “близко” к нулю. Заметим, что r_{ij} не является функцией сходства, так как не выполняется аксиома а) из определения 2.

К выбору коэффициента сходства следует подходить очень осторожно. Рассмотрим пример, кода точки X_1 и X_2 находятся сравнительно “далеко” друг от друга, в то время как их сходство, измеряемое с помощью r_{ij}, равно единице (рис. 5.4.1).

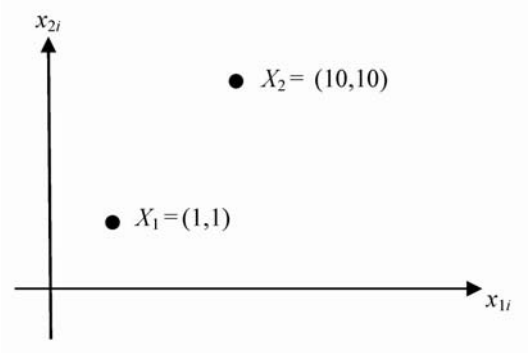

Рис. 5.4.1. Две точки в $\mathbb{E}^2$

Используя метрики (1), (2), (3) из табл. 5.4.1, получаем

$$d_2(X_1, X_2) = [(10 - 1)^2 + (10 - 1)^2]^{1/2} = 9\sqrt{2};$$

$$d_1(X_1, X_2) = [|10 - 1| + |10 - 1|] = 18;$$

$$d_\infty(X_1, X_2) = \sup [|10 - 1|, |10 - 1|] = 9.$$

В данном случае коэффициент корреляции $r_{12} = 1$. Несмотря на то что $X_1 \neq X_2$, с точки зрения критерия r_{12} объекты I_1 и I_2 будут считаться сходными.

5.5. Два приема кластеризации

5.5.1. *Иерархическая кластеризация.* Иерархическая кластеризация делится на объединяющую (*Agglomerative hierarchical clustering*) и разделяющую (*Divisive hierarchical clustering*). Алгоритм объединяющей кластеризации группирует единицы данных путем последовательных объединений, начиная с кластеров, каждый из которых содержит единичный объект, и заканчивая кластером, содержащим все объекты. Алгоритмы разделяющей кластеризации начинают с кластера, содержащего все объекты, последовательно разделяя их до единичных кластеров.

Поясним процедуру объединяющей иерархической кластеризации на следующем примере. Рассмотрим n объектов, где каждый объект соответствует своему кластеру. Затем два наиболее близких объекта объединяются в один кластер и число кластеров становится равным $n-1$. Процесс повторяется до тех пор, пока все n объектов не попадут в один кластер, содержащий все объекты. Следует отметить, что задачи КА можно формулировать и решать как в терминах матрицы расстояний, так и в терминах матрицы сходства.

Один из известных методов представления матрицы расстояний D и матрицы сходства S основан на идее “дерева” (дендрограммы, *Dendrogram*), с помощью которой можно отразить ре-

зультаты последовательной кластеризации при условии, что эта процедура оперирует только с элементами указанных матриц.

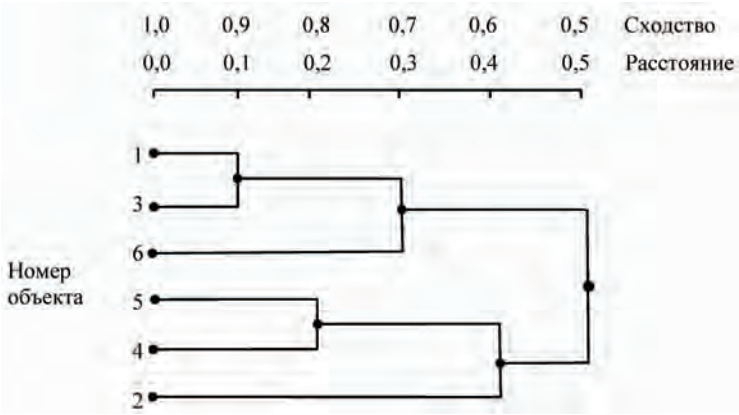

Рис. 5.5.1. Пример дендрограммы

На рис. 5.5.1 представлен случай шести объектов ($n = 6$). Объекты 1 и 3 наиболее близки (наименее удалены друг от друга) и поэтому объединяются в один кластер на уровне сходства, равном 0,9. Объекты 4 и 5 объединяются на уровне сходства 0,8. На этом шаге имеется четыре кластера: (1, 3); (6); (4, 5) и (2). На третьем и четвертом шаге образуются кластеры (1, 3, 6) и (5, 4, 2), соответствующие уровням сходств 0,7 и 0,6. Окончательно все объекты группируются в один кластер на уровне 0,5.

Неотъемлемой частью задачи КА является понятие “оптимального критерия”, т. е. целевой функции, позволяющей установить, когда достигается желательное разбиение. Для введения подобного критерия необходимо определить меру внутренней однородности кластера и меру разнородности кластеров, например расстояние между ними.

Пусть $I=\{I_1, I_2,\ldots,I_{n1}\}$ и $J=\{J_1, J_2,\ldots,J_{n2}\}$ – два кластера; $X=\{X_1, X_2,\ldots,X_{n1}\}$ и $Y=\{Y_1, Y_2,\ldots,Y_{n2}\}$ – множества измерений характеристик , соответствующие I и J.

Определение 1. Обозначим через

$$D = \left\{ \; d(X_i,\; X_j),\; i = 1,\ldots,n_1;\;\; j = 1,\ldots,n_2 \; \right\}$$

множество всех расстояний. Величину

$$D_1(I,J) = \min_{\substack{i=1,\ldots,n_1 \\ j=1,\ldots,n_2}} d(X_i, X_j)$$

будем называть минимальным локальным расстоянием (*Nearest neighbor distance*) между кластерами *I, J*, соответствующим данной функции расстояния *d*.

Определение 2. Пусть $D = \left\{ \; d(X_i,\; X_j) \; \right\}$ определено так же, как в определении 1. Тогда

$$D_2(I,J) = \max_{\substack{i=1,\ldots,n_1 \\ j=1,\ldots,n_2}} d(X_i, X_j)$$

назовем максимальным локальным расстоянием (*Furthest neighbor distance*) между *I, J*.

Определение 3. Величина

$$D_3 = \sum_{j=1}^{n_2} \sum_{i=1}^{n_1} d(X_i, X_j) \big/ (n_1 n_2)$$

есть “среднее расстояние” (*Mean distance*) между кластерами I и J, соответствующее данной функции расстояния *d*.

Определение 4. Величину

$$D_4 = \frac{n_1 n_2}{n_1 + n_2} \left(\; \overline{X} - \overline{Y} \; \right)^{\mathrm{T}} \left(\; \overline{X} - \overline{Y} \; \right),$$

где

$$\overline{X} = \sum_{i=1}^{n_1} x_i / n_1, \quad \overline{Y} = \sum_{i=1}^{n_2} y_i / n_2 \, ,$$

называют статистическим расстоянием (*Statistical distance*) между кластерами I и J. Существует множество объединяющих иерар-

хических алгоритмов, основанных на различных определениях дистанции между кластерами. Примерами могут служить алгоритмы кластеризации, основанные на методах: “одиночной связи” (*Single – linkage clustering*); “полной связи” (*Complete – linkage clustering*); “средней связи” (*Average – linkage clustering*).

Рассмотрим метод одиночной связи, в котором функция расстояния между кластерами X и Y, описывается выражением (см. Определение 1):

$$D(X,\ Y) = \min_{x \in X,\, y \in Y} d(x, y),$$

где $d(x, y)$ – расстояние между элементами x и y. Кластеры объединяются на основе единичных элементов, в то время как другие расстояния между элементами могут быть большими.

Построим простейшую объединяющую схему, которая удаляет ряды и колонки матрицы подобия, соответствующие объединяемым кластерам, и добавляет новые ряды и колонки, просчитывая все расстояния до оставшихся “старых” кластеров. Пусть исходная матрица расстояний имела размерность $N \times N$. Этапам кластеризации последовательно присваиваются номера от 0 до $N - 1$. Обозначим через $L(k)$ уровень k-й кластеризации. Кластер с последовательным номером m обозначим через (m). Расстояние между кластерами обозначим через $d[(r), (s)]$. Алгоритм имеет следующий вид.

Шаг 1. Для первичной матрицы $m = 0$, $L(m) = 0$.

Шаг 2. Найти наиболее близкую пару кластеров. Допустим, это (r), (s), соответственно $d[(r), (s)]$ минимально по всем парам векторов, имеющимся на этом этапе кластеризации.

Шаг 3. Увеличить номер последовательности $m = m+1$. Установить уровень $L(m) = d[(r), (s)]$.

Шаг 4. Обновить матрицу подобия, удалив ряды и колонки, соответствующие кластерам (r), (s), и добавив ряд и колонку, со-

ответствующие новому кластеру (*r*,*s*). Пусть (*k*) – старый кластер. Тогда *d*[(*k*), (*r*,*s*)] = min(*d*[(*k*), (*r*)], *d*[(*k*), (*s*)]).

Шаг 5. Если все объекты оказались в одном кластере, то остановка, иначе − перейти к шагу 2.

Алгоритмы иерархической кластеризации используются программой *Pajek* [Batagelj, Mrvar, 1998] для конструирования карт науки.

5.5.2. *Разделительная кластеризация*. Неиерархическая разделительная кластеризация (*Partitional clustering*) делит множество единиц данных на заранее определенное количество кластеров с помощью итерационной процедуры без выявления иерархической структуры. В качестве примера разделительной кластеризации можно привести простой и эффективный метод кластеризации к-средних (*k-means*), изобретенный Г. Штейнгаузом [Steinhaus, 1956].

Алгоритм к-средних стремится минимизировать суммарное квадратичное отклонение объектов от центров кластеров. Для этого на каждой итерации пересчитывается центр масс каждого кластера, полученного на предыдущем шаге. Затем объекты вновь разделяются на кластеры в зависимости от того, какой из новых центров оказался ближе в соответствии с выбранной метрикой.

Пусть имеется n единиц данных, представленных векторами $X_1, X_2, \ldots, X_n$ размерности d. Алгоритм стремится разделить множество измерений на k кластеров ($k \le n$) $\{C_1, C_2, \ldots C_k\}$, так чтобы минимизировать суммарное квадратичное отклонение точек кластеров от центров этих кластеров:

$$\arg\min \sum_{i=1}^{k} \sum_{X_j \in C_i} || X_j - \mu_i ||^2$$

(μ_i – центр масс векторов $X_j \in C_i$). Как правило, алгоритм использует итеративную технику и состоит из следующих шагов.

Шаг 1. Инициализация. Первоначальное разделение объектов на k кластеров произвольным образом или на основе предварительного знания.

Шаг 2. Подсчет центров масс для каждого кластера, т. е. для каждого C_i, $i \in \{1, 2, \ldots, k\}$,

$$\mu_i = \frac{1}{|C_i|} \sum_{X_j \in C_i} X_j .$$

Шаг 3. Новое разбиение на кластеры, каждый вектор попадает в кластер, центр которого ближе всего к вектору, т.е.

$$X_j \in C_w, \text{ если } \| X_j - \mu_w \| < \| X_j - \mu_i \|$$

для всех $j=1,\ldots,n$, $i \neq w$ и $i=1,\ldots,k$.

Шаг 4. Повторять шаги 2 и 3, до тех пор пока не сложится ситуация, когда не изменилось содержимое ни одного кластера.

Таким образом, алгоритм завершается, когда на какой-то итерации прекращается изменение содержимого кластеров. Это происходит за конечное число итераций, так как количество возможных разбиений конечного множества конечно, а на каждом шаге суммарное квадратичное отклонение уменьшается. Время выполнения одной итерации оценивается как $O(nkd)$.

Назовем некоторые проблемы алгоритма *к*-средних. Во-первых, не существует эффективного и универсального метода первоначального разделения на кластеры и определения их числа. Результат зависит от выбора исходных центров, а их оптимальный выбор не известен. Кроме того, число кластеров необходимо задавать заранее. Во-вторых, алгоритм не гарантирует достижения глобального минимума, а только одного из локальных минимумов. В-третьих, алгоритм чувствителен к выбросам. Несмотря на то что могут существовать объекты, находящиеся далеко от центров всех кластеров, объект приписывается к какому-либо “ближайшему” кластеру.

5.6. Кластеризация *SCI*

В работе [Small, Sweeney, 1985] показана эволюция методов кластеризации *SCI*, используемых для определения исследовательских фронтов (*Research fronts*) и подготовки карт и атласов науки. Передовые исследования отражают узкую научную специализацию и выявляются методами кластерного анализа на основе коцитирования.

Первый эксперимент по кластеризации *SCI* был проведен в 1974 г. на основе данных 1972 г. [Small, Griffith, 1974], [Griffith, et al., 1974]. Сначала составлялся перечень высокоцитируемых (основополагающих) работ за текущий год, по которому устанавливался порог "активности". Затем устанавливался порог "коцитирования", после чего проводилась иерархическая кластеризация методом одиночной связи, основанным на этих целочисленных порогах. Первый уровень содержал все документы, которые были коцитированы хотя бы с одним документом, таким образом, не коцитируемые работы далее не рассматривались. В результате были сформированы фронты исследований в период с 1970 по 1974 гг. [Garfield, et al., 1978].

Применение "волюнтаристского" подхода к определению порога коцитирования приводит к тому что, во-первых, в некоторых областях науки, например биомедицине, высокоцитируемые работы, как правило, связывают вместе очень большое количество работ; во-вторых, высокоцитируемые работы имеют тенденцию быть и высококоцитируемыми, что создает "перекос" при анализе научных областей меньшего масштаба.

Начиная с 1975 г. была введена процедура нормализации данных коцитирования в качестве способа частичного преодоления проблемы высокоцитируемых публикаций и зависимости от размеров областей исследования. Первоначально для нормализации данных использовалась мера сходства Жаккара, затем косинусная

мера сходства Солтона, так как она эффективно разделяет высоко- и слабоцитируемые работы. С использованием нормализованной меры коцитирования стало возможным получать больше кластеров на том же уровне коцитирования и точнее оценивать “большие” и “малые” области.

В период с 1970 по 1980 г. кластеризация в *SCI* проводилась с использованием нормализованных мер коцитирования и порогом цитирования от 15 до 17 цитирований на документ. Пороги нормализованных коцитирований выбирались с учетом получения как можно большего числа кластеров для исходных данных. Наибольшие кластеры содержали порядка 100 цитируемых документов.

Однако проблема наличия порога давала себя знать в части репрезентативности кластеров для различных научных областей. Было сложно получить адекватное представление в таких областях, как математика и техника, наряду с биологией и физикой, которые имели более высокие пороги цитирования. Естественно, считать, что количество кластеров в представлении должно быть пропорционально представительству цитирующих источников. До этого момента пороги цитирования устанавливались в зависимости от того, вся ли база была объектом анализа или некоторая ее часть. При этом все цитирования считались равными, каждое имело значение 1. Появилась идея использовать дробные цитирования, зависящие от количества работ, цитируемых каждым автором. Если в списке ссылок 10 публикаций, то каждое цитирование имеет вес, равный 1/10. Соответственно изменяется количество цитирований документа, которое уже будет суммой взвешенных цитирований. Заметим, что дробное цитирование используется только на начальном этапе выбора высокоцитируемых документов.

Еще одним методологическим приемом является ограничение количества кластеров на каждом уровне кластеризации. Извест-

но, что наиболее просто реализуется кластеризация методом одиночной связи, если устанавливать порог для объединения и формировать все группы, имеющие прочность связи, большую или равную величине порога. Если кластер превосходит размеры, установленные на данном уровне, следует повысить уровень и снова провести кластеризацию. Большой кластер разделится на более мелкие, но поскольку можно устанавливать минимальный начальный порог, то небольшие кластеры могут стать больше.

Итак, для проведения кластеризации требуется определить три параметра: максимальный размер кластера, стартовый уровень для нормализованных коцитирований и уровень приращения, который указывает, насколько должен увеличиться коэффициент коцитирования, в случае если кластер получился слишком большим. Такой метод существует и называется кластеризацией с варьированием уровней (*Variable level clustering*) (рис. 5.6.1). Здесь кластеры представлены в виде прямоугольников, а шкала для нормализованных коцитирований варьируется от 0,1 до 0, 4. Предположим, что на уровне 0,1 разрешены только кластеры, имеющие размер 50, и уровень приращения коэффициента равен 0,1. На стартовом уровне 0,1 имеем кластер *A*, содержащий 100 документов. Следующая попытка произвести кластеры меньшего размера будет проводиться на уровне 0,2. Получаем два кластера: кластер *C* допустимого размера и кластер *B*, количество элементов которого больше дозволенного. Поднимаем уровень до 0,3. Из кластера *B* получаем два кластера: *D* и *E*. На следующем уровне кластер *D* вновь делим на два кластера. В итоге получаем множество кластеров {*C*, *E*, *G*, *F*}, которые представляют собой разукрупнение кластера *A* уровня 0,1. Нормализованное коцитирование для публикаций i и j определяется как $C_{ij} \Big/ \left(C_i \times C_j \right)^{\frac{1}{2}}$,

где C_{ij} – коэффициент коцитирования публикаций i и j; C_i – количество цитирований публикации i (целое число); C_j – количество коцитирований публикации j (целое число).

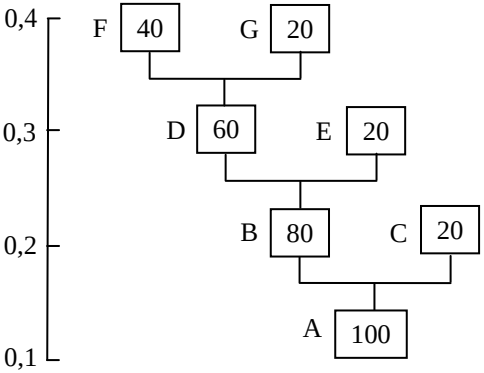

Рис. 5.6.1. Кластеризация с варьированием уровней (буквами обозначены метки кластеров, цифрами – число высокоцитируемых публикаций в кластере)

Эксперимент показал, что дробный коэффициент цитирования является перспективной стратегией для устранения имеющегося перекоса в сторону биомедицины, который получается вследствие введения порогов как базиса для кластеризации коцитирований. Потенциальной проблемой остается то, что чрезмерный вес достается цитирующим работам с небольшим количеством ссылок. Однако из эксперимента выяснилось, что часто такие работы не оказывают влияния на кластеризацию. В целом использование дробных счетчиков цитирования и кластеризации с варьированием уровней улучшило результаты кластеризации междисциплинарных баз данных, таких как *SCI*.

5.7. Кластеризация на основе встречаемости общих слов

В конце XX в. во Франции ряд исследователей занимался разработкой методов лексического анализа, основанных на одновре-

менном появлении в публикациях терминов, ключевых слов и предметных заголовков (см., например, [Callon, et al., 1983], [Turner, et al., 1988], [Law, Whittaker, 1992], [Courtial, 1994]). Был разработан подход, основанный на частотном анализе совместной встречаемости ключевых слов, выбранных из заголовка, краткого содержания или текста. Данный метод развился в отдельное направление, поскольку применяется во многих областях. Кластеризация на основе совместных слов также относится к стандартной методике в *WoS*.

В работе [Lin, et al., 2007] приведена структура системы автоматической кластеризации и ранжирования множества документов, являющихся результатом извлечения информации по тематическому запросу к базе данных *Medline* поисковой машиной *PubMed*. Поисковая машина содержит механизмы для ранжирования извлеченных документов, такие как сортировка по дате, имени автора или журналу. Осуществляется также сортировка по релевантности, однако она может быть не совсем верной относительно запроса пользователя. Для того чтобы помочь исследователю быстро найти нужные публикации, требуется более совершенная система ранжирования. Кроме того, длинный неструктурированный список требует дополнительных усилий для извлечения необходимых знаний.

Предлагается после извлечения результатов запроса из ББД выполнить преобразование массива документов, заключающееся в их кластеризации и последующем ранжировании с целью распределения документов по темам и важности. Это делается следующим образом.

Подготовка данных. Процедура начинается с сопоставления каждому извлеченному документу текста аннотации, названия и списка терминов.

Первичная обработка. Аннотация делится на отдельные слова – термы. При этом от слов отрезаются суффиксы, остаются только корни. Каждый документ представляется в виде *N*-размерного вектора взвешенных термов, в данном случае это вектор

$$(tf\text{-}idf_1, tf\text{-}idf_2, \ldots, tf\text{-}idf_N),$$

где $tf\text{-}idf_i$ – это *tf-idf* вес терма i (см. п. 4.4). Затем строится матрица встречаемости, в которой столбцы соответствуют документам, а строки – термам.

Кластеризация. Матрица нормализуется с помощью косинусной нормализации (см. п. 5.3). Нормализованная матрица выступает в качестве входных данных для кластеризации. Для предварительной оценки числа кластеров использован алгоритм *Spectroscopy* (http://ink.library.smu.edu.sg/sis_research/1018). Кластеризация проводится с помощью программного продукта "*Cluto*" [Karypis, 2003] с применением варианта алгоритма кластеризации *k*-means [Steinbach, et al., 2000].

Определение меток кластеров. Для того чтобы сопоставить кластеру соответствующую тему, генерируется набор информативных слов и ключевых медицинских терминов. Данная техника подробно описана в [Lin, et al., 2007].

Ранжирование. На этом шаге определяется множество важных публикаций, соответствующих тематике кластера. В качестве метрик ранжирования используются количество цитирований за год для каждой публикации и ее импакт-фактор (оба параметра можно получить из *JCR*).

Представленная методика была применена для работы с ББД *Medline*, которая на момент проведения эксперимента содержала свыше 15 млн цитирований. По запросу "*breast cancer*" было получено всего 77784 документа, из них 65 эксперты признали важными. В результате кластеризации было получено 6 кластеров.

Методы кластеризации на основе общих слов, коцитирования и библиографического сочетания предназначены, прежде всего, для описания структуры науки и ее оценки на макроуровне и промежуточном уровне. Кластеризация на основе соавторства и анализ коцитирования авторов представляют собой библиометрическую процедуру для выявления структуры на микроуровне и промежуточном уровне [White, McCain, 1998].

Глава 6. Карты науки

> Человечество всегда мне представлялось в виде множества блуждающих в тумане огоньков, которые лишь смутно чувствуют слияние, рассеиваемое всеми другими, но связаны сетью ясных огненных нитей, каждый в одном, двух, трех направлениях…
>
> *А. Н. Колмогоров*

Теоретические и прикладные исследования в области научной картографии на основе анализа цитирований начались во второй половине XX в. Английский математик и библиотекарь С. Брэдфорд (*Samuel Clement Bradford*, 1878-1948) в своей работе [Bradford, 1948] говорит о "…картине вселенной дискурса, представленной в качестве шара, на котором во взаимосвязанном виде рассеяны отдельные вещи, которые мы видим или о которых мы думаем". Такие карты дают заслуживающее доверия представление о результатах деятельности некоего "невидимого колледжа". На этих картах области, представляющие интерес для изучения, можно представить и сравнить на нескольких уровнях; с помощью карт можно строить предположения относительно прогресса в исследованиях, достигнутого за определенный период времени.

Справка. Невидимый колледж − неинституционализированная группа исследователей, согласованно работающих над общей проблематикой. Термин, введенный Дж. Берналом, был развернут Д. Прайсом в гипотезу о "невидимых колледжах" как коммуникационных объединениях, имеющих определенную достаточно устойчивую структуру, функции и объем. [ЭЭиФН, 2009].

За десятилетие до высказывания С. Брэдфорда в работе [Bernal, 1939] был представлен практический результат, связан-

ный с построением карты науки (*Map of Science*). В этой статье сообщается о вручную созданной карте, имеющей вид таблицы. Колонки соответствуют физическому, биологическому и общественно-научному секторам науки, а ряды представляют фундаментальный и технический подходы. Карта отражает доминантное положение физических наук по отношению к биологическим и доминантное положение биологических наук по отношению к общественным наукам. По горизонтальной оси наблюдается иерархическое отличие фундаментальных наук от прикладных. “Несмотря на то что наука в те годы имела другое распределение, эти карты ближе к более поздним, чем можно ожидать” [Klavans, Boyack, 2009].

И. В. Маршакова-Шайкевич в словаре [ЭЭиФН, 2009] дает развернутое определение термина “картографирование науки”. Цитируем с сокращениями: «В эпистемологии и философии науки интерес к тематическому строению науки был вызван общей эволюцией научного знания. К 70−80 гг. XX в. фронт таких исследований значительно расширяется, интенсивное развитие получают эмпирические исследования, относящиеся к социологии науки, к наукометрии и к научным коммуникациям ... Ученые осознают, что любая теория науки не может абстрагироваться от дисциплинарного строения науки. Именно в тот период в философии науки появляются и широко обсуждаются такие понятия, как “парадигма науки”, “исследовательская программа”, “сплоченная социальная группа” и др. Параллельно с обсуждением этих концепций и моделей развития науки ряд исследований посвящен проблемам формирования и развития новых научных направлений и “картографирования науки”, которые опираются на разнообразные формы анализа цитирования и позволяют изучать тематические группы и типы научной коммуникации (Р. Мертон, Ю. Гарфилд, Г. Смолл и др.). Все эти исследования представляют

собой библиометрический подход к анализу науки и изучают когнитивные и социальные формы репрезентации знания на уровне научной специальности, опираясь на библиографические данные научных публикаций в естественных и социальных науках.»

Д. Прайс отмечал, что научная информация есть нечто гораздо большее, чем только проблема научной литературы и научных библиотек. “Научная статья отнюдь не является единицей информации, которую публикуют, накапливают, находят и выдают по требованию. Она − меняющаяся часть социальной ткани науки, и она производится в одних условиях, а используется в других”. В 1955 г. Дж. Бернал при обсуждении стратегии научных исследований писал, что “…обзоры и карты столь же необходимы в науке, как и в навигации. Они заслуживают того, чтобы им сегодня уделяли больше внимания. Основные области исследования требуют четкого указания границ, а список основных проблем данной отрасли должен пересматриваться в короткие сроки. Это особенно важно для тех отраслей прикладной науки, традиционная практика которых быстро меняется в результате научных исследований” [Бернал, 1956]. Д. Прайс был первым, кто дал представление о том, как пространственные модели науки могут быть сконструированы из библиографических данных. С годами, продвигаясь с помощью этой модели от научных журналов до статей, опубликованных в них, Д. Прайс пришел к выводу, что “карта военных действий в науке” может быть получена из данных цитирования. С библиометрической точки зрения [Small, Garfield, 1985] “карты науки служат инструментом навигации в сфере научной литературы, обозначая пространственные связи между фронтами исследований, которые представляют собой области наиболее активной деятельности. Такие карты часто используются для обозначения распределения этих

областей и придания дополнительной значимости их взаимоотношениям".

Создание карт науки относится к области визуализации информации, которая предполагает интерактивный процесс отображения абстрактных (непространственных) объектов, каковыми являются наборы библиографических данных. Проблемам визуализации информации посвящено множество статей и обзоров в ведущих журналах в области библиометрии: *Journal of the American Society for Information Science* и *Scientometics*, что свидетельствует об актуальности проблемы научной картографии.

В обзоре [Börner, et al., 2003], посвященном визуализации областей знаний, представлена методология создания карт, отражающих структуру научных дисциплин. Процесс порождения карт науки можно представить в виде последовательности следующих основных этапов: извлечение данных; выбор метрик; вычисление коэффициента подобия; присваивание координат каждой единице; визуализация.

6.1. Источники данных

В общем случае можно сказать, что карты науки состоят из элементов и отношений между ними. Такие карты отличаются от простой классификационной системы, поскольку эти элементы в большинстве случаев размещены на плоскости и между парами элементов существуют взаимосвязи, для которых, как правило, указана их сила. Элементами для построения карт могут служить экспертные оценки, библиометрические данные о публикациях, научные журналы, категории журналов (для карт дисциплин). В табл. 6.1.1 приведены примеры работ по построению карт науки из различных исходных данных и их назначение.

Таблица 6.1.1

Карты науки: исходные данные и назначение

Данные	Назначение	Карта науки
Журналы	Изучение структуры науки: относительное положение и взаимосвязь областей науки	[Bassecoulard, Zitt, 1999]
Журналы	Анализ областей науки	[Leydesdorff, 1994] [McCain, 1998] [Ding, et al., 2000]
Категории журналов	Изучение структуры науки	[Moya-Anegón, et al., 2007] [Leydesdorff, Rafols, 2008]
Документы	Изучение сети коммуникаций в науке	[Small, 1999a] [Small, 2000]
Документы	Оценка производительности исследований	[Noyons, et al., 1999] [Noyons, et al., 1999a] [Noyons, Van Raan, 1998] [Noyons, 2001]
Документы	Анализ областей науки, научный ландшафт	[Boyack, et al., 2002] [Boyack, et al., 2005]
Авторы, коцитирование	Изучение интеллектуальной структуры научной области	[White, McCain, 1998] [Ding, et al., 1999] [Chen, 1999] [Lin, Kaid, 2000] [Chen, et al., 2001]
Авторы, соавторство	Социальная интерпретация библиометрической сети	[Mahlck, Persson, 2000]
Авторы, соавторство	Исследование сотрудничества	[Glänzel, DeLange, 1997] [Glänzel, 2001]
Совместная встречаемость слов	Выявление когнитивной структуры областей науки	[DeLooze, Lemarie, 1997] [Bhattacharya, Basu, 1998] [He, 1999]

Выбор исходных данных имеет большое значение, и главным показателем в этом процессе является их полнота и достоверность. Следует руководствоваться двумя положениями. Первое положение состоит в том, что “…имея полностью обновленную

(свежую) двух-трехлетнюю базу данных, можно создать лишь такую карту, которая будет показывать, где находились фронты исследований. Вследствие задержки, обусловленной циклом публикации, исследования уходят вперед к моменту появления соответствующих журнальных статей. Такие карты не отражают общую картину исследований, по ним можно сделать самое общее предположение относительно того, в каком направлении может идти дальнейшее развитие. Однако по мере увеличения размера базы данных до десятилетия и более карта, созданная с помощью анализа цитирования, будет служить в качестве исторического, даже историографического окна в исследуемую область" [Garfield, 1993]. С глобальной точки зрения, эти карты отражают связи между областями или дисциплинами.

Второе положение касается поставщиков актуальной информации о научном цитировании, роль которых в настоящее время выполняют две международные междисциплинарные ББД: *WoS* и *Scopus*. Конечно, не следует сбрасывать со счетов ББД по конкретным областям знаний, например *Chemical Abstract Service* (химия), *MathScNet* (математика), *PubMed* (медицина), *DBLP* (информатика) и национальные ББД, в нашем случае – РИНЦ [Википедия]. Кроме статей и обзоров из научных журналов − основных источников информации – индексируются материалы конференций, диссертации, рефераты и патенты. Из каждого документа, как правило, извлекаются следующие данные: название, автор (авторы), название организации, источник публикации, ключевые слова, аннотация, список цитируемой литературы, дата публикации.

На момент написания этих строк ББД *WoS* (создана в 1961 г., принадлежит корпорации "*Thomson Reuters*") индексировала 12 200 (с точностью до сотни) научных журналов, в то время как ББД *Scopus* (создана в 2004 г. издательским домом "*Elsevier*")

индексировала 17 800 журналов; пересечение множества журналов этих ББД составляет порядка 1 100 наименований. Если рассматривать только журналы, имеющие статус “активных” (т. е. журналы, которые на данный момент индексируются в базе, в отличие от журналов, имеющих статус “пассивных”, которые индексировались ранее, но в данный момент не индексируются по каким-либо причинам), то получим иную картину. Несмотря на стремительный рост базы данных *Scopus*, база *WoS* выигрывает по объему и глубине архива. Кроме того, библиографические записи в *Scopus* до 1996 г. имеют значительные лакуны.

В подавляющем большинстве случаев для построения карт используется информация ББД *WoS* и *Scopus*, которая генерируется в виде таблиц. Отметим, что “…само по себе построение таблицы предполагает, что исследованию задано определенное направление, а об области исследования имеется какое-то предварительное представление. Таким образом, уже на этапе собирания данных делается первый шаг к абстрагированию от конкретной действительности, когда из бесконечного числа способов описывать объекты исследования выбирается один, характеризуемый выбором набора признаков, с помощью которого объекты отделяются друг от друга. За собиранием данных следует их анализ, конечная цель которого – извлечение особого рода *информации*, или, более отвлеченно, *знания* из таблицы” [Зиновьев, 2000].

6.2. Преобразование данных

В этом пункте будут кратко представлены методы преобразования данных, используемые для составления карт науки. Главной целью процесса преобразования является достижение наглядного и лаконичного описания. Применение большинства представленных ниже методов можно найти в монографии [Jain, Dubes, 1988].

Один из методов − определение внутренней структуры исследуемых данных путем кластеризации − рассмотрен в гл. 5 настоящей монографии. Далее рассматриваются приемы понижения размерности данных с целью их визуального представления в 2D- и 3D-пространствах. Значительная часть этих приемов относится к многомерному статистическому анализу − разделу математической статистики, изучающему представление, обработку и систематизацию статистических данных с целью выявления характера и структуры взаимосвязей между компонентами исследуемого многомерного признака.

Пакеты прикладных программ, такие как *SPSS*, *Statistica*, *SAS* и др., снимают трудности в применении многих из рассматриваемых методов, опирающихся на линейную алгебру, теорию вероятностей и математическую статистику.

6.2.1. *Декомпозиция на основе собственных векторов и собственных значений* (*Eigenvalue / Eigenvector Decomposition*). Пусть дана матрица A размерности $n \times n$. Если имеется вектор $\mathbf{v}$ и скалярная величина λ, такие что $A\mathbf{v} = \lambda\mathbf{v}$, то вектор $\mathbf{v}$ называется собственным вектором матрицы A, а λ – собственным значением A, соответствующим $\mathbf{v}$. Разложение на основе собственных векторов (собственных значений) уменьшает размерность при сохранении внутренней структуры.

Например, в работе [Davidson, et al., 1998] технология собственных векторов используется для минимизации функции стоимости, применяемой при определении положения точек в соответствии с коэффициентами подобия:

$$Cost = \sum_{i=1}^{n} \sum_{j=1}^{n} s_{ij} \times d_{ij}^{\,2}.$$

Здесь s_{ij} – коэффициент подобия; d_{ij} – геометрическое расстояние между объектами. Задача сводится к минимизации выражения

$$\sum_{i=1}^{n}\sum_{j=1}^{n} s_{ij} \times \left[(x_i - x_j)^2 + (y_i - y_j)^2 \right],$$

и вычислению n собственных векторов матрицы Лапласа L, определяемой как

$$L(i, j) = \begin{cases} -s_{ij}, & \text{если } i \neq j, \\ \sum_{j=1}^{n} s_{ij}, & \text{если } i = j. \end{cases}$$

6.2.2. *Метод сингулярного разложения* (*Singular value decomposition*). Метод сингулярного разложения основан на разложении матрицы, т. е. на представлении ее в виде произведения других матриц, обладающих определенными свойствами. В данном случае представляет интерес приближение исходной матрицы матрицами более низкого ранга. Такую матрицу можно получить из сингулярного разложения исходной матрицы путем простых преобразований. Применение такой техники к латентно-семантическому анализу (*Latent Semantic Analysis*) текста используется, например, в работе [Deerwester, et al., 1990].

Латентно-семантический анализ (ЛСА) [Landauer, et al., 1998] – это метод обработки информации на естественном языке, основанный на анализе взаимосвязи между коллекцией документов и встречающимися в них терминами, сопоставляющий некоторые факторы (тематики) всем документам и терминам. В качестве исходной информации ЛСА использует матрицу, столбцы которой соответствуют термам, а строки – документам. Эта матрица представляет собой набор данных, используемых для обучения системы анализа. Как правило, элементы матрицы содержат веса, учитывающие частоты использования каждого терма в каждом документе и участие терма во всех документах (*tf-idf*). Наиболее распространенный вариант метода ЛСА основан на использова-

нии разложения диагональной матрицы по сингулярным значениям. Любая вещественная прямоугольная матрица может быть разложена на произведение трех матриц: $A = U \times S \times V^T$, где U и V – ортогональные матрицы, S – диагональная матрица, значения на диагонали которой называются сингулярными значениями матрицы A. Такое разложение обладает следующей особенностью: если в матрице S оставить только k наибольших элементов, а в матрицах U и V – только соответствующие им столбцы, то произведение преобразованных матриц $A^* = U^* \times S^* \times V^{*T}$ будет наилучшим приближением матрицы A к матрице k-го ранга. Матрица A^* отражает основную структуру различных зависимостей, имеющихся в исходной матрице. Структура зависимостей определяется весовыми функциями термов [Википедия].

6.2.3. *Факторный анализ и метод главных компонент* (*Factor analysis*, *Principal component analysis*). Факторный анализ − это многомерный метод, применяемый для изучения взаимосвязей между значениями переменных. Предполагается, что известные переменные зависят от меньшего количества неизвестных переменных (факторов) и случайной ошибки. При анализе в один фактор объединяются сильно коррелирующие между собой переменные, вследствие чего происходит перераспределение дисперсии между компонентами и получается максимально простая и наглядная структура факторов. Таким образом, факторный анализ k случайных векторов выявляет m ($m < k$) общих для всех исходных величин факторов, объясняя оставшуюся после этого дисперсию влиянием специальных факторов. Корреляция компонент внутри каждого фактора между собой выше, чем их корреляция с компонентами из других факторов. Можно выделить две цели факторного анализа: 1) определение взаимосвязей

между переменными (классификация переменных); 2) сокращение числа переменных необходимых для описания данных [Википедия].

Для выявления наиболее значимых факторов и как следствие факторной структуры, наиболее часто применяется метод главных компонент (МГК). Суть метода заключается в замене коррелированных компонент некоррелированными факторами. Еще одной важной характеристикой метода является возможность ограничиться наиболее информативными главными компонентами и исключить из рассмотрения остальные, что упрощает интерпретацию результатов. Достоинство МГК также в том, что это – единственный математически обоснованный метод факторного анализа [Ким, Мьюллер, 1989].

Задача анализа главных компонент для конечных множеств данных имеет три эквивалентные формулировки, которые не выдвигают гипотез о статистическом порождении данных: 1) аппроксимировать данные линейными многообразиями меньшей размерности; 2) найти подпространства меньшей размерности, в ортогональной проекции на которые разброс данных (т. е. среднеквадратичное отклонение от среднего значения) максимален; 3) найти подпространства меньшей размерности, в ортогональной проекции на которые среднеквадратичное расстояние между точками максимально. Вычисление главных компонент сводится к вычислению собственных векторов и собственных значений ковариационной матрицы исходных данных.

Справка. *Ковариационная матрица* (матрица ковариаций) в теории вероятностей – матрица, составленная из попарных ковариаций элементов одного или двух случайных векторов. Ковариационная матрица случайного вектора – квадратная симметричная матрица, на диагонали которой располагаются дисперсии компонент вектора, а внедиагональные элементы являются ковариациями между компонентами.

Пусть $X = (X_1, X_2, \ldots, X_n)$ и $Y = (Y_1, Y_2, \ldots, Y_m)$ − два случайных вектора размерности n и m.Тогда матрица ковариации Σ векторов X, Y определяется равенством

$$\Sigma = \mathrm{cov}(\mathbf{X}, \mathbf{Y}) = \mathrm{E}\left[(\mathbf{X} - \mathrm{E}\mathbf{X})(\mathbf{Y} - \mathrm{E}\mathbf{Y})^{\mathrm{T}} \right]$$

где EX и EY − векторы средних значений. Компоненты ковариационной матрицы σ_{ij} определяются равенством

$$\sigma_{ij} = \mathrm{cov}\left(X_i, Y_j \right) \equiv \mathrm{E}\left[\left(X_i - \mathrm{E}X_i \right)\left(Y_j - \mathrm{E}Y_j \right) \right],$$

здесь i=1,...,n, j=1,...,m.

Если $\mathbf{X} \equiv \mathbf{Y}$, то Σ называется матрицей ковариации вектора $\mathbf{X}$ и обозначается cov($\mathbf{X}$). Сокращенная формула для вычисления матрицы ковариации имеет вид $\mathrm{cov}(\mathbf{X}) = \mathrm{E}[\mathbf{X}\mathbf{X}^{\mathrm{T}}] - \mathrm{E}[\mathbf{X}] \cdot \mathrm{E}[\mathbf{X}^{\mathrm{T}}]$.

МГК выявляет k компонент, объясняющих дисперсию и корреляцию k случайных величин. При этом компоненты строятся в порядке убывания объясняемой ими доли суммарной дисперсии исходных величин, что позволяет ограничиться несколькими первыми компонентами.

Первым шагом в процессе визуализации множества данных является ортогональное проецирование на плоскость первых двух главных компонент (или трехмерное пространство первых трех главных компонент). Плоскость проектирования, по сути, является плоским двумерным “экраном”, расположенным таким образом, чтобы обеспечить “картинку” данных с наименьшими искажениями. Такая проекция оптимальна (среди всех ортогональных проекций на разные двумерные экраны) в трех отношениях: 1) минимальна сумма квадратов расстояний от точек данных до проекций на плоскость первых главных компонент, т. е. экран расположен максимально близко по отношению к облаку точек; 2) минимальна сумма искажений квадратов расстояний между всеми парами точек из облака данных после проецирования точек на плоскость; 3) минимальна сумма искажений квадратов

расстояний между всеми точками данных и их “центром тяжести”. Визуализация данных является одним из наиболее широко используемых приложений МГК и его нелинейных обобщений.

МГК был представлен в работе [Pearson, 1901] для решения задачи наилучшей аппроксимации конечного множества точек прямыми и плоскостями, которая формулируется следующим образом. Дано конечное множество векторов $x_1, x_2, \ldots, x_m \in \mathbb{R}^n$. Для каждого $k = 0, 1, \ldots, n-1$ среди всех k-мерных линейных многообразий в $\mathbb{R}^n$ найти такое $L_k \subset \mathbb{R}^n$, чтобы сумма квадратов отклонений x_i от L_k была минимальна:

$$\sum_{i=1}^{m} \operatorname{dist}^2\left(x_i, L_k\right) \to \min,$$

где $\operatorname{dist}(x_i, L_k)$ – евклидово расстояние от точки до линейного многообразия.

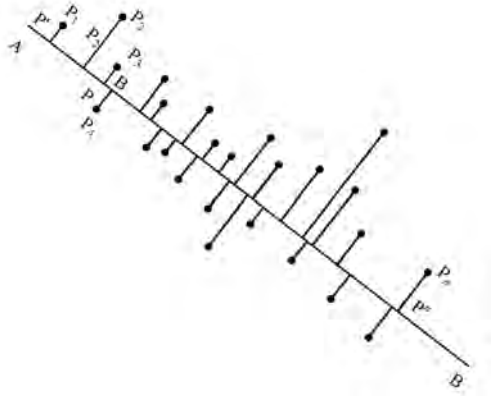

Рис. 6.2.1. Иллюстрация МГК [Pearson, 1901]

На рис. 6.2.1 P_i – точки на плоскости, p_i – расстояние от P_i до прямой AB. Требуется найти прямую AB, минимизирующую сумму $\sum_i p_i^2$ [Википедия].

МГК позволяет трансформировать множество (возможно) связанных переменных во множество меньшего размера несвязанных переменных, называемых главными компонентами. Первая главная компонента определяет как можно большее разнообразие данных. Каждая следующая компонента вновь определяет как можно большее разнообразие оставшихся данных. Таким образом, данные могут классифицироваться по нескольким факторам. Применение этих методик можно найти в ряде работ, посвященных анализу цитирования: [McCain, 1990], [McCain, 1995], [Raghupathi, Nerrur, 1999], [White, McCain, 1998].

6.2.4. *Многомерное шкалирование* (МШ, *Multidimensional scaling*). МШ − это набор статистических методов для представления данных о сходстве изучаемых объектов [Kruskal, 1977]. "…в более узком смысле МШ является упрощением большой и сложной совокупности наблюдений посредством построения пространственного представления, которое позволит увидеть отношения между объектами. Предполагается, что в такой пространственной модели величины близостей (сходств, различий, индексов совместной встречаемости или других мер тесноты или близости) соотносятся простым и прямым способом с расстояниями между попарно сравниваемыми объектами"

Данные для анализа заданы множеством I объектов, на которых определена функция расстояния. Алгоритм МШ начинается с рассмотрения матрицы расстояний $[\delta_{i,j}]$, где $\delta_{i,j}$ − расстояние между объектами i и j. Целью МШ является нахождение I векторов $x_1,\ldots,x_I \in \mathbb{R}^N$, где N определено заранее, таких что

$$\|x_i - x_j\| \approx \delta_{i,j} \text{ для всех } i, j \in I$$

($\|\cdot\|$ − норма вектора, в МШ это евклидово расстояние). Иными словами, МШ пытается отобразить I объектов в пространстве $\mathbb{R}^N$

таким образом, чтобы расстояния между объектами соответствовали исходным различиям.

Существуют различные методы определения векторов x_i. Как правило, МШ формулируется как задача определения такого набора (x_1,...,x_I), при котором достигается минимум функции расхождения (нелинейная модель)

$$\min_{x1,\ldots,xI} \sum_{i<j} \left(\| x_i - x_j \| - \delta_{i,j} \right)^2 .$$

Достоинством МШ является то, что он может использоваться для любого типа дистанции или матрицы подобия. Технология МШ используется для построения карт на основе документов. Примеры применения МШ можно найти в работах [Дэйвисон, 1988], [Chalmers, 1992], для анализа коцитирования – в работах [White, McCain, 1998], для построения карт науки – в работе [Small, 1999].

6.2.5. *Масштабирование на основе PFnet* (*Pathfinder Network Scaling*)**.** Это метод пространственного представления структуры данных на основании подобия элементов. В сетях класса Pathfinder (*PFnet*) узлы соответствуют единицам данных, а связи определяются на основании подобия [Schvaneveldt, et al., 1989]. Алгоритмы построения сети *PFnet* в качестве входных данных используют оценки близости между парами объектов и определяют сетевое представление объектов с сохранением только “важных” связей. Объекты являются узлами сети, пары объектов связываются направленными или ненаправленными ребрами, в зависимости от того, является отношение между объектами симметричным или несимметричным.

Топология конкретной сети *PFnet*, соответствующей матрице подобия, зависит от двух параметров: q и r. Параметр q ограничивает количество непрямых связей. Параметр r, определяющий

метрику (см. расстояние Минковского, табл. 5.4.1), используется при вычислении протяженности пути. Заметим, что путь из вершины a_i в вершину a_j – обход без повторения вершин.

Пусть n – количество единиц данных, тогда параметр q может меняться от 2 до $n-1$, параметр r – от 1 до ∞. Пусть имеется путь P, содержащий k ребер, имеющих веса w_1, w_2, $\ldots w_k$. Определим вес пути:

$$W(P) = \left[\sum_{i=1}^{k} w_i^r \right]^{\frac{1}{r}}.$$

Вес пути можно рассматривать как длину пути, если в сети *PFnet* соблюдаются следующие условия: 1) расстояние от документа до самого себя равно нулю; 2) матрица подобия симметрична; 3) выполняются условия неравенства треугольника для всех путей, содержащих не более k ребер. Например,

$$w_{ae} \le (w_{ab}{}^r + w_{bc}{}^r + \ldots + w_{de}{}^r)^{1/r}.$$

Расстояние между узлами (геодезическое) определяется как длина пути минимального веса. Параметр q определяет максимальную длину такого пути, удовлетворяющего условиям неравенства треугольника. Применение методики *PFnet* позволяет сократить количество пар объектов в представлении данных. В работе [Schvaneveldt, et al., 1989] приведены алгоритмы построения сетей и методика проверки соответствия полученной сети исходным данным.

Техника *PFnet* используется в универсальном анализе подобия (*Generalized Similarity Analysis*) для структуризации и визуализации распределенных информационных ресурсов [Chen, 1998a], [Chen, 1998b], для анализа коцитирования авторов [Chen, 1999], для визуализации областей знания [Chen, Paul, 2001].

6.2.6. *Самоорганизующиеся карты* (*Self-Organizing Maps, SOM*). Самоорганизующаяся карта − это искусственная соревновательная нейронная сеть с “обучением без учителя”, предназначенная для кластеризации и визуализации данных, представленных векторами в *n*-мерном пространстве [Kohonen, 1995]. Карта состоит из компонентов, называемых узлами или нейронами. Их количество, задаваемое аналитиком, обычно меньше, чем количество входных данных. Как правило, узлы располагают в вершинах регулярной решетки с квадратными или шестиугольными ячейками. Каждый узел описывается двумя векторами. Первый называется вектором веса m и имеет такую же размерность, что и векторы входных данных. Второй вектор r соответствует координатам узла на карте.

Перед началом обучения карты необходимо проинициализировать весовые коэффициенты нейронов. Удачно выбранный способ инициализации может существенно ускорить обучение и привести к получению более качественных результатов. Существует три варианта инициализации векторов веса: 1) инициализация случайными значениями, когда всем весам присваиваются малые случайные значения; 2) инициализация примерами, когда в качестве начальных значений задаются значения случайно выбранных примеров из обучающей выборки; 3) линейная инициализация. В этом случае веса инициируются значениями векторов, линейно упорядоченных вдоль линейного подпространства, расположенного между двумя главными собственными векторами исходного набора данных.

В процессе обучения векторы веса узлов приближаются к векторам входных данных. Для каждого шага итерации выбирается новый образец данных и наиболее похожий по вектору веса узел решетки (обычно похожесть понимается как расстояние в евкли-

довом пространстве), так называемый нейрон-победитель. После того как найден “нейрон-победитель”, проводится корректировка весов нейросети. Изменяется вектор веса, описывающий “нейрон-победитель”, и векторы, описывающие его соседей в сетке. Мера соседства определяется с помощью специальной функции, зависящей от координат векторов в решетке r и шага итерации. Исходный вектор, исследуемый на шаге итерации, и мера соседства с “нейроном-победителем” используются при определении функции корректировки векторов веса “нейрона-победителя” и его ближайших соседей. Таким образом, если во множестве входных данных два наблюдения были схожи, на карте им будут соответствовать близкие узлы. Циклический процесс обучения, перебирающий входные данные, заканчивается по достижении картой допустимой (заранее заданной аналитиком) погрешности или по совершении заданного количества итераций. После окончания процесса обучения все векторы исходных данных могут быть отображены на ближайшие нейроны, т. е. производится классификация.

Примеры использования *SOM* для классификации данных на основании встречаемости термов можно найти, например, в работах [Lin, 1997], [Lin, et al., 1991]. В работе [Polanco, et al., 2001] расширение *SOM* используется для картографирования науки.

6.3. Проецирование данных

Распространенным видом карт являются географические карты, выполненные в меркаторской проекции. Толковый словарь В. Даля дает следующее определение термина “меркаторская карта”: “…[карта], на которой градусы широты и долготы увеличиваются постепенно к полюсу, для взаимной соразмерности на

плоской бумаге”. Соблюдение “взаимной соразмерности” является едва ли не главной темой в научной картографии.

Справка. Меркаторская проекция − это способ изображения сферической поверхности Земли на плоскости (карты). Меркаторская проекция является равноугольной картографической проекцией, т. е. сохраняет правильность углов и направлений, но не сохраняет правильности размеров. Недостаток ее в том, что с увеличением широты растет и масштаб. Но есть и неоценимое достоинство − простота построения и нанесения на нее точек и линий. Кроме того, прямая линия, проведенная в произвольном направлении, является линией постоянного курса (пересекает меридианы под одинаковым углом). Именно поэтому меркаторскую проекцию применяют для навигационных карт. Изобретена Г. Меркатором − фламандским картографом и математиком XVI в. [Wikipedia]

Как правило, процесс проецирования данных, заключается в отображении множества данных, представленных в виде таблицы подобия, в пространство $\mathbb{R}^2$ таким образом, чтобы подобные данные располагались ближе друг к другу, нежели “не подобные”. В результате каждый объект получает две координаты.

6.3.1. *Триангуляция* (*Triangulation*). Триангуляция − это техника размещения объектов, при которой точки *n*-мерного пространства отображаются на двумерное пространство [Lee, et al., 1977]. Процесс начинается с выбранной случайным образом точки в первоначальной системе координат. Она размещается на плоскости. Затем выбирается наиболее близкая точка, которая помещается на определенном расстоянии от первоначальной точки. Размещение третьей точки проводится в зависимости от расстояния от первых двух на основе углов треугольника. Триангуляция использовалась, например, Г. Смоллом для построения многомерного ландшафта науки на пяти уровнях агрегации кластеров [Small, 1999].

Справка. В математике триангуляция [Wikipedia] – это метод разбиения топологического пространство на фигуры, для плоскости это треугольники. Одна базовая сторона и прилегающие к ней углы измеряются, а длины других сторон вычисляются с помощью тригонометрии. Рассмотрим пример для плоскости: Пусть l – длина стороны AB треугольника ACB (рис. 6.3.1), α и β –примыкающие к ней углы. Необходимо определить расстояние d от вершины C до стороны AB при условии, что CD перпендикулярна AB.

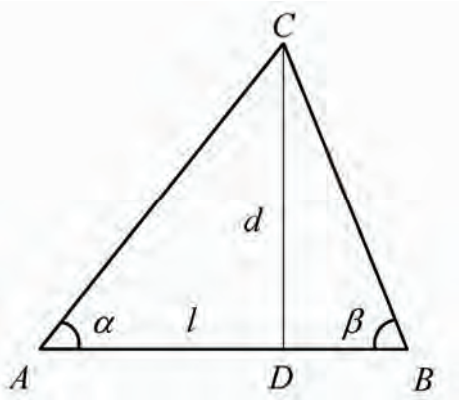

Рис. 6.3.1. Треугольник ACB

Задача решается следующим образом. Определим

$$l = \frac{d}{\tan\alpha} + \frac{d}{\tan\beta}.$$

Тогда

$$d = l \Big/ \left(\frac{1}{\tan\alpha} + \frac{1}{\tan\beta} \right).$$

Поскольку

$$\tan\alpha = \sin\alpha / \cos\alpha,$$

$$\sin(\alpha + \beta) = \sin\alpha \cos\beta + \cos\alpha \sin\beta,$$

получаем

$$d = \frac{l \times \sin\alpha \times \sin\beta}{\sin(\alpha + \beta)}.$$

6.3.2. *Силовые алгоритмы* (*Force directed placement – FDP*). Метод *FDP* используется для сортировки случайно расположенных объектов в требуемом порядке соответственно данным отношениям подобия и удобству (равномерное распределение, унифицированная длина ребер, симметрия, отсутствие перекрывающихся областей, минимальное пересечение ребер). В основу методики положен закон Гука (*R. Hooke*) – уравнение теории упругости, связывающее напряжение и деформацию упругой среды. Поскольку закон Гука записывается для малых напряжений и деформаций, он имеет вид простой пропорциональности. В словесной форме закон звучит следующим образом: сила упругости, возникающая в теле при его деформации, прямо пропорциональна величине этой деформации: $F = kl$. Здесь F – сила, с которой растягивают (сжимают) стержень, l – абсолютное удлинение (сжатие) стержня, k – коэффициент упругости (жесткости), зависящий от материала [Википедия].

При построении графа рассматривается виртуальная динамическая система, в узлах которой расположены физические тела, связанные между собой пружинами (или взвешенными дугами), устанавливающими натяжения. Основой любого силового алгоритма является модель динамической системы и алгоритм, вычисляющий состояние ее равновесия. Узлы перемещаются с учетом действующих на них сил, до тех пор пока не будет достигнут минимум локальной энергии. Состояние равновесия для этой системы достигается, когда длины дуг имеют тенденцию стать одинаковыми, а не связанные дугами узлы – располагаться достаточно далеко.

Опишем кратко модель, приведенную в работе [Kamada, Kawai, 1989]. Данный алгоритм реализован в системе визуализации *Pajek*. Предположим, нужно разместить n вершин: $v_1,\ldots,v_n$. Пусть $p_1, \ldots, p_n$ – точки на плоскости, соответствующие вершинам. Следует расположить точки таким образом, чтобы получилась сбалансированная система. Необходимо минимизировать степень дисбаланса, возникающего вследствие натяжения пружин, которая определяется по формуле

$$\frac{1}{2}\sum_{i=1}^{n-1}\sum_{j=i+1}^{n} k_{ij}\left(\left|\, p_i - p_j \,\right| - l_{ij}\right)^2,$$

где $k_{ij} = K / d(i, j)^2$; K – константа; l_{ij} – желаемое расстояние между вершинами i и j, пропорциональное кратчайшему пути между вершинами $d(i, j)$:

$$l_{ij} = L \times d(i, j),$$

L – "желаемая" длина дуги.

Глобальный минимум труднодостижим. Алгоритм достижения локального минимума начинается с произвольного размещения точек и подсчета всех k_{ij} и $l_{ij.}$ На каждом шаге по определенным правилам выбирается лишь одна точка, вычисляется, на какое расстояние ее можно переместить. Метод работает достаточно медленно, поскольку на каждом шаге итерации пересчитываются силы напряжения между всеми узлами. Другие реализации метода представлены в работах [Eades, 1984], [Di Batista, et al., 1994], [Fruchterman, Reingold, 1991]. Модификация метода, приведенная в [Davidson, et al., 2001], значительно ускоряет процесс. Еще одна модификация приведена в работе [Feng, Börner, 2002] и используется для кластеризации в группы семантически связанных документов [Boyack, et al., 2002]. Методы преобразования данных, рассмотренные в данном параграфе, представлены в табл. 6.3.1.

Таблица 6.3.1

Методы преобразования данных

Название метода	**Справочный материал**	**Применение**
Декомпозиция на основе собственных векторов и собственных значений	[Гантмахер, 1988]	[Davidson, et al., 1998]
Сингулярное разложение	[Гантмахер, 1988] [Логинов, 1996]	[Deerwester, et al., 1990]
Факторный анализ и метод главных компонент	[Айвазян и др., 1989] [Лоули, Максвелл, 1967] [Хартман, 1972]	[McCain, 1995] [Raghupathi, Nerrur, 1999] [White, McCain, 1998]
Многомерное шкалирование	[Kruskal, 1977] [Терехина, 1986] [Айвазян и др., 1989]	[Дэйвисон, 1988] [Chalmers, 1992] [White, McCain, 1998] [Small, 1999].
Масштабирование на основе *PFnet*	[Schvaneveldt, 1990]	[Chen, 1998a] [Chen, 1999] [Chen, Paul, 2001]
Самоорганизующиеся карты	[Kohonen, 1995] [Lin, 1997]	[Chen, Rada, 1996] [Chen, et al., 1996] [Polanco, et al., 2001].
Триангуляция	[Lee, et al., 1977]	[Small, 1999].
Силовые алгоритмы	[Di Battista, 1999] [Kamada, Kawai, 1989]	[Feng, Börner, 2002] [Boyack, et al., 2002].

6.4. Карты *ISI*

“Всеохватывающие карты науки *ISI* представляют собой новую двухмерную библиометрическую модель науки в целом. Исследовательские фронты, как минимальная структурная единица тематического строения науки, агрегируются и систематизируются в научные специальности; те, в свою очередь − в более крупные исследовательские направления, области знания, которые в совокупности и представляют иерархическую когнитивную

структуру современной науки. Первые карты науки, сформированные в *ISI* к 1990, выявили циклический характер науки, неоднородную ее структуру, маргинальный статус отдельных областей, в которых, как правило, происходит национальная специализация научных направлений" [ЭЭиФН, 2009].

В работах [Small, Sweeney, Greenlee, 1985], [Small, 1997] приведены истории создания карт науки на основе *SCI*, задачи, которые они призваны выполнять, а также развитие применяемых методов.

6.4.1. *Исторический экскурс*. Один из первых подходов к анализу науки путем картографирования на основе библиографических данных предложен Д. Прайсом [Price, 1966]. Объектом анализа были журналы, на основе индексов цитирования которых строились пространственные "картины" науки. Модель претерпевала изменения начиная с так называемой "военной карты" (*War map*), далее следовали "головоломка" (*Jigsaw puzzle*), луковичная кожура (*Onion skin*) и, наконец, "рваная рыболовная сеть"(*Torn fish net*). Особое развитие данный подход получил в работе [Narin, 1976], в которой рассматривается модель создания так называемых карт влияния (*Influence map*). В качестве мер влияния использовались нормализованные счетчики цитирования журналов. Однако с помощью таких подходов строились, скорее, карты влияния журналов.

Одним из направлений создания карт является использование фамилии автора как единицы анализа методом коцитирования [White, Griffith, 1982]. Большинство таких исследований посвящено изучению общественных наук на материале *SSCI* (*Social Science Citation Index*). Подход имеет специфические методологические проблемы, связанные с омографами и сокращением списка авторов до первого автора, вследствие этого использова-

ние данного метода для картографирования всей науки остается проблематичным.

Построение карт науки на основе ББД *SCI* с использованием метода коцитирования было начато в 1974 г. с картографирования биомедицинских областей науки. На основе кластеризации публикаций, индексированных базой данных *SCI* (см. 5.6), строились графы, которые показывали взаимосвязи кластеров. Выбирались наибольшие кластеры в определенной научной области (в данном случае – в биомедицине), определялся порог коцитирования кластеров и строился граф оставшихся связей. Мера коцитирования кластеров (*Cluster co-citation*) определялась как сумма всех коцитирований документов, относящихся к кластерам. Граф должен был быть как можно более простым, поэтому размещались только узловые области, чтобы не допускать перекрещивания связей. Это явилось первой попыткой построить карту исследовательских областей в рамках биомедицины на основе квартального файла *SCI*. В работе [Small, Sweeney, 1985] показано, как изменялся подход к кластеризации при изменении задач картографирования.

После того как анализ коцитирования был расширен до полных годовых файлов *SCI*, удалось получить обстоятельную карту областей исследования в биомедицине. Принцип построения остался прежним. Когда были построены карты за несколько последовательных лет, стали видны структурные сдвиги, касающиеся реструктуризации областей и возникновения новых центральных областей.

Появление подхода, основанного на коцитировании, частично обусловлено тем, что требовалось найти ответы на следующие вопросы:

1) Каковы реальные структурные единицы науки? Подобны ли они общепринятым дисциплинам, таким как физика и химия.

2) Как эти структурные единицы соотносятся друг с другом?

3) Каковы силы, определяющие эти структурные единицы и их взаимодействие? В какой мере карта науки является и картой научных знаний? Какую роль играют социальные факторы?

4) Как с течением времени меняется структура науки на макро- и микроуровнях?

Тем временем для анализа отдельных областей стала использоваться техника многомерного шкалирования, основанная на понятии различия между объектами, обратном понятию сходства, т. е. на вычислении расстояния между объектами. Применение многомерного шкалирования к наибольшим кластерам низшего уровня привело к возможности построения мультидисциплинарных карт науки [Garfield, et al., 1978]. Как правило, эти карты позволяли построить одномерную структуру – линию, на концах которой располагались физика и медицина, посередине – химия. Многомерное шкалирование позволило объективно оценить структурные изменения карт с течением времени.

6.4.2. *Иерархия кластеров*. Применение многоуровневой кластеризации обеспечило возможность анализа карт путем рассмотрения естественной иерархической структуры кластеров. В первом издании атласа науки *ISI* использовалась концепция вложенных карт, когда сначала предоставлялась биомедицинская карта, а затем − несколько вложенных карт, соответствующих точкам первоначальной карты.

Для получения иерархической структуры кластеров потребовалась новая техника − кластеризация кластеров. Многоуровневый подход можно представить следующим образом:

Уровень 1: на входе − цитируемые документы, на выходе − кластеры C_1;

Уровень 2: на входе − кластеры уровня C_1, на выходе – кластеры кластеров C_2;

Уровень 3: на входе − кластеры уровня C_2, на выходе – кластеры кластеров C_3.

Процесс можно продолжать. В любом случае выход предыдущего шага является входом следующего, его выход дает больший уровень агрегации.

На первом уровне коцитирование определяется как коэффициент коцитирования двух документов; начиная со второго уровня − как количество коцитирований документов различных кластеров. На каждом уровне осуществляется нормирование коэффициентов с помощью косинусной формулы Солтона, что обеспечивает их согласование. На каждом уровне кластеризация выполняется с варьированием коэффициентов, для того чтобы кластеры не превосходили заданного размера, это уменьшает вероятность получения аморфных макрокластеров. Определяется окно цитирования и публикаций, а также пороги полученных цитирований и сделанных в рамках окна ссылок. На каждой итерации получения новых кластеров из кластеров более низкого уровня применяется комбинированная мера сходства.

“Самый высокий уровень кластеризации − пятый (C_5) − называется макроуровнем. Карта науки этого уровня представляет собой глобальную карту областей исследования в естественных и социальных науках. Начиная с макроуровня, нужно двигаться как бы вниз: выбрав какой-то узел (кластер), можно на следующем уровне карты науки видеть карту этого кластера, также состоящую из совокупности кластеров − подобластей или дисциплин, и т. д. до второго уровня картографирования C_2, где выделенные кластеры представляют собой активные исследовательские фронты, над которыми работает научное сообщество. Нижний, или первый уровень C_1 включает кластеры “ядерных” публикаций, образующих исследовательские фронты и являющихся, по сути, классическими работами в этих исследовательских направлениях.

Карты науки *ISI* представляют собой новую двухмерную библиометрическую модель науки в целом. Исследовательские фронты, как минимальная структурная единица тематического строения науки, агрегируются и систематизируются в научные специальности; те, в свою очередь − в более крупные исследовательские направления, области знания, которые в совокупности и представляют иерархическую когнитивную структуру современной науки" [ЭЭиФФ, 2009].

6.4.3. *Комбинированная мера.* Методы кластеризации данных, основанные исключительно на коцитировании, могут быть подвергнуты справедливой критике относительно степени охвата. Для предупреждения этой ситуации была предложена новая комбинированная мера определения сходства, учитывающая коцитирование, прямое цитирование, опосредованное цитирование и библиографическое сочетание [Small, 1997]. Это нормализованная мера, изменяющаяся от нуля до единицы. На рис. 6.4.1 представлена комбинированная мера (*Combined linkage, Cl*), вычисляемая по формуле:

$$Cl(A, B)=(2dc+cc+lc+bc) / ((A.links + 1) \times (B.links+1))^{1/2}.$$

Для значений, приведенных на рис. 6.4.1, получаем

$$(2+3)/((4+1) \times (4+1))^{1/2} = 1.$$

На рис. 6.4.1 заштрихованные круги A и B − документы, связанные прямым цитированием (*dc, Direct citation*) и тремя видами непрямого цитирования: коцитированием (*cc, Co-citation*), опосредованным цитированием (*lc, Longitudinal coupling*) и библиографическим сочетанием (*bc, Bibliographic coupling*). Нормализованная мера получается путем взвешивания: прямое цитирование увеличивается в два раза, остальные просто суммируются. *A.links* и *B.links* в знаменателе – общее количество связей документов A

и B соответственно. Поскольку других связей между документами A и B в данном случае не существует, документы получают наибольшее значение коэффициента, равное единице.

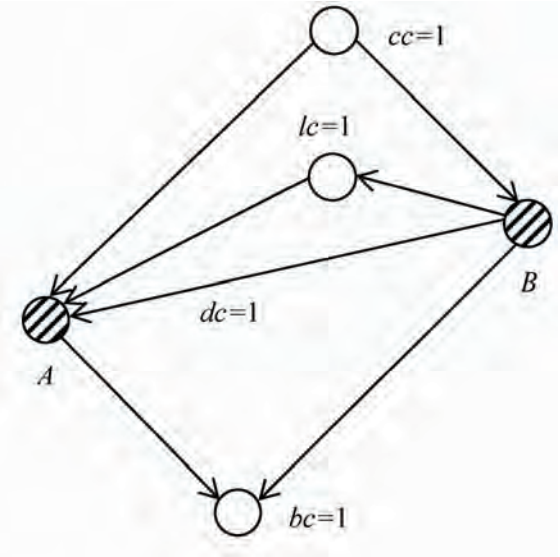

Рис. 6.4.1. Комбинированная мера

6.4.4. *Позиционирование.* Центральной для картографирования и визуализации информации является процедура позиционирования объектов в пространстве. Существует две стратегии решения проблемы крупномасштабного картографирования: 1) масштабирование с помощью одного из классических методов размещения, таких как многомерное шкалирование или анализ методом главных компонент; 2) разбиение базы данных на группы документов с помощью иерархической кластеризации, а затем преобразование групп в единую систему. Достоинством первого метода является использование одного процесса для порождения карты, недостатком − увеличение компьютерного времени с ростом количества объектов. Недостатком второго подхода является двуста-

дийность, однако время его исполнения существенно меньше зависит от роста количества объектов.

Для размещения данных в двумерном или трехмерном пространстве комбинированная мера подобия (*Similarity*) преобразуется в дистанцию с учетом порога (*Threshold*) сходства, используемого для формирования кластера. Были протестированы три способа преобразования: линейный (*Linear distance*), логарифмический (*Logarithmic distance*) и инверсный (*Inverse distance*). Формулы для получения относительного расстояния имеют вид:

linear distance = (1 − *similarity*) / (1 – *threshold*);

logarithmic distance = log(*similarity*) / log(*threshold*);

inverse distance = ((1/ *similarity*) − 1) / ((1/ *threshold*) − 1).

Во всех случаях, документы с наибольшим сходством находятся на нулевой дистанции. Сравнение показало, что для метода триангуляции, используемого для размещения объектов в координатном пространстве, наиболее эффективно линейное расстояние. Процесс триангуляции начинается с позиционирования начальной точки, затем позиционируется объект, наиболее сильно связанный с начальной точкой. После этого выбирается объект, наиболее связанный с этими двумя точками по сумме. Если существует возможность построения треугольника на этих величинах, то третий объект размещается в соответствующей вершине. Если не получается, то новый объект размещается как можно дальше от центра, и т. д. Данная методика также претерпевала изменения. Мерой полноты отображения связей является сравнение количества отображенных и возможных связей [Small, 1997].

6.4.5. *SciViz.* Рассмотрим схему создания крупномасштабных карт *SciViz* [Small, 1998], основанную на приведенных выше методиках. Тип визуализации, используемый в системе *SciViz*, − комбинация иерархической и географической моделей [Lin, 1997],

в которой иерархически организованные карты размещаются в координатной области. Данному типу можно дать определение с помощью термина “вложенное картирование”.

Процедура вложенного картирования выполняется в три этапа: 1) организация многоуровневой иерархии кластеров или секций, берущая начало с определенных документов или других объектов; 2) размещение объектов внутри каждого кластера в иерархическом виде с использованием двумерного или трехмерного представления каждой группы; 3) интеграция структур кластеров в глобальную структуру − общее координатное пространство. Последний этап включает растяжение, преобразование и вращение конфигурации на каждом уровне.

Связывание объектов. Для связывания объектов Г. Смолл использует метод, отличающийся от *VSM*. Сначала определяются связи между объектами. Если объектами являются документы, то используется прямая связь объектов, такая как связь по цитированию, или непрямая, такая как использование одинаковых лингвистических атрибутов (например, совместной встречаемости слов [Callon, Law, Rip, 1986]) или коцитирование [Small, 1995]. Нормализация этих мер связывания означает устранение влияния различного размера объектов или частоты их встречаемости в базе данных. Предполагается, что нормализация проведена таким образом, чтобы окончательные коэффициенты сходства варьировались от нуля до единицы для всех пар объектов: нуль означает, что связь отсутствует, единица – полное подобие.

Иерархия кластеров. Кластеризация включает предварительный анализ данных и используется для отделения сильно связанных объектов от несвязанных объектов. Обобщающая категория таких методов – это алгоритмы неперекрывающейся иерархической объединяющей кластеризации (см. 5.5). Большинство подобных алгоритмов использует эквивалент матрицы связей или

массива связей и пересчитывает сходство по мере организации кластеров. Предполагается, что определяются уровни иерархии и рассчитываются меры близости кластеров на каждом уровне.

Для получения визуальной картины объектов, организованных в пространстве, предпочтение отдается неперекрывающимся кластерам, так как в этом случае каждый объект принадлежит одному кластеру, т. е. ему соответствует единственная точка пространства. Г. Смолл считает, что иерархическая кластеризация предпочтительнее, так как позволяет включить в крупномасштабную структуру структуры меньшего масштаба. На каждом уровне иерархии можно организовать размещение объектов, распространяющееся на следующий уровень иерархии. Размещение объекта более высокого уровня может использоваться как центр для размещения относящихся к нему объектов более низкого уровня. Это открывает путь для финальной геометрии карты, отражающей иерархическую структуру. В работе [Johnson, Shneiderman, 1991] описываются варианты такого стиля, называемого *Tree-map*.

Одной из проблем является существование объектов, оставшихся не связанными на некотором уровне. Их расположение относительно всей иерархии произвольно. Вариантом решения является объединение таких объектов в отдельный мир, не связанный с основным.

Преобразование подобия в дистанцию. Для создания визуального образа требуется определение пространственных размеров объектов и связей. В системе *SciViz* коэффициент сходства, изменяющийся от нуля до единицы, превращается в дистанцию с использованием линейных (простейшее), обратных и логарифмических преобразований мер подобия. Таким образом, полное сходство превращается в нулевую дистанцию.

Желательно также нормализовать дистанцию, так чтобы изображение кластеров приблизительно соответствовало количеству

содержащихся в нем объектов. Одним из подходов является использование того же порога, который используется для меры сходства при формировании кластеров. Таким образом, порог определяет самую слабую связь внутри кластера, соответствующая дистанция может быть принята за единицу. Эту единицу близости назвали "*Garfield*" в знак уважения к изобретателю *SCI*. Она позволяет изображать кластеры в одном масштабе.

Размещение. Размещение – это процесс позиционирования объектов в пространстве. Для визуализации результатов кластеризации подходит двумерное или трехмерное решение. Система *SciViz* использует современный встроенный метод, базирующийся на методе геометрической триангуляции, впервые использованном в системе *SCI-map*[*] [Small, 1994]. Поскольку сначала проводится кластеризация, можно размещать каждый кластер отдельно. После размещения координаты каждого объекта преобразуются, так что среднее, или центр, становится началом координат. Все кластеры на каждом уровне иерархии размещаются раздельно. Результат размещения определяет местоположение объектов внутри кластера. Так как кластер более высокого уровня является объединением кластеров более низкого уровня, то центр кластера более высокого уровня используется в качестве центра для набора кластеров низшего уровня.

Создание общего координатного пространства. После определения иерархии объектов и размещения для каждого кластера остается объединить их в единую структуру. Это − процесс размещения по-разному организованных групп в общее координатное пространство, отражающее иерархическое отношение как геометрическую позицию и сохраняющее отношение уровней.

[*] *SCI-map* – система, разработанная для построения карт на основе выбранного пользователем набора документов, в то время как *SciViz* предоставляет исследователю предварительно картографированную базу данных, которая может использоваться для навигации.

Создание общего координатного пространства является ключевым моментом системы *SciViz*. Стратегия заключается в размещении сначала самого низкого, первого уровня агрегации. В нем располагаются документы и кластеры документов. Каждое координатное пространство расширяется по мере продвижения на следующие уровни иерархии к корню. После достижения корня начинается движение в обратном направлении по иерархии с перемещением центров каждого кластера на новые места. Позиция объектов более высокого уровня становится центром объектов более низкого.

Для определения соответствующего фактора расширения можно использовать несколько подходов. Слишком большое расширение может привести к формированию большого пустого пространства, слишком маленькое − к перекрытию или скоплению. В *SciViz* фактор расширения для каждого объекта находится путем определения пределов каждого более низкого объекта иерархии. Он представляет собой достаточно большой круг (или сферу для 3D) для включения всех объектов. Расстояние между объектами до расширения известно, следовательно, можно вычислить, насколько необходимо расширить координаты, чтобы наибольшие смежные круги или сферы в худшем случае соприкасались. Чтобы предотвратить излишнее расширение, проверяются на перекрытие только соседние объекты в минимальном связующем дереве.

Исключение возможности перекрытия. В визуальной презентации необходимо, чтобы объекты имели ограниченные размеры, позволяющие им быть хорошо различимыми, а не просто математическими точками. Простейшим методом является определение последовательности объектов внутри кластеров (например, связующее дерево) и тестирование объектов последовательности на предмет перекрытия со всеми предыдущими объектами последо-

вательности, после чего перекрывающиеся объекты перемещаются в направлении от центра кластера. Такая процедура гарантирует преобразование в неперекрывающуюся конфигурацию.

Ориентация кластеров. Поскольку внутри кластера объекты размещаются относительно центра, то кластер можно вращать.

Отображение и навигация. Исследователь сначала видит всю карту. Каждый круг помечен генеральной темой или ключевыми словами, а размер круга соответствует количеству и распределению тем внутри него. На самом высоком, наиболее агрегированном уровне карты находится коллекция неперекрывающихся кругов, располагающихся на разных расстояниях друг от друга. При приближении кругов обнаруживается подобная сеть кругов меньшего размера, имеющих более специфическую пометку темы. Последовательно можно дойти до уровня, на котором располагаются сами документы, имеющие вид кругов, помеченных автором или названием. На рис. 6.4.2 представлен фрагмент вымышленной карты, приведенный в работе [Cahlik, 2000], цифры соответствуют меткам кластеров.

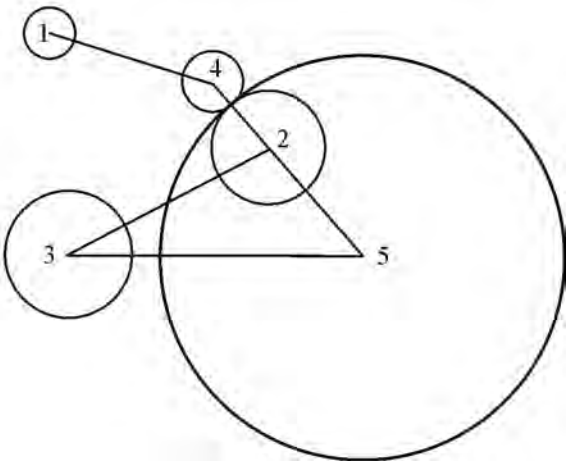

Рис. 6.4.2. Вымышленная карта

6.5. Сравнение карт

Монография [Börner, 2010] посвящена вопросам возникновения и эволюции карт науки. В ней приводятся и иллюстрируются успешные технологии в области научной картографии. Они отражают работу исследовательских групп всего мира и представляют различные подходы к проблеме техники создания карт и базы данных, что свидетельствует об актуальности этого вида исследований.

В работе [Klavans, Boyack, 2009] изучается вопрос сравнения результатов построения карт в терминах “конвергенции” и “консенсуса”. Под конвергенцией (*Convergence*) авторы понимают возможность одинаковой формы, содержания и связей карт, отображающих одну и ту же часть науки. Консенсус (*Consensus*) – более слабое условие существования возможности построения на основе информации ряда карт согласованной карты, разделяющей достаточное количество свойств с исходными картами и являющейся представительной. Авторы [Klavans, Boyack, 2009] утверждают, что конвергенция отсутствует, а консенсус возможен. В работе представлена методика организации согласованной карты на основе 20 карт науки. Рассматриваемые карты построены различными способами: на основе экспертной оценки, на основе коцитирования документов, журналов и проч. Выявлено три основных типа карт науки. Первый тип – иерархические карты (*H-map*), хорошо отражающие линейные последовательности элементов с возможными ветвлениями. Второй тип – карты центрической формы (*C-map*), на которых одна дисциплина находится в центре радиальной сети, имеющей высокую степень ветвления. Третий тип карт, как правило, имеет кольцевую структуру и отличается от иерархического лишь тем, что структура смыкается, образуя кольцо (*R-map*).

В табл. 6.5.1 приведена информация, выбранная из 20 карт, построенных на основе анализа информации из ББД относительно статей и журналов.

Таблица 6.5.1

Карты на основе ББД *Thomson* и *Scopus*

Источник	**Исходные данные**	**ББД и год**	**Тип**
[Griffith, et al., 1974]	1 150 статей	*SC*, 1972	*C-map*
[Small, Garfield, 1985]	11 000 статей	*SC+SS*, 1983	*H-map*, *C-map*
[Bassecoulard, Zitt, 1999]	~2 000 журналов	*SC/JCR*, 1993	*H-map*, *C-map*
[Small, 1999a]	36 720 статей	*SC+SS*, 1995	*H-map*
[Boyack, et al., 2005]	7 121 журнал	*SC+SS*, 2000	*R-map*
[Klavans, Boyack, 2007]	1,9 М статей	*SC+SS*, 2004	*R-map*
[Klavans, Boyack, 2007]	2,1 М статей	*Scopus*, 2004	*R-map*
[Klavans, Boyack, 2008]	800 К статей	*SC+SS*, 2003	*R-map*
[Rosvall, Bergstrom, 2008]	6 116 журналов	*SC+SS*, 2004	*R-map*
[Boyack, et al., 2009]	7 227 журналов	*SC+SS*, 2002	*R-map*
[Boyack, 2009]	8 667 журналов	*SC+SS+PR*, 2003	*R-map*

Примечание. Использованы следующие сокращения относительно ББД T*homson*: *SC – Scientific's Science; SS – Social Science; PR – Proceedings Citation*.

Согласованная карта строится путем декомпозиции двадцати карт. Она основана на данных 16 отраслей науки, представленных в табл. 6.5.2.

Цель настоящей работы − примерное определение расположения указанных отраслей на рассматриваемых картах. Связь между отраслями считалась согласованной, если присутствовала более чем в половине рассматриваемых карт (табл. 6.5.3.).

Таблица 6.5.2

Отрасли науки

1. M-Математика (*Mathematics*)	9. B-Биология (*Biology*)
2. CS-Компьютерные науки (*Computer Science*)	10. I-Инфекционные заболевания (*Infectious disease*)
3. P-Физика (*Physics*)	11. MD-Медицина (*Medical specialties*)
4. PC-Физическая химия (*Physical chemistry*)	12. HS-Медицинское обслуживание (*Health services*)
5. C-Химия (*Chemistry*)	13. N-Нейронаука (*Brain research*)
6. E-Технические науки (*Engineering*)	14. PS-Психология / Психиатрия (*Psychology / Psychiatry*)
7. G-Науки о Земле (*Earth sciences*)	15. SS-Социальные науки (*Social sciences*)
8. BC-Биохимия (*Biochemisrty*)	16. H-Гуманитарные науки (*Humanities*)

Таблица 6.5.3

Распределение пар отраслей наук

Ранг	**Пара**	***N***	***N*-poss**	**Доля, %**
1, 2	*B-BC, I-MD*	20	20	100,0
3	*H-SS*	8	8	100,0
4	*C-PC*	19	20	95,0
5, 6	*HS-MD, PS-SS*	16	17	94,1
7	*P-PC*	18	20	90,0
8, 9	*MD-N, E-G*	16	18	88,9
10	*B-G*	17	20	85,0
11	*BC-I*	16	20	80,0
12, 13	*E-PC, N-PS*	14	18	77,8
14	*CS-M*	13	18	72,2
15, 16	*BC-MD, BC-C*	14	20	70,0
17	*E-P*	12	18	66,7
18	*B-I*	13	20	65,0
19	*CS-SS*	10	16	62,5
20	*H-PS*	5	8	62,5
21	*M-P*	11	19	57,9

22	*C-E*	10	18	55,6
23	*C-P*	11	20	55,0
24	*HS-N*	8	15	55,3
25	*CS-E*	9	17	52,9
26	*C-G*	10	20	50,0
27	*HS-PS*	8	16	50,0

Примечание. Параметр *N-poss* указывает число карт, на которых присутствует данная пара наук, *N* – число карт, на которых эта пара наук не только присутствует, но и находится в состоянии взаимной связи. Для примера рассмотрим пару *E-PC* (технические науки и физическая химия), одновременно присутствующую на 18 картах из 20. Из этих 18 карт они взаимосвязаны на 14 картах, следовательно, они встречаются совместно в 77,8 % возможных случаев.

Процесс формирования табл. 6.5.3. выполнялся по шагам. На первом шаге зафиксированы четыре фундаментальные отрасли: математика, физика, химия и биология, и шесть возможных комбинаций. В результате оказалось, что только две комбинации – физическая химия и биохимия – присутствуют на всех картах.

Далее исследовались такие прикладные по отношению к математике, физике и химии отрасли, как информатика, техника и наука о Земле. Другие три – инфекционные болезни, медицина и нейронаука – являются прикладными по отношению к биологии.

Еще три прикладные области – здравоохранение, психология и общественные науки – относятся, скорее, к социальным, чем к точным наукам. Это очень обширные и разнообразные области, которые недостаточно точно представлены на исследуемых экспертных и библиографических картах, основанных на ББД *WoS*. Подключение информации ББД *Scopus* позволяет лучше представить роль этих областей в науке, особенно это касается здравоохранения. Наконец, гуманитарные науки могут рассматриваться как фундаментальные по отношению к общественным. Однако эта область присутствует на недостаточном числе карт.

В работе [Klavans, Boyack, 2009] приведены два представления согласованных карт – одно- и двумерное. На рис. 6.5.1 приведено одномерное изображение согласованной карты типа *R-map*.

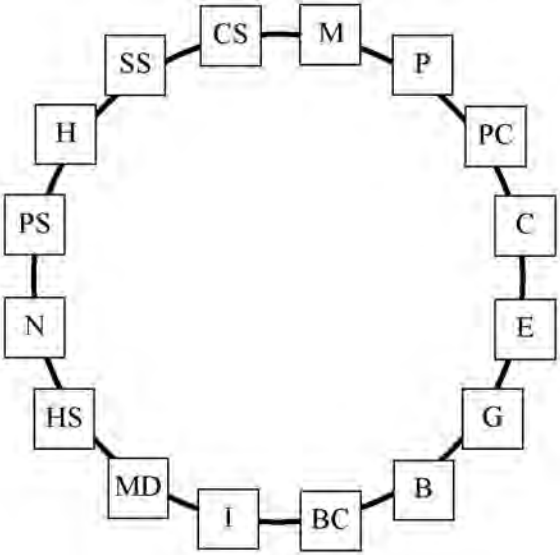

Рис. 6.5.1. Согласованная карта типа *R-map*

В работе [Cahlik, 2000] рассмотрены методы сравнения карт науки, построенных на основе анализа цитирования документов и совместной встречаемости слов. Обсуждаются две задачи.

Первая − сравнение карт одной и той же научной области, базирующихся на публикациях, относящихся к различным периодам времени. Сравнение позволяет обнаруживать динамику эволюции научной области. Если рассматривать карты на основе совместной встречаемости слов, то основной методологической проблемой является идентификация исследовательских тематик на различных картах, поскольку происходит изменение не только положения тематики на карте, но и набора ключевых слов, характеризующих тематику. Решение проблемы выявления идентично-

сти лежит в определении (как правило, субъективном) порога встречаемости общих слов.

Вторая задача – сравнение карт за один и тот же период, построенных с использованием различных методов анализа документов, − имеет следующие решения.

Карты, построенные путем анализа цитирования и анализа совместной встречаемости слов, используют различные типы представления. Предлагается способ преодоления проблемы путем приведения карты, полученной на анализе цитирования, к виду, используемому при анализе совместной встречаемости слов, а именно к так называемым стратегическим диаграммам (*Strategic diagram*).

Идентификация исследовательских тематик, представленных на картах, выполняется путем нахождения релевантных публикаций. При анализе на основе встречаемости слов они выявляются путем определения встречаемости ключевых слов, соответствующих тематике, и сравнения с некоторым порогом. При анализе цитирования релевантные документы определяются путем вычисления совместной цитируемости с публикациями, определяющими тематику. Темы на разных картах считаются соответствующими, если имеют заданное порогом количество общих релевантных публикаций. В работе [Cahlik, 2000] приведены примеры согласования карт, отражающих исследования в области *Water Resources* за 1994−1999 гг.

6.6. Средства визуализации

Рассмотрим программные продукты, предназначенные для анализа и визуализации библиографической информации.

6.6.1. *HistCite*. В работе [Garfield, et al., 2010] приведен пример анализа и визуализации библиографических данных с использованием пакета программ *HistCite*. Основная задача пакета –

преобразование библиографических данных в хронологическую диаграмму, с помощью которой можно получить ответы на следующие вопросы: 1) Как много работ опубликовано в некоторой научной области? 2) Когда и какие страны внесли наибольший вклад в данную область? На каких языках издавались работы? 3) Какие журналы наиболее “популярны” в данной области? 4) Какие из них наиболее важны? 5) Какие наиболее авторитетные авторы работают в рассматриваемой области? 6) Какие публикации наиболее важны? 7) Как различные участники влияют друг на друга?

Графическая форма ответа облегчает восприятие больших объемов информации. Так называемый историограф (*Historiograph*) – это диаграмма связей публикаций с учетом временных периодов. Историографы базируются на информации о цитировании. Каждая публикация представляется в виде символа, выбираемого пользователем. Символы располагаются вдоль временных меток, соответствующих дате публикации. В зависимости от запроса получается историография для выбранного периода.

Исходными данными для анализа являются загружаемая библиография, подобранная пользователем в качестве представительной литературы. Далее необходимо выполнить следующие действия: 1) выбрать данные из ББД; 2) очистить их от дублирования; 3) привести к единообразному виду (формату); 4) добавить недостающие публикации; 5) выполнить сортировку, например по количеству цитирований. На данном этапе предлагается использовать индикатор цитирования *PRI* (*Percentile rank index*) [Pudovkin, Garfield, 2009], вычисляемый по формуле

$$PRI = (N - R + 1) / N \times 100,$$

где N – количество публикаций в годовом выпуске журналов; R – ранг публикаций, упорядоченных по убыванию полученных

цитирований, наиболее цитируемая публикация получает наивысший ранг (равный единице) и имеет *PRI*=100.

Результатом выполнения предварительных процедур является *Masterfile* − основа для построения диаграмм двух основных видов:

− *LCS* (*Local citation score*) отображает важные для работы цитирующие ее публикации из исходной библиографии;

− *GCS* (*Global citation score*) отображает важные для работы цитирующие ее публикации из всей базы.

На рис. 6.6.1 приведена *LCS*-диаграмма, отражающая влияние статьи [Van Raan, 1990] - цитируемые и наиболее влиятельные цитирующие ее работы. Из ББД *WoS* по запросу (van raan a* or vanraan a*) было извлечено и проанализировано 104 публикации, из них 14 признаны наиболее влиятельными. Статья [Van Raan, 1990] помечена кругом с меткой 6. Величина кругов соответствует *PRI*-коэффициенту.

6.6.2. *Pajek*. Это пакет программ для анализа больших сетей, включающих тысячи узлов и десятки тысяч связей [Batagelj, Mrvar, 1998] (см. сайт: vlado.fmf.uni-lj.si/pub/networks/default.htm). Мотивацией для создания пакета стало наличие исходных данных для построения существующих сетей, представленных в машинно-читаемом виде. Визуализацию можно назвать *Graph-based*, поскольку она базируется на библиотеках графических структур данных и алгоритмов. К основным задачам, решаемым с помощью этого пакета, следует отнести достижение уровня абстракции путем разложения большой сети на малые сети, к которым легче применимы алгоритмы; обеспечение пользователя мощными средствами визуализации; возможность выбора эффективных алгоритмов анализа. *Pajek* позволяет производить кластеризацию, рассматривать вершины кластера изолированно или объединять

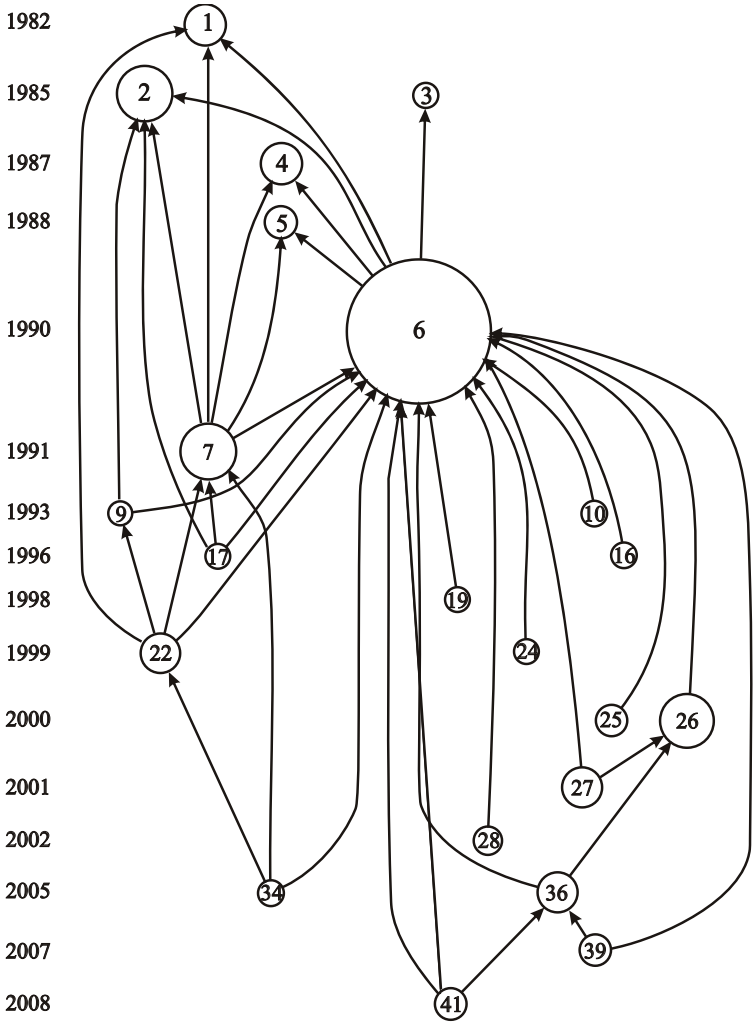

Рис. 6.6.1. *LCS*-историограф работы [Van Raan, 1990]

их в кластер, а также отображать связи между кластерами. В качестве входного и выходного файла, представляющего сеть, выступает легко читаемый и корректируемый текстовый файл, который затем преобразуется в формат *Pajek.net*.

Помимо функций визуализации и графического преобразования в пакете реализовано множество алгоритмов анализа сетей, в том числе принадлежащие базовому множеству [Ахо и др., 1979], [Knuth, 1993], [Rogers, Kincaid, 1981], [Tarjan, 1983].

Графические возможности пакета позволяют выполнять различные виды визуализации сети, следуя известным алгоритмам преобразования. Для автоматической генерации размещения элементов графа используются алгоритмы *Kamada − Kawai* [Kamada, Kawai, 1989] (двумерное изображение), *Fruchterman − Reingold* [Fruchterman, Reingold, 1991] (двумерное и трехмерное), относящиеся к технологии силовых алгоритмов (*Force-directed graph drawing*) и *Lanczos algorithm* [Wikipedia], основанный на технике собственных векторов и собственных значений.

Рассмотрим возможности *Pajek* в области визуализации и анализа библиометрических данных. В пакете имеются средства преобразования “сырых” данных, представленных в виде таблицы. Например, строки соответствуют авторам, а столбцы − признакам. В частности, существует несколько способов автоматического сопоставления веса с дугами, которые отображаются в виде толщины линий на графе. Имеется возможность отображать плотные сети в виде матриц с применением алгоритма кластеризации.

Данные, извлекаемые из ББД, таких как *WoS*, *Scopus* и др., можно автоматически преобразовать (с помощью специализированных программ и дополнительных средств проверки на несоответствия и устранение цикличности) в сети по выбранным темам: работы × авторы (сеть *WA*), работы × ключевые слова (сеть *WK*), работы × классификация (сеть *WC*), работы × работы (сеть *Ci*). Существует возможность сортировки статей по журналам, издательствам и году публикации. Поскольку основой всех четырех сетей являются работы (*W*), из них можно получить другие сети с

помощью операций транспонирования и умножения, например сеть сотрудничества

$$Co = (WA)^{T} \times WA$$

или сеть цитирования авторов

$$Ca = (WA)^{T} \times Ci \times WA.$$

Для сети Ca элемент $w(i, j)$ – это количество цитирований из работ, автором которых является i, на работы автора j. При анализе вклада отдельного автора учитывается соавторство. Предлагаются способы нормализации сетей относительно весов, что позволяет сравнивать различные сети. Анализ сетей может проводиться на нескольких уровнях на основании подразделения множества единиц на группы, например авторов в соответствии с организациями или странами.

Анализ сети цитирования заключается в выявлении важных структурных единиц. Он базируется на определении веса (важности) вершин / дуг. Методы вычисления весов предложены в работах [Hummon, Doreian, 1989], [Hummon, Doreian, 1990]. Типичные шаги и результаты анализа приведены в работе [Batagelj, 2003].

Пример использования пакета *Pajek* в исследованиях по библиометрии можно найти в работе [Leydesdorff, 2007], в которой приведены результаты исследования по определению позиции профессора Т. Брауна (*Tibor Braun*), редактора журнала *Scientometrics*. Проанализировано 183 статьи, в которых Т. Браун является, по меньшей мере, соавтором. Применялся метод БиС. На основе общедоступных программных продуктов, использующих данные *ISI/WoK*, строятся две матрицы транзакций, которые могут использоваться программой визуализации *Pajek*: матрица совместной встречаемости и косино-нормализованная матрица. Матрица совместной встречаемости – это обобщенное название

матриц коцитирования, библиографического сочетания, совместной востребованности слов и т. д.

Результатом анализа является, во-первых, граф, описывающий связи профессора Т. Брауна с другими авторами публикаций на основании библиографических сочетаний со 183 работами. Поскольку граф является обозримым, можно сделать вывод, что связи Т. Брауна не ограничиваются журналом *Scientometrics*. Во-вторых, было проведено исследование сотрудничества автора с учеными из других областей науки. Для этого рассматривалась косино-нормализованная матрица, отражающая связи между журналами (236), в которых опубликованы работы, библиографически сочетающиеся с работами профессора Брауна. Из 236 только 18 относятся к ядру области *Scientometrics*. В работе также представлен график распределения журналов, составленный на основе цитирований из работ Т. Брауна. Если рассмотреть граф, представляющий связи между журналами, и выделить журналы, для которых мера "*Betweenness centrality*" [Wikipedia] более 1 % (21 журнал), то распределение соответствует степенному закону, однако "хвост" (207 журналов) убывает по экспоненциальному закону. В другом контексте при рассмотрении всех цитирований "журнал × журнал" [Leydesdorf, Bensman, 2006], наоборот, "хвост" убывает по степенному закону.

6.6.3. *VOSviewer*. Это компьютерная программа (см. сайт http://www.vosviewer.com/), предназначенная для конструирования и визуализации карт на основе библиометрических данных. В ее состав включен анализатор, принцип работы которого описан в [Van Eck, Waltman, 2007a].

В отличие от *Pajek*, базирующегося на теории графов, *VOSviewer* строит карты на основе расстояния между двумя элементами, которое отражает силу связи между ними. Чем меньше расстояние, тем больше подобие. Во многих случаях элементы на

такой карте распределены неровно, так что метки кластеров могут налагаться друг на друга. Однако на таких картах хорошо видна сила связи. В *VOSviewer* применяется техника МШ [Van Eck, et al., 2010].

Реализовано четыре способа визуализации карт:

1) Рассеянный вид (*Scatter view*) – элементы помечаются кружками без меток, возможно, с использованием цвета, что позволяет выявлять общую структуру;

2) Помеченный вид (*Label view*) – элементы по умолчанию помечены кружками-метками и, возможно, цветом; чем важнее элемент, тем больше круг; возможны перекрытия, тогда не все метки будут отражены;

3) Вид с учетом плотности (*Density view*) – элементы отображаются как и в случае *Label view*, цвет элемента зависит от плотности элементов, расположенных рядом, что способствует выявлению ядра исследований;

4) Отображение кластеров с учетом плотности (*Cluster density view*) – используется в случае, когда проведена кластеризация; плотность элементов для каждого кластера отображается индивидуально; цвета внутри кластера определяются с учетом плотности составляющих его элементов; способ эффективен при рассмотрении взаимосвязей кластеров.

Построение карты состоит из трех основных шагов.

Шаг 1. На основе матрицы совместной встречаемости путем нормализации строится матрица подобия. Для нормализации используется коэффициент ассоциативности (*Association strength*), имеющий вид

$$s_{ij} = \frac{c_{ij}}{w_i w_j},$$

где s_{ij} – коэффициент подобия документов i и j; c_{ij} – частота совместной встречаемости документов i и j; w_i, w_j – коэффициент

встречаемости документов i и j или коэффициент совместной встречаемости (см. 5.3).

Шаг 2. На основе матрицы подобия строится карта с использованием *VOS*-технологии, т. е. конструируется двумерное представление, при этом элементы располагаются таким образом, чтобы максимально точно отразить коэффициент сходства s_{ij} для каждых двух. Смысл заключается в минимизации взвешенной суммы квадратов евклидова расстояния между всеми парами. Чем больше сходство, тем выше вес суммируемого расстояния. Иными словами, требуется минимизировать функцию

$$V(x_1,...,x_n) = \sum_{i<j} s_{ij} \left\| x_i - x_j \right\|^2 ,$$

где n – количество точек, вектор $\boldsymbol{x}_i=(x_{i1}, x_{i2})$ задает координаты размещения элемента i на плоскости.

Шаг 3. Карта преобразуется с целью получения удобного вида и отображается.

Поскольку решение, получаемое на шаге 2, не является единственным, то применяются несколько стадий трансформации карты, так чтобы для каждой индивидуальной матрицы подобия получилась одна и та же карта. Визуализация представления с учетом плотности требует определения функций плотности для элементов и кластеров. Подробную информацию можно найти в [Van Eck, Waltman, 2010], программные компоненты представлены в работах [Van Eck, Waltman, 2007], [Van Eck, et al., 2006] и [Van Eck, et al., 2008].

Глава 7. Индекс Хирша

Ничто не может быть сильнее идей,
время которых пришло.
В. Гюго

7.1. Общие сведения

Индекс Хирша, или *h*-индекс (*h-index*), предложен Х. Хиршем в 2005 г. [Hirsch, 2005a], [Hirsch, 2005b] в качестве меры продуктивности ученого, основанной на распределении цитирования его работ. Определение *h-индекса* ученого: *ученый имеет h-индекс, равный h, если h из его* N_p *публикаций имеют, по крайней мере, h цитирований каждая, а остальные* ($N_p - h$) *работ имеют не больше, чем h цитирований каждая*. Множество всех публикаций автора, удовлетворяющих этому определению, называют *h*-ядром (*Hirsch core*).

Введение данной метрики Х. Хирш аргументировал тем, что этот критерий оценки с помощью единичного числа предпочтительнее других подобных метрик, таких как количество работ, деленное на общее количество цитирований или количество цитирований, приходящихся на одну работу. Однако Х. Хирш считал, что применение одной этой количественной меры может дать только грубую аппроксимацию индивидуального профиля ученого, ее следует применять в случае, если дело касается грантов или подтверждения статуса ученого. Х. Хирш предложил при подсчете *h*-индекса в случае соавторства делить количество цитирований работы на количество соавторов и сопоставлять полученную долю с каждым автором.

В работе [Glänzel, 2006] отмечено, что преимущество *h*-индекса перед другими показателями подобного рода заключается в учете как количества публикаций, так и их востребованности (число цитирований этих публикаций). Таким образом, *h*-индекс

является результатом баланса между количеством публикаций и количеством цитирований, полученных каждой публикацией. Например, он учитывает, что ученый (индивидуально или в соавторстве) может опубликовать одну или несколько выдающихся работ, которые будут много цитироваться, однако не обеспечат равномерную производительность автора. В то же время *h*-индекс "поощряет" ученых, имеющих непрерывный поток публикаций и постоянное влияние или влияние выше среднего [Bornmann, Daniel, 2005]. В. Гланзель [Glänzel, 2006] считает, что сила этого индекса − в потенциальной возможности оценивать небольшие наборы публикаций, для которых часто неприменимы традиционные библиометрические индикаторы.

Правомерность использования данной метрики подтверждается следующими примерами. *h*-индекс нобелевских лауреатов оказался высоким (84 % имели *h*-индекс, равный, по крайней мере, 30), а члены Национальной академии наук по физике и астрономии США, выбранные в 2005 г., имели *h*-индекс, равный в среднем 46.

Сравнение *h*-индекса с другими библиометрическими метриками проводилось многими авторами. Так, в работе [Bornmann, Daniel, 2005a] подтверждается соответствие величины индекса оценке коллег. В работе [Cronin, Meho, 2006] выявлена корреляция *h*-индекса с оценкой, полученной на основе подсчета общего количества цитирований. В работе [Van Raan, 2006] подтверждается корреляция *h*-индекса с несколькими библиометрическими метриками, а также с оценкой коллег. В работе [Kelly, Jennions, 2006] обнаружена корреляция *h*-индекса с общим количеством публикаций. Эти исследования подтверждают правомерность использования *h*-индекса.

В работе [Lehnmann, et al., 2005] высказывается сомнение в корректности применения индекса для вычисления научной

производительности. Автор работы [Sidiropuolos, et al., 2007] отмечает, что *h*-индекс имеет ряд недостатков, в основном связанных с его неспособностью дифференцировать активных и неактивных ученых, выявлять важные работы, созданные в прошлом, а также работы, которые задают тенденцию и продолжают влиять на научное мышление. Х. Хирш к недостаткам *h*-индекса относил зависимость от продолжительности научной карьеры ученого и области исследований, имеющей свои традиции цитирования.

Тем не менее, предоставление *h*-индекса ББД в качестве индикатора (менее чем через два года после определения!) является показателем того, что он стал общепринятой мерой академических достижений. На основе *h*-индекса определено большое количество новых индексов, предназначенных для преодоления недостатков и использования совместно с *h*-индексом. Тем более что в работах ряда авторов высказывается предположение о невозможности использования одномерной метрики в многомерном пространстве библиометрии [Glänzel, 2006].

7.2. Метрика *hJ*

Классическое определение индекса Хирша предназначено для сравнения производительности авторов. В работе [Braun, et al., 2006] приводится мнение, что широкому распространению *h*-индекса в качестве меры производительности ученых мешают как недостатки (отсутствие стандартов цитирования в научных дисциплинах, адекватное взвешивание соавторства и др.), так и естественное и обоснованное нежелание научного сообщества использовать для оценки численные индексы. Однако существуют области библиометрии, в которых меры, основанные на цитировании, получили большее признание. Одной из таких областей является анализ цитирования журналов.

В работе [Braun, et al., 2006] определяется метрика *hJ*, которая является *h*-индексом для журналов и вычисляется следующим образом. Пусть журнал *J* имеет *N* публикаций за рассматриваемый период времени *T*. Публикации упорядочим в порядке убывания цитирований. Из этой последовательности выберем *h* статей, имеющих, по крайней мере, *h* цитирований, так чтобы каждая из оставшихся (*N*–*h*) статей имела число цитирований, меньшее или равное *h*. В этом случае говорят, что журнал *J* за период *T* имеет индекс Хирша журнала *hJ*. Авторы считают, что эта метрика выгодно дополняет факторы важности журналов как мера, устойчивая к выбросам.

В работе [Harzing, Wal, 2008] проводится сравнение двух метрик: индекса *hJ*, вычисляемого на основе *Google Scholar*, и импакт-фактора, традиционно вычисляемого на основе ББД *WoS*. Импакт-фактор *IF* определяется как среднее количество цитирований, полученных в рассматриваемом году на публикации за два предшествующих года. Отмечается, что, во-первых, *hJ* не связан с фиксированным временным горизонтом. Его можно вычислять, рассматривая любой временной отрезок. Во-вторых, *hJ* ослабляет влияние одной высокоцитируемой работы, т. е., как и в случае ранжирования авторов, является устойчивой метрикой, отражающей непрерывную и долговременную деятельность. В-третьих, журнал, печатающий большее количество работ, имеет большую вероятность получить более высокий *h*-индекс. Это соответствует тому, что журнал, публикующий большое количество высокоцитируемых статей, имеет большее влияние в научной области. Если индекс − мера важности, то данный фактор является преимуществом, так как позволяет оценивать не публикации, а журнал в целом.

Авторы [Harzing, Wal, 2008] исследовали указанные выше метрики на примере 838 журналов по экономике и бизнесу, для которых на сайте Harzing (*www.harzing.com/pop.htm*) ведется *Journal Quality List* − информация о ранжировании журналов согласно 20 метрикам. Список содержит журналы, относящиеся к 15 различным направлениям, и считается достаточно влиятельным. Было выделено множество журналов (536 названий), для которых вычислены обе метрики за 2003−2006 гг. Для сравнения выбран коэффициент ранговой корреляции Спирмена (*Spearman's rank correlation coefficient*), так как обе метрики имеют распределение, отличающееся от нормального. Среднее значение коэффициента оказалось равным 0,718 (колебание от 0,633 до 0,891 для различных подобластей). Для метрик, вычисленных на основе различных баз данных с учетом различных временных интервалов, корреляция выражена существенно. Кроме того, обнаружено, что примерно 50 журналов, составляющих половину высоко ранжированных относительно *hJ*, не индексированы в *WoS*, и большинство из них издается в Европе.

Несмотря на строгую корреляцию, отмечается значительное различие. Причины различия можно разделить на пять групп:

а) "высокий" индекс оперативности оказывает значительное влияние на значение *IF* (например, у журналов в области "Психология");

б) отдельные высокоцитируемые работы повышают *IF*;

в) цитирование материалов, не охваченных базой данных *WoS*, повышает *hJ*;

г) ограниченное количество публикуемых работ может понизить *hJ*, несмотря на высокую цитируемость работ в журнале;

д) менее существенные причины, такие как повышение *IF* вследствие учета в числителе цитирования от источников,

которые не учитываются в знаменателе, или возможность ошибки при поиске информации, обусловленная, например, наличием омографов.

Авторы [Harzing, Wal, 2008] делают вывод, что для журналов, для которых не существует покрытия базой данных *WoS*, можно пользоваться аналогичной метрикой *hJ*. В целом эта метрика преодолевает статистические ограничения, свойственные *IF*. Метрика *hJ*, как правило, применяется для оценки более широкого влияния, чем академическое, и рассматривается как мера социальной и экономической важности.

Несмотря на положительные качества, метрику *hJ* следует применять с осторожностью при учете индивидуальной производительности ученых. Исследования показали, что высокоцитируемые работы не обязательно печатаются в самых влиятельных журналах, и наоборот, существенная часть работ, опубликованных в самых влиятельных журналах, не является высокоцитируемой.

7.3. Математика *h*-индекса

Строгое определение *h*-индекса для автора приведено в работе [Glänzel, 2006a]. Оно формулируется следующим образом. Предположим, что некоторый автор опубликовал n трудов и каждая i-я публикация имеет X_i цитирований (i=1,2,…,n). Упорядочим публикации по “рангу”, т. е. по убыванию значений X_i

$$X_1^* \geq X_2^* \geq \ldots \geq X_n^*, \qquad (7.3.1)$$

где X_1^* – число цитирований, полученное наиболее цитируемой публикацией; X_n^* – число цитирований, полученное наименее цитируемой публикацией. Будем считать, что значение *h*-индекса для автора вычисляется по формуле

$$h = \max\{j : X_j^* \geq j\}.$$

7.3.1. *Аксиоматика.* В работе [Woeginger, 2008] приводятся результаты исследования необходимых и достаточных условий, уникально характеризующих *h*-индекс. Условия сформулированы в виде аксиом, учитывающих изменения в количестве и качестве: количестве публикаций и цитируемости.

Исследователь с $n \geq 0$ публикациями формально описывается с помощью вектора $x = (x_1, x_2, \ldots, x_n)$ с неотрицательными компонентами $x_1 \geq x_2 \geq \ldots \geq x_n$; k-я компонента вектора определяет количество цитирований, полученных k-й работой, работы упорядочены по количеству цитирований. Если исследователь не имеет работ, вектор пуст. Пусть X – множество таких векторов. Мы говорим, что вектор $y = (y_1, y_2, \ldots, y_m)$ доминирует над вектором $x = (x_1, x_2, \ldots, x_n)$, если $m \geq n$ и для $\forall k$, $0 \leq k \leq n$ верно $x_k \leq y_k$. Для обозначения этой ситуации будем писать $x \preceq y$.

Приведем определения, отражающие элементарные желаемые свойства индекса научной важности.

Определение 1. Индекс научной важности – это функция f, отображающая множество X на множество неотрицательных целых чисел $\mathbb{N}_0$ и удовлетворяющая следующим двум условиям:

а) если $x = (0, \ldots, 0)$ или x – пустой вектор, то $f(x) = 0$;

б) монотонности: если $x \preceq y$, то $f(x) \leq f(y)$.

Определение 2. Индекс Хирша h – это индекс научной важности $h: X \to \mathbb{N}_0$, который присваивает вектору $x = (x_1, \ldots, x_n)$ значение $h = \max\limits_{m \leq k} \{k : x_m \geq k\}$.

Теперь сформулируем аксиомы, касающиеся добавления единичной публикации. Увеличение индекса происходит только за счет публикации с количеством цитирований большим, чем индекс.

*Аксиома A*1. Если вектор y размерности $(n+1)$ получен из вектора x размерности n путем добавлением одной публикации с $f(x)$ цитированиями, то $f(y) \leq f(x)$.

*Аксиома A*2. Если вектор y размерности $(n+1)$ получен из вектора x размерности n добавлением одной публикации с $(f(x)+1)$ цитированиями, то $f(y) > f(x)$.

Рссмотрим аксиомы, отражающие добавление цитирований к старым работам. Минимальные изменения не должны приводить к большим изменениям индекса.

Аксиома B. Если вектор y размерности n получен из вектора x размерности n путем увеличения количества цитирований единичной публикации, то $f(y) \leq f(x) + 1$.

Аксиома C. Если вектор y размерности n получен из вектора x размерности n путем добавления не более одного цитирования к каждой публикации, то $f(y) \leq f(x) + 1$.

И финальная аксиома, отражающая изменения и в количестве публикаций, и в количестве цитирований.

Аксиома D. Если вектор y размерности $(n + 1)$ получен из вектора x размерности n добавлением одной публикации с $f(x)$ цитированиями, а затем увеличением количества цитирований каждой публикации по меньшей мере на единицу, то $f(y) > f(x)$.

Заметим, что приведенные выше аксиомы не являются независимыми, так, из аксиомы *A*2 следует аксиома *D*. Показано, что если индекс удовлетворяет аксиомам *A*1 и *A*2, то он не должен удовлетворять аксиоме *B* и что ни один индекс научной важности не может удовлетворять всем аксиомам. Доказана теорема: индекс научной важности f: $X \to \mathbb{N}_0$ удовлетворяет аксиомам *A*1, *B* и *D* тогда и только тогда, когда это h-индекс.

В ряде работ продолжено развитие аксиоматики. Так, в работах [Quesada, 2009], [Quesada, 2010] предложена аксиоматика для индекса научной важности в случае, когда индекс задает отображение множества X на множество неотрицательных вещественных чисел. В работе [Miroiu, 2013] доказано, что h-индекс может быть аксиоматизирован с учетом только одного типа изменений, а именно с учетом изменения количества публикаций.

7.3.2. *Модель Хирша.* В своей основополагающей работе [Hirsch, 2005a] Х. Хирш предложил математическую модель h-индекса, в которой связал h-индекс с общим количеством цитирований ($N_{c,tot}$). В ядре Хирша число цитирований не менее h^2, поэтому целесообразно определить коэффициент пропорциональности a, такой что

$$N_{c,tot} = \alpha h^2,$$

а значит

$$h = \sqrt{\frac{N_{c,tot}}{\alpha}}.$$

Этот коэффициент определяет долю неучтенных цитирований и зависит от индивидуального распределения. Эмпирически определено, что значение константы a находится в диапазоне от 3 до 5.

В рамках простейшей линейной модели предполагается, что количество новых цитирований публикации автора за год есть некоторая константа c, а количество публикаций за год – p. Тогда количество цитирований за $(n+1)$ год выражается формулой

$$N_{c,tot} = \sum_{j=1}^{n} pcj = \frac{pcn(n+1)}{2}.$$

Предположим, что все публикации, вплоть до года y, вносят вклад в вычисление h, тогда верны два равенства: $py = h$, $(n - y)\,\mathrm{c} = h$. В результате для достаточно большого n имеем

$$h = \frac{c}{1+\frac{c}{p}}n \quad \text{и} \quad N_{c,tot} \approx \frac{\left(1+\frac{c}{p}\right)^2}{\frac{2c}{p}}h^2 .$$

Отсюда следует, что коэффициент a зависит от количества публикаций и цитирований на публикацию, прирастающих за год. Для модели, в которой ученый публикует работы эквивалентного качества и в постоянном темпе, справедливо выражение $h \sim mn$.

Х. Хирш предположил, что отношение h к n может использоваться для сравнения ученых с различной продолжительностью карьеры. При этом первая публикация не всегда может быть подходящей точкой отсчета, так как может пройти время до того, как ученый начнет получать устойчивые результаты. Кроме того, коэффициент не подходит для оценки ученых, не поддерживающих уровень продуктивности в течение всей карьеры. В работе [Harzing, 2008] указано, что коэффициент m неустойчив на ранних стадиях карьеры, и для молодых ученых небольшие отличия в h-индексе могут привести к большим отличиям относительно m. Тем не менее в некоторых случаях m может использоваться в качестве дополнительной метрики.

Если в представленной линейной модели рассматривать публикации по убыванию количества цитирований (7.3.1), X_r^* является линейной функцией от r (X_0^* – число цитирований, полученное наиболее цитируемой публикацией). В этом случае распределение цитирований можно выразить формулой

$$X_r^* = X_0^* - \left(\frac{X_0^*}{h} - 1\right) r .$$

7.3.3. *Модель Schubert – Glänzel.* В работе [Glänzel, 2006a] предпринята попытка интерпретировать теоретически некоторые

свойства *h*-индекса. Исследовалась зависимость *h*-индекса от параметров распределения и размеров выборки с использованием теории вероятностей экстремальных значений Э. Гумбеля [Гумбель, 1965]. В работе [Schubert, Glänzel, 2007] исследуется соответствие теоретических и практических взаимосвязей нескольких метрик.

Пусть X – случайная переменная. В нашем случае это скорость цитирования публикации. Рассмотрим плотность распределения X: $p_k = P(X = k)$ и функцию распределения X: $F(k) = P(X < k)$.

Определим

$$G_k = G(k) := 1 - F(k) = P(X \geq k).$$

Предположим, имеется выборка из n элементов $\left(\{X_i\}, \right.$ $\left. i = 1, \ldots, n \right)$, где все компоненты независимы и имеют функцию распределения F. Для выборки размером n r-е характеристическое наибольшее значение Гумбеля (u_r) определяется следующим образом:

$$u_r := G^{-1}\left(r/n \right) = \max\left\{ k : G_k \geq r/n \right\}$$

Ранговая статистика $R(r) = X_r^*$ ($X_1^* \geq X_2^* \geq \ldots \geq X_n^*$ – элементы выборки, упорядоченные по убыванию) может рассматриваться как статистическая оценка r-го наибольшего значения u_r.

Согласно работе [Glanzel, 2006a] теоретический *h*-индекс можно определить следующим образом:

$$h := \max\left\{ r : u_r \geq r \right\} = \max\left\{ r : \max\left\{ k : G_k \geq r/n \right\} \geq r \right\},$$

$$\text{если } n > 0,\ X_1 \geq 1.$$

Если существует индекс r, такой что $u_r = r$, то $h := r$ и, следовательно, $h := u_h$.

Рассмотрим важный случай, а именно дискретные распределения Парето (*Pareto distribution*) с конечным математическим ожиданием. К этой категории принадлежит большинство распределений, используемых для моделирования публикационной активности и процесса цитирования.

Будем считать, что распределение случайной величины является распределением Парето, если оно асимптотически подчиняется закону Ципфа (*Zipf's law*), т. е. если k стремится к бесконечности, то $G_k k^{-\alpha}$ – константа. Асимптотически случайные переменные, имеющие распределение Парето, удовлетворяют этому условию, так как

$$p_k = P(X = k) \approx d(N + k)^{-(\alpha+1)},$$

если $k \gg 1$; $\alpha > 1$; N и d – положительные константы. Далее рассматриваются только такие распределения. Для $k \gg N$ выполняются соотношения

$$p_k = P(X = k) \approx dk^{-(\alpha+1)}, \quad G_k = P(X \geq k) \approx d_1 k^{-\alpha},$$

где d_1 – положительная константа. Значит, имеем ожидаемое значение

$$E(X) = \sum_{k=0}^{\infty} kp_k = \sum_{k=0}^{\infty} G_k < \infty, \text{ если } \alpha > 1.$$

Произведя элементарные манипуляции с функцией распределения, можем получить следующую аппроксимацию из определения r-х характеристических экстремумов Гумбеля

$$u_r \approx c_1 \left(\frac{n}{r} \right)^{\frac{1}{\alpha}}, \qquad (7.3.2)$$

где c_1 – положительная константа. Применяя эти приближения для случая h-индекса, получаем следующее свойство:

$$h = u_h \approx c_1 \left(\frac{n}{h} \right)^{\frac{1}{\alpha}}, \text{ если } n \gg 1. \qquad (7.3.3)$$

Следовательно,

$$h \approx c_2 n^{1/(\alpha+1)}, \text{ если } n \gg 1, \qquad (7.3.4)$$

где $c_2 = c_1^{\frac{\alpha}{\alpha+1}}$ – положительная константа. Иными словами, *h*-индекс приблизительно пропорционален корню (α+1)-й степени из числа публикаций.

Существует проблема определения *h*-индекса в случае журналов через параметры распределения цитирований: прежде всего через математическое ожидание − среднее значение числа цитирований на одну публикацию (обозначим его *CPP*) и "размер выборки". В случае распределения Парето (второго типа) с двумя параметрами *N*, α ожидаемое значение выражается равенством

$$CPP = \frac{N}{\alpha - 1} \qquad (7.3.5)$$

(константа в выражении (7.3.2) $c_1 = N$). Тогда согласно (7.3.3) и (7.3.4)

$$h = u_h \approx c_1 \left(\frac{n}{h} \right)^{\frac{1}{\alpha}} \approx N^{\alpha/(\alpha+1)} n^{1/(\alpha+1)} \qquad (7.3.6)$$

В случае α = 2, что соответствует закону Лотки (*Lotka's law*) с экспонентой 3, согласующемуся с предположениями, принятыми в библиометрии, выражение (7.3.5) можно записать как $CPP = N$. Тогда, преобразуя (7.3.6), получаем

$$h = cn^{1/3} CPP^{2/3}, \qquad (7.3.7)$$

где *c* – положительное число порядка 1.

Связь метрик *h*, *n* и *CPP* изучалась на практике. Результаты приведены в работе [Schubert, Glänzel, 2007]. Вычислялись

метрики: количество публикаций *n*, средняя скорость цитирования *CPP*, и *h*-индекс 6406 журналов за 2001 г. и 6481 журнала за 2002 г. Данные взяты из базы данных *WoS*. Было выбрано трехгодовое окно цитирования (год публикации и два предшествующих), чтобы иметь возможность вычислять как *h*-индекс, так и импакт-фактор *IF* (в этом случае *CPP* = *IF*). Рассматривались четыре типа документов: статьи (*Articles*), письма (*Letters)*, заметки (*Notes*) и обзоры (*Reviews*).

Показано, что не только *h* имеет строгую линейную корреляцию с произведением $n^{1/3}IF^{2/3}$, но и что константа *c* независима от научной области и года рассмотрения и $c \approx 0{,}75$.

Таким образом, теоретическая связь между метриками *h*, *n* и *CPP*, выведенная из модели Парето для распределения цитирований публикаций журналов и представленная в выражении (7.3.7), полностью соответствует эмпирическим данным о цитировании за год и окне цитирования, равном трем годам.

Заметим, что единственный параметр *c* практически не зависит от научной области. Отсюда следует, что индекс *h* не имеет особой зависимости от научной области, кроме хорошо известной зависимости от характерных для области количества публикаций и скорости цитирования. В этом смысле выражение (7.3.7) обеспечивает некоторую “трансформацию подобия” *h*-индекса между различными областями.

Индекс Хирша и другие индексы, определенные на его основе (например, *g*, *A*, *R*), значительно расширяют горизонты библиометрии, их использование для оценки журналов является перспективным, поскольку связано с системным и статистическим анализом.

В работе [Glänzel, 2008] продолжено исследование статистических особенностей *h*-индекса. Рассматривается непрерывное распределение Парето второго типа как моделирующее распределе-

ние цитирований. В этом случае вместо F_x, G_x будем писать $F(x)$ и $G(x)$. Неотрицательная случайная переменная X имеет распределение Парето второго типа [Kleiber, Kotz, 2003], называемое также распределением Ломакса (*Lomax*), если выполняется условие

$$G(x) = P(X \geq x) = \frac{N^{\alpha}}{(N+x)^{\alpha}} \text{ для всех } x \geq 0.$$

При достаточно больших x ($x \gg N$) можно записать

$$G(x) \sim \frac{N^{\alpha}}{x^{\alpha}} \text{ для } x \gg N.$$

Рассматривая статистический образец размера n с распределением Ломакса, получаем:

$$G(u_r) \approx \frac{N^{\alpha}}{u_r^{\alpha}} = \frac{r}{n}, \text{ если } n \gg r.$$

Следовательно, $r u_r^{\alpha} = N^{\alpha} n$ и

$$\zeta(r) := r^{\frac{1}{\alpha+1}} u_r^{\frac{\alpha}{\alpha+1}} = N^{\frac{\alpha}{\alpha+1}} n^{\frac{1}{\alpha+1}}. \quad (7.3.8)$$

Так как правая часть (7.3.8) не зависит от ранга r, то левая часть должна быть константой. Более того, $\zeta(h) = h$, так как

$$\zeta(h) = h^{\frac{1}{\alpha+1}} h^{\frac{\alpha}{\alpha+1}} = h.$$

Следовательно, $\zeta(r)$ – это константная функция, более того,

$$\zeta(r) \equiv h \text{ для всех } r \ll n. \quad (7.3.9)$$

Соотношение (7.3.8) является центральной частью дальнейших исследований: левая часть – основа для анализа взаимосвязи с другими индикаторами, правая часть предоставляет средства для статистического анализа хвоста распределения цитирований в контексте h-индекса.

Из свойств, описанных в выражениях (7.3.8), (7.3.9), следует важный результат, принимая во внимание, что математическое ожидание распределения Ломакса вычисляется по формуле

$$E(X) = \frac{N}{\alpha - 1}.$$

Согласно работе [Schubert, Glänzel, 2007]

$$h = c\left(\alpha\right)^{*} E(X)^{\frac{\alpha}{\alpha+1}} n^{\frac{1}{\alpha+1}},$$

если $\alpha > 1,\ c\left(\alpha\right)^{*} = \left(\alpha - 1\right)^{\frac{\alpha}{\alpha+1}}$ – положительное действительное число, зависящее только от параметра α. Учитывая, что модель на основе непрерывного распределения Ломакса плохо соотносится с эмпирическим дискретным целочисленным распределением, хорошая корреляция между теоретическими и эмпирическими значениями невозможна. Однако в работе [Schubert, Glänzel, 2007] показано, что существует корреляция между h и значением выражения

$$\overline{x}^{\frac{\alpha}{\alpha+1}} n^{\frac{1}{\alpha+1}}, \qquad (7.3.10)$$

где $\overline{x}$ – средняя скорость цитирования научных журналов. Следует отметить, что эмпирическое значение $c(\alpha)^{*}$ несколько меньше теоретического значения. Этот результат оказался поразительно стабильным для малого окна цитирования и независимым от научной области. Однако параметр α не проявил инвариантности относительно временного интервала. Для небольших периодов, охватывающих три года после публикации, подходит значение $\alpha = 2$. Для более длинных периодов подходит меньшее значение $\alpha = 1{,}5$.

Теоретический и эмпирический анализы показали, что h-индекс и выражение (7.3.10) строго коррелируют и индекс h может

рассматриваться как составной индикатор, основанный на публикационной активности автора и средней скорости цитирования, однако он не рассматривается в качестве замены h-индекса. Подобный составной индикатор определен в работе [Lindsey, 1978], где предложен индекс, который учитывает всю продукцию и может использоваться для сравнения в различных научных областях, а именно индекс *Corrected Quality Ratio* (*CQ*), определяемый следующим образом

$$CQ = \frac{C}{P} \cdot \left(CP \right)^{\frac{1}{2}} = \left(\frac{C^3}{P} \right)^{\frac{1}{2}}.$$

В нашем обозначении это $n\overline{x}^{\frac{3}{2}}$ (7.3.10). Применяя преобразование $CQ \to CQ^{0,4}$, получаем $CQ^{0,4} = n^{0,4} \times \overline{x}^{0,4}$, что совпадает с выражением, соответствующим составному индикатору при $\alpha = 1{,}5$. Поскольку степенная функция является монотонной, трансформация не влияет на ранжирование с помощью CQ.

Рассмотрим левую часть равенства (7.3.8). Заменим u_r на соответствующий ранжированный элемент и определим

$$z(r) := r^A R(r)^{1-A}, \quad A = \frac{1}{1+\alpha}$$

($z(r)$ может рассматриваться как возможная статистическая оценка выражения в правой части равенства (7.3.8) при $r \ll n$).

В работе [Glänzel, 2008] показано, что медиана M эмпирического распределения $z(r)$ является хорошей статистической оценкой для h, несмотря на то что отклонение от $\zeta(r) \equiv h$ является значительным для конкретных значений r. Z-статистики могут использоваться для анализа хвоста распределений цитирования в контексте h-индекса. В то же время h-индекс оказался полезен как точка сечения для анализа ранговых зависимостей, например

применения z и связанных с ней статистик к публикациям в ядре Хирша [Burrell, 2007].

7.3.4. *Модель Egghe – Rousseau.* В работе [Egghe, Rousseau, 2006a] h-индекс определяется в рамках обобщенной концепции *Information Production Processes* (*IPP*s, процессы производства информации) в терминах "источник – объект" (*Source − item*). Примером пары "источник − объект" могут служить пары "журнал − статья" или "статья − цитирование".

Рассмотрим *IPP*, состоящий из источников и объектов. Пусть $R(r)$ – функция ранжирования этой системы: если источники упорядочены в убывающем порядке количества объектов, то в дискретном случае $R(r)$ – количество объектов, "производимых" источником ранга r (будем считать эту функцию непрерывной). Переформулируем определение h-индекса для данной концепции, используя T вместо N_p из первоначального определения. В непрерывном случае функция R, определенная на отрезке [0, T] и задающая плотность объектов, полагается строго положительной, строго убывающей и непрерывной. Тогда h-индекс – это r, такое что

$$R(r) = r.$$

В работе [Egghe, Rousseau, 2006a] показано, что каждый *IPP* имеет единственный h-индекс.

Теперь в системе, в которой выполняется степенной закон Лотки, рассмотрим функцию зависимости между размером и частотой (*Size-frequency function*)

$$f: [1,\infty[\to]0,C]$$

вида:

$$f(j) = \frac{C}{j^{\alpha}},$$

где $C > 0$, $\alpha > 1$ – экспонента Лотки. В дискретном варианте функция определяет количество источников продуктивности j. В непрерывном варианте функция интерпретируется как плотность. Показано, что если рассматривать закон Лотки, то для заданного количества источников T верно

$$h = T^{\frac{1}{\alpha}}$$

Показано также, что в системе с A объектов, подчиняющейся закону Лотки, h-индекс равен:

$$h = \left(\frac{\alpha - 2}{\alpha - 1} A \right)^{\frac{1}{\alpha}}.$$

7.3.5. *Сравнение моделей.* В работе [Ye, 2009] проведен сравнительный анализ трех представленных выше моделей на материале БД *WoS* и *Essential Science Indicators* (*ESI*). Для однообразия переформулируем представление h-индекса в соответствии с математическими моделями. Пусть C − общее количество цитирований, P – количество публикаций.

В модели Хирша $h = \left(C/a \right)^{\frac{1}{2}}$, где $3 < a < 5$ – константа. В модели Egghe − Rousseau $h = P^{1/\alpha}$, где α – экспонента Лотки. В модели Schubert − Glänzel $h = cP^{\frac{1}{3}}(CPP)^{\frac{2}{3}}$, где $CPP = C/P$ – среднее количество цитирований, которое для журналов ассоциируется с импакт-фактором *IF*, c – константа.

Проведены исследования следующих оценок h-индекса, соответствующих моделям, при $\alpha = 2$, $a = 5$:

$$h_c \sim \sqrt{\frac{C}{5}}, \; h_p \sim \sqrt{P}, \; h_{pc} \sim cP^{\frac{1}{3}} \left(C/P \right)^{\frac{2}{3}}.$$

Данные об эмпирическом *h-индексе* брались из ББД *WoS* в рамках временного окна, равного десяти годам, а данные о *P*, *C*, *CPP* – из ББД *ESI*. При вычислении h_{pc} константа c устанавливалась равной $c = 0{,}9$ для журналов и $c = 1$ для организаций.

Результаты анализа показали, что оценка Schubert − Glänzel ближе к реальным значениям h-индекса как для журналов, так и для организаций. Поскольку при вычислении h-индекса учитываются и цитирования, и публикации, модель Schubert − Glänzel наиболее близка к реальности. В большинстве случаев сравнительная характеристика для журналов и организаций имеет вид

$$h_p < h \sim h_{pc} < h_c.$$

В работе [Ye, 2009] приведены коэффициенты корреляции Пирсона, вычисленные на множестве наиболее цитируемых журналов и организаций.

7.4. *h*-последовательности

Существенными недостатками h-индекса являются недооценка высокоцитируемых публикаций в h-ядре и работ, получивших не намного меньше чем h цитирований; невозможность учета длительности карьеры ученого; потребность более точной оценки процесса цитирования, которую трудно получить с помощью единственной метрики. Существует ряд исследований, направленных на исправление перечисленных недостатков и повышение эффективности h-индекса.

В работе [Schreiber, et al., 2012] приведен неполный список h-подобных индексов (*h-type indices*), часть из них представлена в табл. 7.4.1. Пусть c_r – количество цитирований публикации ранга r. Обозначим через $C(r) = \sum_{j=1}^{r} c_j$ общее количество цитирований на публикации ранга $1 \div r$, $C=C(n)$ – общее количество цитирований. Заметим, что $C(h) \geq h^2$.

Таблица 7.4.1

Некоторые *h*-подобные индексы

Индекс	Определение, формула
h [Hirsch, 2005a]	$h=\max\{j: c_j \geq j\}$
A [Jin, 2006]	$A=1/h\ C(h)$ (среднее количество цитирований для публикаций в *h*-ядре)
g [Egghe, 2006]	$g=\max\{j: C(j) \geq j^2\}$
$\boldsymbol{h^{(2)}}$ [Kosmulski, 2006]	$h^{(2)}=\max\{j: c_j \geq j^2\}$
ħ [Miller, 2006]	$\hbar = \sqrt{\frac{C}{2}}$
R [Jin, et al., 2007]	$R=\sqrt{C(h)}$
x [Kosmulski, 2007]	$x = \max\{i \times c_i,\ i \leq n\}$ (максимум среди произведений ранга на количество цитирований публикации данного ранга)
m [Bornmann, et al., 2008]	m – медиана цитирований, полученных публикациями в *h*-ядре
$\boldsymbol{h_w}$ [Egghe, Rousseau, 2008]	Взвешенный *h*-индекс $h_w = \sqrt{\sum_{i=1}^{r_0} c_i}$, где $r_0 = \max\left\{ j: \sum_{i=1}^{j} \left(c_i \Big/ h \right) \leq c_j \right\}$
$\boldsymbol{h_T}$ [Anderson, et al., 2008]	Индекс построен на основе представления цитирований всех публикаций в виде диаграммы Ферре (*Ferrers diagram*). Каждое цитирование получает вес, вес каждой публикации равен сумме весов его цитирований. Вес публикации ранга j:

	$h_{T(j)} = \begin{cases} \dfrac{c_j}{2j-1}, & c_j \le j, \\ \dfrac{j}{2j-1} + \displaystyle\sum_{i=j+1}^{c_j} \dfrac{1}{2i-1}, & c_j > j, \end{cases}$ $h_T = \displaystyle\sum_{j=1}^{n} h_{T(j)}$
f [Tol, 2009]	$f = \max\{r : r \Big/ \sum_{i=1}^{r} \frac{1}{c_i} \ge r\}$ (максимальный ранг *r*, такой что гармоническое среднее цитирований публикаций рангов от 1 до *r* больше или равно *r*)
e [Zhang, 2009]	$\sqrt{C(h) - h^2}$ учитывает дополнительные цитирования в *h*-ядре
π [Vinkler, 2009]	1/100 общего количества цитирований, полученных множеством, состоящим из $\sqrt{n}$ наиболее цитируемых публикаций
w [Wu, 2010]	$w = \max\{j: c_j \ge 10j\}$
h*²*-lower, ***h*²*-center,*** ***h*²*-upper*** [Bornmann, et al., 2010]	Оценка с помощью трех индексов $h^2\text{-}lower = \dfrac{\sum_{j=h+1}^{n} c_j}{\sum_{j=1}^{n} c_j} \cdot 100,$ $h^2\text{-}center = \dfrac{h \cdot h}{\sum_{j=1}^{n} c_j} \cdot 100,$ $h^2\text{-}upper = \dfrac{\sum_{j=1}^{h} \left(c_j - h \right)}{\sum_{j=1}^{n} c_j} \cdot 100$

	(*h^2-upper* – доля не учтенных при построении *h*-индекса цитирований публикаций в *h*-ядре; *h^2-center* – доля учтенных цитирований в *h*-ядре; *h^2-lower* – доля неучтенных цитирований публикаций вне *h*-ядра)
ch [Ajiferuke, Wolfram, 2010]	Пусть ct_j – количество авторов, цитирующих публикацию j в своих работах, и публикации упорядочены в порядке убывания количества цитирующих их авторов. $ch = \max\{j: ct_j > j\}$

На основе *h*-индекса разработаны новые методы и метрики, восполняющие информацию о цитировании, утерянную при определении *h*-индекса.

7.4.1. *h-последовательность Liang.* Определение *h*-последовательности (*h-sequence*) как средства для изучения механизма изменения *h*-индекса во времени и сравнения ученых с различными периодами карьеры введено в работе [Liang, 2006]. *h*-последовательность – последовательность *h*-индексов, вычисленных за увеличивающиеся промежутки времени в обратном порядке по годам, начиная с текущего года. Например, фиксируем 2004-й год, вычисляем *h*-индекс, получаем h_1 – первый элемент *h*-последовательности. Вычисляем *h*-индекс за период 2004–2003 гг., получаем h_2 – второй элемент *h*-последовательности. Далее вычисляем *h*-индекс за 2004–2002 гг., получаем h_3 – третий элемент *h*-последовательности. Выполняем аналогичные вычисления до некоторого фиксированного года. Например, если рассматривается период 2004–1976 гг., получится последовательность $h_1, h_2, \ldots, h_{29}$.

Для ранжирования ученых по этому принципу построим *h*-матрицу, столбцы которой соответствуют ученым, а строки – элементу *h*-последовательности: $h_1, h_2, \ldots, h_n$.

h-последовательность выявляет механизм увеличения *h*-индекса, а *h*-матрица позволяет сравнивать результативность ученых с

различной научной карьерой, если взять за начало год, до которого все ученые уже опубликовали первую статью. Можно построить новую матрицу для всех ученых за первые n лет. Можно для каждого ученого выбрать n самых продуктивных лет, а затем провести сравнение. В этом случае следует выбрать окно цитирования, например считать цитирования через m лет после публикации.

7.4.2. *h-последовательность Egghe.* В работе [Egghe, 2009] предложено строить h-последовательности с начала карьеры ученого. Пусть период карьеры описывается временными отрезками $t = 1, 2,\ldots, t_m$, где $t = 1$ соответствует первому году карьеры (первой публикации), а t_m – последнему (ограничившись, например, текущим годом). h-последовательность строится следующим образом. Для $t = 1$ рассматриваются только публикации и цитирования текущего года. Вычисляем h-индекс h_1. Для $t = 1$ и $t = 2$ вместе вычисляем h-индекс h_2, рассматривая все цитирования и публикации за этот период, и т. д. Последовательность $h_1, h_2,\ldots, h_{t_m}$ дает динамическое представление о карьере ученого и может использоваться для сравнения результативности ученых. Заметим, что h-индексы для h-последовательности Liang (обозначим как h_1^*, $h_2^*,\ldots, h_{t_m}^*$) автоматически получаются на основе данных из базы данных *WoS*, в то время как h-индексы для h-последовательности Egghe необходимо вычислять дополнительно.

В работе [Egghe, 2009] рассматривается непрерывный временной интервал $t \in \mathbb{R}^+$ и вычисляются указанные два типа последовательностей в рамках среды, где выполняется закон Лотки, связывающий источники и элементы (здесь публикации и цитирования):

$$f(j) = \frac{C}{j^{\alpha}}$$

где $C > 0$; $\alpha > 1$; $f(j)$ – плотность публикаций с плотностью j для цитирований [Egghe, 2005]. Предполагается, что α – константа для каждого периода времени. Поскольку временной интервал является непрерывным, вместо $h_1, h_2, \ldots, h_{t_m}$ рассматривается $h(t)$, вместо $h_1^*, h_2^*, \ldots, h_{t_m}^* - h(t)^*$.

Показано, что в общем случае графики функции различаются, т. е. последовательности ведут себя неодинаково. Это подтверждено эмпирически.

Внесено предложение включать в *WoS* вычисление h-последовательностей в автоматическом режиме, так как в отличие от единичного h-индекса они иллюстрируют эволюцию карьеры ученого. Очевидно, что ученый может самостоятельно накапливать h-индексы в течение карьеры, а затем строить h-последовательность, однако это требует желания и терпения.

7.4.3. *h-последовательность Randić.* В работе [Randić, 2009] предложен иной подход к расширению информации, используемой при вычислении h. Рассмотрим ранжированный по количеству цитирований список публикаций автора. При вычислении h ранги нумеруются от 1 до n (количество работ) и вычисляется $h = \max\{r: X_r^* \geq r\}$. Обозначим $H_0 = h$. Для построения H_1 удваиваем ранги и вновь вычисляем h. Вычисления продолжаем до тех пор, пока есть работы, цитирование которых превосходит ранг. Получаем последовательность $\{{}^0H, {}^1H, \ldots\}$, из которой вытекает H-последовательность $\{{}^0H, {}^0H + {}^1H, {}^0H + {}^1H + {}^2H, \ldots\}$. Последняя сумма и есть H-индекс H. Теперь вычисляем коэффициент $Q = H/h$, который рассматривается как мера потенциала ученого, построенная на основе истории цитирования.

В работе [Egghe, 2010] данный подход изучается в рамках системы, в которой выполняется закон Лотки. Определяются

формулы для *H*-индекса и *H*-последовательности, а также определяются условия, при которых в случае равенства *h*-индексов двух ученых (h_1=h_2) *H*-индексы будут не равны. *H*-индексы и *H*-последовательности несут дополнительную информацию о манере цитирования и позволяют ранжировать ученых, имеющих равные *h*-индексы.

7.4.4. *CSS-метки Egghe.* В работе [Egghe, 2010a] также ставится задача расширения возможностей оценки, основанной на *h*-индексе. Использована методика *Characteristic Scores and Scales* (*CSS*), представленная в работе [Glänzel, Schubert, 1988], применительно к *h*-индексу. Основная идея *CSS* заключается в том, что на основе ранжированного распределения цитирований определяется последовательность точек ν_k (k=1,2,..) и β_k (k=1,2,...), таких что ν_k – ранг публикации, а β_k=$\gamma(\nu_k)$ – количество цитирований публикации ранга $r = \nu_k$ (*Characteristic scores*). Здесь γ – ранговая функция частоты (*Rank-order frequency function*), $\gamma(r)$ − количество цитирований публикации ранга r.

Методология базируется на том факте, что многие публикации, а именно имеющие количество цитирований менее h, не принимаются к рассмотрению. Поэтому выполняется следующая процедура. Вычисляется h_0 − h-индекс на всем списке работ. Из списка работ удаляются относящиеся к h-ядру. Вычисляется h_1 − h-индекс для полученного списка. Вновь удаляются работы, относящиеся к h_1-ядру, вычисляется h_2 и т. д. Величины h_0, $h_0 + h_1$, $h_0 + h_1 + h_2$ и т. д. представляют ранги $\nu_k = \sum_{i=0}^{k-1} h_i$, а β_k=$\gamma(\nu_k)$. Заметим, что $\beta_1 = \nu_1 = h_0 = h$. Маркирующие точки хорошо отражают важность всего списка работ.

Список литературы

[Айвазян и др., 1989] Айвазян С. А., Бухштабер В. М., Енюков И. С., Мешалкин Л. Д.. Прикладная статистика. Классификация и снижение размерности М.: ФиС, 1989. 607 с.

[Ахо и др., 1979] Ахо А., Хопкрофт Дж., Ульман Дж. Построение и анализ вычислительных алгоритмов. М.: Мир, 1979. 536 с.

[Бериков, Лбов, 2008] Бериков В. С., Лбов Г. С. Современные тенденции в кластерном анализе // Всероссийский конкурсный отбор обзорно-аналитических статей по приоритетному направлению “Информационно-телекоммуникационные системы”. [Электрон. ресурс]. http://www.ict.edu.ru/ft/005638/62315e1-st02.pdf.

[Бернал, 1956] Бернал Дж. Наука в истории общества. М.: ИЛ, 1956. 735 с.

[Браверман, Мучник, 1983] Браверман Э. М., Мучник И. Б. Структурные методы обработки эмпирических данных. М., Наука, 1983. 464 с.

[Бредихин, Кузнецов, 2012] Бредихин С. В., Кузнецов А. Ю. Методы библиометрии и рынок электронной научной периодики. Новосибирск, Москва: ИВМиМГ СО РАН, НЭЙКОН, 2012. 254 с.

[Википедия] Википедия, свободная энциклопедия. [Электрон. ресурс]. http://ru.wikipedia.org/wiki/Main_Page.

[Гантмахер, 1988] Гантмахер Ф. Р. Теория матриц. М.: Наука, 1988. 548 с.

[Гарфилд, 1982] Гарфилд Ю. Можно ли выявлять и оценивать научные достижения и научную продуктивность? // Вестн. Акад. наук СССР. 1982. № 7. С. 42–50.

[Гмурман, 2004] Гмурман В. Е. Тория вероятностей и математическая статистика. М.: Высш. шк., 2004. 480 с.

[Грановский, 2000]. Грановский Ю. В. Можно ли измерять науку? Исследования В. В. Налимова по наукометрии. [Электрон. ресурс]. http://vivovoco.rsl.ru/vv/papers/bio/nalimov2.htm.

[Гумбель, 1965] Гумбель Э. Статистика экстремальных значений. М.: Мир, 1965. 451 с.

[Дэйвисон, 1988] Дэйвисон М. Многомерное шкалирование. Методы наглядного представления данных. М.: ФиС, 1988. 254 с.

[Дюран, Оделл, 1977] Дюран Б., Оделл П. Кластерный анализ. М.: Статистика, 1977. 128 с.

[Жамбю, 1988] Жамбю М. Иерархический кластер-анализ и соответствия. М.: ФиС, 1988. 345 с.

[Зиновьев, 2000] Зиновьев А. Ю. Визуализация многомерных данных. Красноярск: Изд-во. гос. техн. ун-та, 2000. 180 с.

[История ВИНИТИ]. История ВИНИТИ. [Электрон. ресурс]. http://www2.viniti.ru/.

[Кара-Мурза, 1989] Кара-Мурза С. Г. Проблемы интенсификации науки. М.: Наука, 1989, 248 с.

[Ким и др., 1989] Ким Дж. О., Мьюллер Ч. У., Клекка У. Р. Факторный, дискриминантный и кластерный анализ: Пер. с англ. М.: ФиС, 1989. 216 с.

[Королюк и др., 1985] Королюк В. С., Портенко Н. И., Скороход А. В., Турбин А. Ф. Справочник по теории вероятностей и математической статистике. М.: Наука, 1985. 640 с.

[Лившиц, 1976] Лившиц В. Скорость переработки информации человеком и факторы сложности среды // Тр. по психологии ТГУ. Тарту, 1976. С. 139—146.

[Логинов, 1996] Логинов Н. В. Сингулярное разложение матриц. М.: МГАПИ. 1996. 80 с.

[Лоули, Максвелл, 1967] Лоули Д., Максвелл А. Факторный анализ как статистический метод. М.: Мир, 1967. 144 с.

[Мандель, 1988] Мандель И. Д. Кластерный анализ. М.: ФиС, 1988. 176 с.

[Мандельброт, 2002] Мандельброт Б. Фрактальная геометрия природы. М.: Ин-т компьютерных исследований, 2002. 676 с.

[Маршакова, 1973] Маршакова И.В. Система связей между документами, построенная на основе ссылок: по данным Science Citation Index // НТИ. сер.2. 1973. № 6. С. 3−8.

[Маршакова, 1981]. Маршакова И. В. Сети цитирования: Информационные модели системы научных публикаций // Обзоры по электронной технике. Сер. Экономика и системы управления. М.: ЦНИИ “Электроника”, 1981. Вып. 1 (760). С. 1−48.

[Мат. энц.] Ветвящийся процесс. [Электрон. ресурс]. http://dic.academic.ru/contents.nsf/enc_mathematics/.

[Мирская, 2005] Мирская Е. З. Р. К. Мертон и этос классической науки. [Электрон. ресурс]. http://iph.ras.ru/page48033174.htm.

[Михайлов и др., 1968] Михайлов А. И., Черный А. И., Гиляревский Р. С. Основы информатики. М.: Наука, 1968. 757 с.

[Найссер, 1981] Найссер У. Познание и реальность. М.: Прогресс, 1981. 232 с.

[Налимов, 1966] Налимов В. В. Количественные методы исследования процесса развития науки // Вопросы философии. 1966. № 12. С. 38−47.

[Оре, 2008] Оре О. Теория графов. М.: УРСС, 2008. 352 с.

[Поппер, 1983] Поппер К. Логика и рост научного знания. М.: Прогресс, 1983. 330 с.

[Прайс, 1966] Прайс Д. Малая наука, большая наука. Сб. статей. Наука о науке. М.: Прогресс, 1966. С. 281−384.

[Прайс, 1971] Прайс Д. Квоты цитирования в точных и неточных науках, технике и не-науке // Вопросы философии. 1971. № 3. С. 149–155.

[РИНЦ, 2005] [Электрон. ресурс]. http://elibrary.ru/defaultx.asp/.

[Солтон, 1979] Солтон Дж. Динамические библиотечно-поисковые системы. М.: Мир, 1979. 557 с.

[Терехина, 1986] Терехина А.Ю. Анализ данных методами многомерного шкалирования. М.: Наука, 1986. 168 с.

[Хартман, 1972] Хартман К. Современный факторный анализ. М.: Статистика, 1972. 444 с.

[Шеннон, 1963] Шеннон К. Э. Математическая теория связи. Сб. К. Шеннон “Работы по теории информации и кибернетике”. М.: ИЛ, 1963. С. 243−322. 830 с.

[ЭЭиФН, 2009] Энциклопедия эпистемологии и философии науки. М.: “Канон+”, РООИ “Реабилитация”. 2009. 1248 с.

[Abt, 1980] Abt H. A. The cost-effectiveness in terms of publications and citations of various optical telescopes at the kitt peak national observatory // Publications of the Astronom. Soc. of the Pacific. 1980. V. 92. P. 249–254.

[Adler, Ewing, Taylor, 2009] Adler R., Ewing J., Taylor P. Citation statistics // Statistical Sciences. 2009. V. 24, P. 1−14. Пер. с англ. Адлер Р., Эвинг Д., Тэйлор П. Статистики цитирования // Игра в цыфирь, или как теперь оценивают труд ученого. М. МЦНМО, 2011. 72 с.

[Ahlgren, et al., 2003] Ahlgren P., Jarneving B., Rousseau R. Requirement for a citation similarity measure, with special reference to Pirson’s correlation coefficient // J. Amer. Soc. Inform Sci. Tech. 2003. V. 54, iss. 6. P. 550−560.

[Ajiferuke, Wolfram, 2010] Ajiferuke I., Wolfram D. Citer analysis as a measure of research impact: library and information science as a case study // Scientometrics. 2010. V. 83, iss. 3. P. 623−638.

[Anderson, et al., 2008] Anderson T. R., Hankin R. K. S., Killworth P. D. Beyond the Durfee square: Enhancing the h-index to score total publication output // Scientometrics. 2008. V. 76, iss. 3. P. 577–588.

[Arunachalam, 1998] Arunachalam S. Citation analysis: Do we need a theory? // Scientometrics. 1998. V. 43, iss. 1. P. 141–142.

[Baldi, 1998] Baldi S. Normative versus social constructivist processes in the allocation of citations: A network-analytic model // Amer. Sociologic. Rev. 1998. V. 63, P. 829–846.

[Barlup, 1969] Barlup J. Relevancy of cited articles in citation indexing // Bull. of the Medical Library Association. 1969. V. 57. P. 260–63.

[Barton, Kebler, 1960] Barton R. F., Kebler R. W. The half-life of some scientific and technical literatures // Amer. Doc. 1960. V. 11, N 1. P. 8−12.

[Bassecoulard, Zitt, 1999] Bassecoulard E., Zitt M. Indicators in a research institute: A multilevel classification of journals // Scientometrics. 1999. V. 44, iss. 3. P. 323–345.

[Batagelj, 2003] Batagelj V. Efficient algorithms for citation network analysis // arXiv:cs/0309023, 2003, 27 p.

[Batagelj, Mrvar, 1998] Batagelj V., Mrvar A. Pajek − program for large network analysis // Connections. 1998. V. 21, iss. 2. P. 47−57.

[Bateson, 1980] Bateson G. mind and nature. N.Y.: Bantam, 1980. 259 p.

[Bavelas, 1978] Bavelas J. B. The social psychology of citations // Canad. Psychol. Rev. 1978. V. 19. P. 158–163.

[Bayer, Folger, 1966] Bayer A., Folger J. Some correlates of a citation measure of productivity in science // Sociology of Education. 1966. V. 39. P. 381−390.

[Bazeman, 1988] Bazerman C. Shaping written knowledge: The genre and activity of the experimental article in science. Madison: Univ. of Wisconsin Press, 1988. 358 p.

[Bellardo, 1980] Bellardo T. The use of co-citations to study science // Library Res. 1980. V. 2. P. 231–237.

[Bensman, 1982] Bensman S. J. Bibliometric laws and library usage as a social phenomenon // Library Res. 1982. V. 4. P. 279–312.

[Bernal, 1939] Bernal J. D. The social function of science. London: George Routledge & Sons, 1939. 482 p.

[Bhattacharya, Basu, 1998] Bhattacharya S., Basu P. Mapping a research area at the micro level using co-word analysis // Scientometrics. 1998. V. 43, iss. 3. P. 359–372.

[Borgman, 1990] Borgman C.L. Editors' Introduction. In: Scholarly communication and bibliometrics. Newbury Part: Sage, 1990. P. 10–27.

[Börner, 2010] Börner K. Atlas of science: Visualizing what we know. London UK: MIT Press, 2010. 254 p.

[Börner, et al., 2003] Börner K., Chen C., Boyack K. W. Visualizing knowledge domains // Annual Rev. Inform. Sci. Technol. (ARIST). 2003. V. 37. P. 179–255.

[Bornmann, Daniel, 2005] Bornmann L., Daniel H-D. What do we know about the h index? // J. Amer. Soc. Inform. Sci. Tech. 2007. V. 58, iss. 9. P. 1381–1385.

[Bornmann, Daniel, 2005a] Bornmann L., Daniel H-D. Does the h-index for ranking of scientists really work? // Scientometrics. 2005. V. 65, iss. 3. P. 391–392.

[Bornmann, Daniel, 2008] Bornmann L., Daniel H-D. What do citation counts measure? A review of studies on citing behavior // J. Documentation. 2008. V. 64, iss. 1. P. 45–80.

[Bornmann, et al., 2008] Bornmann L., Mutz R., Daniel H-D. Are there better indices for evaluation purposes than the h index? A comparison of nine different variants of the h index using data from biomedicine // J. Amer. Soc. Inform. Sci. Tech. 2008. V. 59, iss.5. P. 830–837.

[Bornmann, et al., 2010] Bornmann L., Mutz R., Daniel H-D. The h index research output measurement: Two approaches to enhance its accuracy // J. Informetrics. 2010. V. 4, iss. 3. P. 407−414.

[Bornstein, 1991] Bornstein R. F. The predictive validity of peer review: A neglected issue // Behavioral and Brain Sciences. 1991. V. 14. P. 138−139.

[Boyack, 2009] Boyack K. W. Using detailed maps of science to identify potential collaborations // Scientometrics. 2009. V. 79, iss. 1. P. 27–44.

[Boyack, et al., 2002] Boyack K. W., Wylie B. N., Davidson G. S. Domain visualization using *VxInsight* for science and technology management // J. Amer. Soc. Inform. Sci. Tech. V. 53, iss. 9. P. 764–774.

[Boyack, et al., 2005] Boyack K. W., Klavans R., Börner K. Mapping the backbone of science // Scientometrics. 2005. V. 64, iss. 3. P. 351–374.

[Boyack, et al., 2009] Boyack K. W., Börner K., Klavans R. Mapping the structure and evolution of chemistry research // Scientometrics. 2009. V. 79, iss. 1. P. 45–60.

[Boyack, Klavans, 2010] Boyack K. W., Klavans R. Co-citation analysis, bibliographic coupling and direct citations. Which citation approach represents the research front most accurately // J. Amer. Soc. Inform. Sci. Tech. 2010. V. 61, iss. 12. P. 2389−2404.

[Braam, et al., 1988] Braam R. R., Moed H. F., Van Raan A. F. J. Mapping of science: critical elaboration and new approaches, a case study in agricultural biochemistry // Informetrics 87/88. Select Proc. of the 1st Intern. conf. on bibliometrics and theorical aspects of infor-

mation retrieval, Diepenbeek, August 25−28, 1987. Amsterdam: Elsevier Science, 1988. P. 15−28.

[Bradford, 1948] Bradford S C. Documentation. London: Crosby Lockwood & Sons, 1948. 137 p.

[Braun, et al., 2006] Braun T., Glänzel W., Schubert A. A Hirsch-type index for journals // Scientometrics. 2006. V. 69, iss. 1. P. 169−173.

[Broadus, 1977] Broadus R. N. An investigation of the validity of bibliographic citations // J. Amer. Soc. Inform. Sci. 1983. V. 34, iss. 2. P. 132–135.

[Brooks, 1985] Brooks T. A. Private acts and public objects: An investigation of citer motivations // J. Amer. Soc. Inform. Sci. 1985. V. 36, iss. 4. P. 223−229.

[Brooks, 1986] Brooks T. A. Evidence of complex citer motivations // J. Amer. Soc. Inform. Sci. 1986.V. 37, iss. 1. P. 34–36.

[Burrell, 2007] Burrell Q. L. On the h-index, the size of the Hirsch core and Jin's A-index // J. of Informetrics. 2007. V. 1, iss. 2. P. 170−177.

[Cahlik, 2000] Cahlik T. Comparison of the maps of science // Scientometrics. 2000. V. 49, iss. 3. P. 373–387.

[Callon, et al., 1983] Callon M., Courtial J-P., Turner W., Brain S. From translations to problematic networks: An introduction to co-word analysis // Social Sci. Inform. 1983. V. 22. P. 191−235.

[Callon, et al., 1986] Callon M., Law J., Rip J. Qualitative scientometrics. Mapping in dynamics of science and technology. London: The MacMillan Press, 1986. P. 103–123.

[Callon, Law, Rip, 1986] Callon M., Law J., Rip A. How to study the force of science. Mapping in Dynamics of Science and Technology. London: The MacMillan Press, 1986. P. 3–18.

[Cawkell, 1974] Cawkell A. E. Search strategy, construction and use of citation networks, within a socio-scientific example: Amorphous

constructor and S.R. Ovshinsky // J. Amer. Soc. Inform. Sci. 1974. V. 25, iss 2. P. 123−130.

[Chalmers, 1992] Chalmers M. BEAD: Explorations in information visualization // SIGIR'92. Copenhagen: ACM Press, 1992. P. 330–337.

[Chen, 1998a] Chen C. Bridging the gap: The use of Pathfinder networks in visual navigation // J. Visual Languages and Comput. 1998. V. 9, iss. 3. P. 267–286.

[Chen, 1998b] Chen C. Generalized similarity analysis and pathfinder network scaling // Interacting with Computers. V. 10, N 2. P. 107–128.

[Chen, 1999] Chen C. Visualizing semantic spaces and author co-citation networks in digital libraries // Inform. Proc. Management. 1999. V. 35, iss. 2. P. 401–420.

[Chen, et al., 1996] Chen H. C., Schuffels C., Orwig R. Internet categorization and search: A self-organizing approach // J. Visual Communication and Image Representation. 1996. V. 7, iss. 1. P. 88–102.

[Chen, et al., 2001] Chen C., Paul R. J., O'Keefe B. Fitting the jigsaw of citation: Information visualization in domain analysis // J. Amer. Soc. Inform. Sci. Technol. 2001. V. 52, iss. 4. P. 315–330.

[Chen, Lien, 2011] Chen L., Lien Y. Using co-citation analysis to examine the intellectual structure of e-learning. A MIS perspective // Scientometrics. 2011. V. 89, iss. 3. P. 867−886.

[Chen, Paul, 2001]. Chen C., Paul R. J. Visualizing a knowledge domain's intellectual structure // Computer, 2001. V. 34, iss. 3. P. 65–71.

[Chen, Rada, 1996] Chen C., Rada R. Modelling situated actions in collaborative hypertext databases // J. of Computer-Mediated Communications. V. 2, N 3. [Electron. resource]. http://jcmc.indiana.edu/vol2/issue3/chen.html.

[Chubin, Moitra, 1975] Chubin D. E., Moitra S. D. Content analysis of references adjunct or alternative to citation counting? // Social Studies of Sci. 1975. V. 5, N 4. P. 426−441.

[Cole, 1992] Cole S. Making science: Between nature and society. Cambridge, MA: Harvard Univ. Press. 1992. 290 p. [Electron. resource]. http://books.google.com.

[Cole, Cole, 1967] Cole S., Cole J. R. Scientific output and recognition. A study in the operation of the reward system in science // American Sociological Rev. 1967. V. 32. P. 377−390.

[Cole, Cole, 1971] Cole J. R., Cole S. Measuring the quality of sociological Resource: problem in the use of the Science Citation Index // The Amer. Sociologist. 1971. V. 6. P. 23–29.

[Cole, Cole, 1972] Cole J. R., Cole S. The Ortega hypothesis // Science. 1972. V. 178, N 4059. P. 368−375.

[Cole, Cole, 1973] Cole J. R., Cole S. Social stratification in science. Chicago: Univ. of Chicago Press. 1973. 283 p.

[Collins, 1986] Collins H. M. Changing order: Replication and induction in scientific practice. Chicago: Univ. of Chicago Press. 1986. 199 p.

[Courtial, 1994] Courtial J. P. A co-word analysis of scientometrics // Scientometrics. 1994. V. 31, iss. 3. P. 251−260.

[Cozzens, 1981] Cozzens S. E. Taking the measure of science: A review of citation theories // Intern. Soc. Sociology of Knowledge Newsletter. 1981. V. 7, N 1/2. P. 16–20.

[Cozzens, 1989] Cozzens S. E. What do citation counts? The rhetoric-first model // Scientometrics. 1989. V. 15, iss. 5-6. P. 437–447.

[Cozzens, et al., 1990] Cozzens S. E., Healy P., Rip A., Ziman J. The research system in transition. Dordrecht: Kluwer Academic. 1990. P. 387−401.

[Cronin, 1981] Cronin B. The need for a theory of citing // J. Documentation. 1981. V. 37, iss. 1. P. 16–24.

[Cronin, 1981a] Cronin B. Transatlantic citation patterns in educational psychology // Education Libraries Bull. 1981. V. 24. P. 48–51.

[Cronin, 1982] Cronin B. Norms and functions in citation: the view of journal editors and referees in psychology // Social Sci. Information Studies. 1982. V. 2, iss. 2. P. 65–78.

[Cronin, 1984] Cronin B. The citation process: The role and significance of citations in scientific communication. London: Taylor Graham. 1984. 103 p. [Electron. resource]. http://books.google.com.

[Cronin, 1994] Cronin B. Tiered citation and measures of document similarity // J. Amer. Soc. Inform. Sci. 1994. V. 45, iss. 7. P. 537−538.

[Cronin, Meho, 2006] Cronin B., Meho L. Using the h-index to rank influential information scientists // J. Amer. Soc. Inform. Sci. Tech. 2006. V. 57, iss. 9. P. 1275–1278.

[Davidson, et al., 1998] Davidson G. S., Hendrickson B., Johnson D. K., et al. Knowledge mining with VxInsight: Discovery through interaction // J. Intelligent Inform. Systems. 1998. V. 11, iss. 3. P. 259–285.

[Davidson, et al., 2001] Davidson G. S., Wylie B. N., Boyack K. W. Cluster stability and the use of noise in interpretation of Clustering // Proc. IEEE Information Visualization. N.Y.: IEEE Press, 2001. P. 23–30.

[Davies, 1970] Davies D. Citation Idiosyncrasies (letter to the editor) // Nature. 1970. V. 228. P. 1356.

[Deerwester, et al., 1990] Deerwester S., Dumais S. T., Landauer T. K., Furnas G. W., Harshman R. A. Indexing by latent semantic analysis // J. Amer. Soc. Inform. Sci. 1990. V. 41, iss. 6. P. 391–407.

[DeLooze, Lemarie, 1997] DeLooze M. A., Lemarie J. Corpus relevance through co-word analysis: An application to plant proteins // Scientometrics. 1997. V. 39, iss. 3. P. 267–280.

[Di Battista, 1999] Di Battista G. Graph drawing: Algorithms for the visualization of graphs / G. Di Battista, et al. N.J.: Prentice-Hall, 1999. 397 p.

[Di Battista, et al., 1994] Di Battista G., Eades P., Tamassia R., Tollis I. G. Algorithms for drawing graphs: An annotated bibliography // Computational Geometry: Theory and Applications. 1994. V. 4, iss. 5. P. 235–282.

[Ding, et al., 1999] Ding Y., Chowdhury G., Foo S. Mapping the intellectual structure of information retrieval studies: an author co-citation analysis 1987–1997 // J. Inform. Sci. 1999. V. 25, iss.1. P. 67–78.

[Ding, et al., 2000] Ding Y., Chowdhury G., Foo S. Journal as markers of intellectual space: Journal co-citation analysis of information Retrieval area 1987–1997 // Scientometrics. 2000. V. 47, iss. 1. P. 55–73.

[Dubes, Jain, 1976] Dudes R. C., Jain A. K. Clustering techniques: The user's dilemma // Pattern Recognition. 1976. V. 8. P. 247–260.

[Duhem-Quine Thesis] Тезис Дюгема–Куайна. [Electron. resource]. http://enc-dic.com/sociology/Tezis-Djugema-Kuajna-9459.html.

[Eades, 1984] Eades P. A heuristics for graph drawing // Congress Numeratium. 1984. V. 42. P. 149–160.

[Earle, Vickery, 1969] Earle P., Vickery B. Social science literature use in the UK as indicated by citations // J. Documentation. 1969. V. 25, iss. 2. P. 123–141.

[Edge, 1977] Edge D. Why I am not a co-citationist // Society for Social Studies in Sci.: Newsletter 2. 1977. P. 13−19.

[Edge, 1979] Edge D. Quantitative measures of communication in science: a critical review // History of Sci. 1979. V. 17. P. 102–134.

[Edwards, 1973] Edwards S., McCarrey M. Measuring the performance of researchers // Research Management. 1973. V. 16, N 1. P. 34–41.

[Egghe, 1994] Egghe L. Special features of the author-publication relationship and a new explanation of Lotka's law based on convolution theory // J. Amer. Soc. Inform. Sci. 1994. V. 45, iss. 6. P. 422–427.

[Egghe, 2005] Egghe L. Power laws in the information production process: Lotkian informetrics. Oxford, UK: Elsevier, 2005. 427 p.

[Egghe, 2006] Egghe L. Theory and practice of the g-index // Scientometrics. 2006. V. 69, iss. 1. P. 131–152.

[Egghe, 2009] Egghe L. Mathematical study of h-index sequences // Inform. Proc. Management. 2009. V. 45, iss. 2. P. 288–297.

[Egghe, 2010] Egghe L. Mathematical results on the Randić's H-index and H-sequence // Research evaluation. 2010. V. 19, iss. 3. P. 203–207.

[Egghe, 2010a] Egghe L. Characteristic scores and scales based on h-type indices // J. Informetrics. 2010. V. 4, iss. 1. P. 14–22.

[Egghe, Michel, 2002] Egghe L., Michel C. Strong similarity measures for ordered sets of documents in information Retrieval // Inform. Proc. Management. 2002. V. 38, iss. 6. P. 823–848.

[Egghe, Michel, 2003] Egghe L., Michel C. Construction of weak and strong similarity measures for ordered sets of documents using fuzzy set techniques // Inform. Proc. Management. 2003. V. 39, iss. 5. P. 771–807.

[Egghe, Rousseau, 1990] Egghe L., Rousseau R. Introduction to informetrics: Quantitative methods in library, documentation and information science. Amsterdam: Elsevier Science Publishers. 1990. 450 p.

[Egghe, Rousseau, 2006a] Egghe L., Rousseau R. An informetric model for the Hirsch index // Scientometrics. 2006. V. 69, iss. 1. P. 121–129.

[Egghe, Rousseau, 2006]. Egghe L., Rousseau R. Classical retrieval and overlap measures satisfy the requirements for rankings based on a

Lorenz curve // Inform. Proc. Management. 2006. V. 42, iss. 1. P. 106–120.

[Egghe, Rousseau, 2008] Egghe L., Rousseau R. An h-index weighted by citation impact // Inform. Proc. Management. 2008. V. 44, iss. 2. P. 770–780.

[Faigley, 1986] Faigley L. Competing theories of process // College English. 1986. V. 48, N 6. P. 527−542.

[Feng, Börner, 2002] Feng Y., Börner K. Using semantic treemaps to categorize and visualize bookmark files // Proc. of the SPIE, San Jose (CA), January 20–25 2002. SPIE Digital Library. 2002. V. 4655. P. 218–227. [Electron. resource]. http://proceedings.spiedigitallibrary.org/volume.aspx?volumeid=3265.

[Ferber, 1986] Ferber M. Citations: are they an objective measure of scholarly merit? // SIGNS. 1986. V. 11, N 2. P. 381–389.

[Fisher, Van Ness, 1971] Fisher L., Van Ness J. W. Admissible clustering procedures // Biometrika. 1971. V. 58, P. 91–104.

[Fruchterman, Reingold, 1991] Fruchterman T. M. J., Reingold E. M. Graph drawing by force-directed placement // Software – Practice & Experience. 1991. V. 21, N 11. P. 1129–1164.

[Gaillard, 1989] Gaillard J. La science du tiers monde est-elle visile? // La Recherche. 1989. N 210. P. 636–640.

[Garfield, 1964] Garfield E. The citaion index, − A new dimension in indexing // Science. 1964. V. 144, N 3619. P. 649–654.

[Garfield, 1965] Garfield E. Can Citation indexing be automated? Washington, DC: National Bureau of Standards, 1965. P. 189−192.

[Garfield, 1970] Garfield E. Citation indexing, historio-bibliography, and the sociology of science // Proc. of the 3d Intern. Congr. of Medical Librianship, Experta Medica, Amsterdam, P. 187−204. Reprinted

in: Essays of an Information Scientist. V. 1. Philadelphia: ISI Press, 1977. P. 158−174.

[Garfield, 1973] Garfield E. The new ISI journal citation reports should significantly affect the future course of scientific publication // Essays of an Information Scientist. Philadelphia: ISI Press, 1973. V. 1. P. 473–474.

[Garfield, 1977] Garfield E. To cite or not to cite: A note of annoyance // Current Contents. 1977. N 35. P. 5–8.

[Garfield, 1979] Garfield E. Citation indexing: Its theory and applications in science, technology, and humanities. N.Y.: Wiley, 1979. 274 p. [Electron. resource]. http://books.google.com.

[Garfield, 1979a] Garfield E. Journal citation studies 32. Canad. journals. Pt. 2: the analysis of Canad. resource published at home and abroad // Current Contents. 1979-80. N 32. P. 5–9.

[Garfield, 1980] Garfield E. Is information retrieval in the arts and humanities inherently different from that in science // Library Quarterly. 1980. V. 50, N 1. P. 40–57.

[Garfield, 1981] Garfield E. Introducing the ISI Atlas of Science: biochemistry and molecular biology, 1978/80 // Current Contents. 1981. N 42. P. 5−13. Reprinted in: Essays of an Information Scientist. Philadelphia: ISI Press, 1983. V. 5. P. 279−287.

[Garfield, 1982] Garfield E. More on the ethics of scientific publication: abuses of authorship attribution and citation amnesia undermine the reward system of science // Current Contents. 1982. N 30. P. 5–10.

[Garfield, 1983] Garfield E. How to use citation analysis for faculty evaluations, and when is it relevant? P. 2 // Current Contents. 1983. N 45. P. 5–14.

[Garfield, 1987] Garfield E. Launching the ISI Atlas of Science: for the new year, a new generation of reviews // Current Contents. 1987. N 1. P. 3−8.

[Garfield, 1989] Garfield E. Citation behavior – An aid or hindrance to information retrieval? // Current Contents. 1989. N 18. P. 3–8.
[Garfield, 1993] Garfield E. Co-citation analysis of the scientific literature: Henry Small on mapping the collective mind of science // Current Contents. 1993. N 19. P. 3−13. [Electron. resource]. http://www.garfield.library.upenn.edu/essays/v15p293y1992–93.pdf.

[Garfield, et al., 1978] Garfield E., Malin M. V., Small H. G. Citation data as science indicator. Towards a Metric of Science. N. Y.: John Wiley, 1978. P. 179−207.

[Garfield, et al., 2010] Garfield E., Pudovkin A. I., Paris S. A bibliometric and histographic analysis of the work of Tony van Raan: a tribute to a scientometrics pioneer and gatekeeper // Research evaluation. 2010. V. 19, iss. 3. P. 161–172.

[Garfield, Pudovkin, Istomin, 2003] Garfield E., Pudovkin A. I., Istomin V. S. Why do we need Algorithmic Historiography? // J. Amer. Soc. Inform. Sci. Tech. 2003. V. 54, iss. 5. P. 400–412.

[Gieryn, 1978] Gieryn T. F. Problem retention and problem change in science // Sociological Inquiry. 1987. V. 48. P. 96−115.

[Gilbert, 1977] Gilbert G. N. Referencing as persuasion // Social Studies of Sci. 1977. V. 7, N 1. P. 113−122.

[Gillie, 1980] Gillie O. Burt: the scandal and the cover-up // Bull. British Psychol. Soc. 1980. V. 33. P. 9–16.

[Gipp, 2011] Gipp B. Identifying related work and plagiarism by citation analysis // IEEE-TCDL Bull. 2011. V. 7.

[Gipp, Beel, 2009] Gipp B., Beel J. Citation Proximity Analysis (CPA) – A new approach for identifying related work based on Co-Citation Analysis // Proc. of the 12th Intern. conf. “Scientometrics and informetrics” (ISSI-09), Rio de Janeiro (Brazil), 2009. V. 2. P. 571−575.

[Glänzel, 2001] Glänzel W. National characteristics in international scientific coauthorship relations // Scientometrics. 2001. V. 51, iss. 1. P. 69–115.

[Glänzel, 2006] Glänzel W. On the opportunities and limitations of the H-index // Sci. Focus. 2006. V. 1, N 1. P. 10–11.

[Glänzel, 2006a] Glänzel W. On the h-index: A mathematical approach to a new measure of publication activity and citation impact // Scientometrics. V. 67, iss. 2. 2006. P. 315–321.

[Glänzel, 2008] Glänzel W. On some new bibliometric applications of statistics related to the h-index // Scientometrics. 2008. V. 77, iss. 1. P. 187–196.

[Glänzel, Czerwon, 1996]. Glänzel W., Czerwon H. J. A new methodological approach to bibliographic coupling and its application to the national, regional and institutional level // Scientometrics. 1996. V. 32, iss. 2. P. 195−221.

[Glänzel, DeLange, 1997] Glänzel W., De Lange C. Modelling and measuring multilateral co-authorship in international scientific collaboration. P. 2. A comparative study on the extent and change of international scientific collaboration links // Scientometrics. 1997. V. 40, iss. 3. P. 605–626.

[Glänzel, Schubert, 1988] Glänzel W, Schubert A. Characteristic scores and scales in assessing citation impact // J. Inform. Sci. 1988. V. 14, iss. 2. P. 123−127.

[Glänzel, Schubert, 2003] Glänzel W., Schubert A. A new classification scheme of science fields and subfields designed for bibliometric evaluation purposes // Scientometrics. 2003. V. 56, iss. 3. P. 357–367.

[Goodrich, Roland, 1977] Goodrich J., Roland C. Accuracy of published medical reference citations // J. Technical Writing and Communications. 1977. V. 7. P. 15–19.

[Griffith, et al., 1974] Griffith B. C. Small H. G., Stonehill J. A., Dey S. Structure of scientific literatures II: Towards a macro- and macrostructure for science // Science Studies. 1974. V. 4. P. 339−365.

[Hagstrom, 1971] Hagstrom W. O. Inputs, outputs, and the prestige of American Univ. Science. Departments // Sociology of Education. 1971. V. 44, N 4. P. 375–398.

[Harris, 1963] Harris T. E. The theory of branching processes. Berlin: Springer, 1963. 230 p.

[Harter, 1992] Harter S. P. Psychological relevance and information science // J. Amer. Soc. Inform. Sci. 1992. V. 43, iss. 9. P. 602−615.

[Harter, Nisonger, Weng, 1993] Harter S. P., Nisonger T. E., Weng A. Semantic relationships between cited and citing articles in library and information science journals // J. Amer. Soc. Inform. Sci. 1993. V. 44, iss. 9. P. 543−552.

[Harzing, 2008] Harzing A-W. Reflections on the h-index. [Electron. resource]. www.harzing.com/pop-hindex.htm.

[Harzing, Wal, 2008] Harzing A-W., van der Wal R. A Google Scholar H-Index for journals: A better metric to measure journal impact in economics & business // Acad. Management Annual Meet., Anaheim (US), Aug. 8–13, 2008. [Electron. resource]. http://www.harzing.com/papers.htm#hjournals.

[He, 1999] He Q. Knowledge discovery through co-word analysis // Library Trends. 1999. V. 48, N 1. P. 133–159.

[Hicks, Potter, 1991] Hicks D., Potter J. Sociology of scientific knowledge: A reflexive analysis of science disciplines and disciplining science // Social Studies of Sci. 1991. V. 21, N 3. P. 459–501.

[Hirsch, 2005a] Hirsch J. E. An index to quantify an individual's scientific research output // Proc. of the National Acad. Sci. USA. 2005. V. 102, N 46. P. 16569–16572.

[Hirsch, 2005b] Hirsch J. E. An index to quantify an individual's scientific research. [Electron. resource]. http://xxx.arxiv.org/abs/physics/0508025.

[Hjørland, 1997] Hjørland B. Information seeking and subject representation. An activity theoretical approach to information science. – Westport CT; London: Greenwood Press, 1997. P. 55–103.

[Hjørland, 2000] Hjørland B. Relevance research: The missing perspective(s): "Non-relevance" and "epistemological relevance" // J. Amer. Soc. Inform. Sci. 2000. V. 51, iss. 2. P. 209–211.

[Hjørland, 2002] Hjørland B. Epistemology and the socio-cognitive perspective in information science // J. Amer. Soc. Inform. Sci. Tech. 2002. V. 53, iss. 4. P. 257–270.

[Hummon, Doreian, 1989] Hummon N. P., Doreian P. Connectivity in a citation network: The development of DNA theory // Social Networks. 1989. V. 11, N. 1. P. 39–63.

[Hummon, Doreian, 1990] Hummon N. P., Doreian P. Computational methods for social network analysis // Social Networks. 1990. V. 12, N 4. P. 273–288.

[Inhaber, Alvo, 1978] Inhaber H., Alvo M. World science as an input-output system // Scientometrics. 1978. V. 1, iss. 1. P. 43–64.

[Jain, Dubes, 1988] Jain A. K., Dubes R. C. Algorithms for clustering data. N. J.: Prentice-Hall, 1988. 334 p.

[Jain, et al., 1999] Jain A. K., Murty M. N., Flynn P. J. Data clustering: A review // ACM Computing Surveys. 1999. V. 31, N. 3. P. 264−323.

[Jarneving, 2008] Jarneving B. A variation of the calculation of the first author cocitation strength in author cocitation analysis // Scientometrics. 2008. V. 77, iss. 3. P. 485–504.

[Jasanoff, 1990] Jasanoff S. The fifth branch: Science advisers as policymaker. Cambridge, MA: Harvard Univ. Press. 1990.

[Jin, 2006] Jin B.-H. H-index: An evaluation indicator proposed by scientist // Sci. Focus. 2006. V. 1, N 1. P. 8–9.

[Jin, et al., 2007] Jin B.-H., Liang L., Rousseau R., Egghe L The R- and AR-indices: Complementing the h-index // Chinese Sci. Bull. 2007. V. 52. P. 855–863.

[Johnson, Shneiderman, 1991] Johnson B., Shneiderman B. Tree-maps: a space-filling approach to the visualization of hierarchical information structures // Proc. of the IEEE conf. on visualization, San Diego (USA), 1991. IEEE Computer Society Press, 1991. P. 284–291.

[Kamada, Kawai, 1989] Kamada T., Kawai S. An algorithm for drawing general undirected graphs // Inform. Proc. Letters. 1989. V. 31, iss. 1. P. 7–15.

[Kaplan, 1965] Kaplan N. The norms of citation behavior: Prolegomena to the footnote // Amer. Documentation. 1965. V. 16, iss. 3. P. 179−184.

[Karypis, 2003] Karypis G. CLUTO: a clustering toolkit // Techn. rep. Department of Computer Sci. Univ. of Minnesota, 2003.

[Kelly, Jennions, 2006] Kelly C. D., Jennions, M. D. The h-index and career assessment by numbers // Trends in Ecology & Evolution. 2006. V. 21, N 4. P. 167–170.

[Kessler, 1963] Kessler M. M. Bibliographic coupling between scientific papers // Amer. Documentation. 1963. V.14, iss.1. P. 10–25.

[Kessler, 1963a] Kessler M. M. An experimental study of bibliographic coupling between technical papers // IRE Transactions PGIT. 1963. IT-9. P. 49.

[Kessler, 1965] Kessler M. M. Comparison of the results of bibliographic coupling and analytic subject indexing // Amer. Document. 1965. V. 16, iss. 3. P. 223−233.

[Klavans, Boyack, 2007] Klavans R., Boyack K. W. Is there a convergent structure of science? A comparison of maps using the ISI and Scopus databases // Proc. of the 11th Intern. conf. of the International Society for scientometrics and informetrics. Madrid (Spain). ISSI, 2007. P. 437–448.

[Klavans, Boyack, 2008] Klavans R., Boyack K. W. Thought leadership: A new indicator for national and institutional comparison // Scientometrics. 2008. V. 76, iss. 2. P. 239–250.

[Klavans, Boyack, 2009] Klavans R., Boyack K. W. Towards a consensus map of science // J. Amer. Soc. Inform. Sci. Tech. 2009. V. 60, iss. 3. P. 455–476.

[Kleiber, Kotz, 2003] Kleiber C., Kotz S. Statistical size distributions in economics and actuarial sciences. Hoboken: John Wiley & Sons, 2003. 353 p.

[Knorr-Cetina, 1981] Knorr-Cetina K. The manufacture of knowledge: An essay on the constructivist and contextual nature of science. Oxford: Pergamon. 1981.

[Knuth, 1993] Knuth D. E. The Stanford GraphBase. N. Y.: ACM Press, 1993. 482 p.

[Kochen, 1974] Kochen M. Principles of information retrieval. Los Angeles: Melville, 1974. 74 p.

[Kohonen, 1995] Kohonen T. Self-organizing maps. Berlin: Springer, 1995. 501 p.

[Kosmulski, 2006] Kosmulski M. A new Hirsch-type index saves time and works equally well as the original h-index // ISSI Newsletter. 2006. V. 2, N 3. P. 4–6.

[Kosmulski, 2007] Kosmulski M. MAXPROD – A new index for assessment of the scientific output of an individual, and a comparison with the h-index // Cybermetrics. 2007. V. 11, N 1. Paper 5.

[Krauze, McGinnis, 1979] Krauze T. K., McGinnis R. A matrix analysis of scientific specialties and careers in science // Scientometrics. 1979. V. 1, iss. 5−6. P. 419−444.

[Kruskal, 1964] Kruskal J. B. Multidimensional scaling by optimizing goodness-of-fit to a non-metric hypothesis // Psychometrika. V. 29. 1964. P. 1−27.

[Kruskal, 1977] Kruskal J. B. Multidimensional scaling and other methods for discovering structure. Statistical methods for digital computers. N. Y.: Wiley. 1977. P. 296–339.

[Landauer, et al., 1998] Landauer T., Foltz P., Laham D. Introduction to latent semantic analysis // Discourse Processes. 1998. V. 25. P. 259–284.

[Latour, 1987] Latour B. Science in action: How to follow scientists and engineers through society. Cambridge, MA: Harvard Univ. Press. 1987. 274 p. [Electron. resource]. http://books.google.com.

[Latour, Woolgar, 1986] Latour B., Woolgar S. Laboratory life: The construction of scientific facts. N. J.: Princeton, 1986. 297 p.

[Laudan, 1977] Laudan L. Progress and its problems: Toward a theory of scientific growth. Berkeley, CA: Univ. of California Press, 1977. 257 p. [Electron. resource]. http://books.google.com.

[Laudan, 1996] Laudan L. Beyond positivism and relativism: Theory, method, and evidence. Oxford: Westview Press, 1996. 277 p. [Electron. resource]. http://books.google.com.

[Law, Whittaker, 1992] Law J., Whittaker J. Mapping acidification research: A test of the co-word method // Scientometrics. 1992. V. 23, iss. 3. P. 417−461.

[Lawani, 1982] Lawani S. M. On the heterogeneity and classification of author self-citations // J. Amer. Soc. Inform. Sci. 1982. V. 33, iss. 5. P. 281–284.

[Lecler, Gagné, 1994] Lecler M., Gagné J. International scientific co-operation: The continentalization of science // Scientometrics. 1994. V. 31, iss. 3. P. 261–292.

[Lee, et al., 1977] Lee R. C. T., Slagle J. R., Blum H. A triangulation method for the sequential mapping of points from N-Space to Two-Space // IEEE Transact. Computer. 1977. V. 26. P. 288–292.

[Lehnmann, et al., 2005] Lehnmann S., Jackson A. D., Lautrup B. E. Measures and mismeasures of scientific quality. [Electron. resource]. http://arxiv.org/abs/physics/0512238.

[Leydesdorff, 1987] Leydesdorff L. Towards a theory of citation // Scientometrics. 1987. V 12, iss. 5−6. P. 305–309.

[Leydesdorff, 1994] Leydesdorff L. The generation of aggregated journal–journal citation maps on the basis of the CD-ROM version of the Science Citation Index // Scientometrics. 1994. V. 31, iss. 1. P. 59–84.

[Leydesdorff, 1998] Leydesdorff L. Theories of citation? // Scientometrics. 1998. V. 43, iss. 1. P. 5–25.

[Leydesdorff, 2007] The position of Tibor Braun's Œuvre: Bibliographic journal coupling. In the Multidimensional World of Tibor Braun. [Electron. resource]. http://www.issisociety.info/tiborbraun75/tiborbraun75.pdf.

[Leydesdorff, Bensman, 2006] Leydesdorff L., Bensman S. J. Classification and powerlaws: The logarithmic classification // J. Amer. Soc. Inform. Sc. Tech. 2006. V. 57, iss. 11. P. 1470–1486.

[Leydesdorff, Rafols, 2008] Leydesdorff L., Rafols I. A global map of science based on the ISI subject categories // J. Amer. Soc. Inform. Sci. Tech., 2009. V. 63, iss. 1. P. 373–362.

[Liang, 2006] Liang L. h-index sequence and h-index matrix: Constructions and applications // Scientometrics. 2006. V. 69, iss. 1. P. 153–159.

[Lin, 1997] Lin X. Map displays for information retrieval // J. Amer. Soc. Inform. Sci. 1997. V. 48, iss. 1. P. 40–54.

[Lin, et al., 1991] Lin X., Soergel D. Marchionini G. A self-organizing semantic map for information retrieval // Proc. of the 14th. Annual Intern. ACM/SIGIR Conf. on research & design in information retrieval, Chikago, N. Y.: ACM Press, 1991. P. 262–269.

[Lin, et al., 2007] Lin Y., Li W., Chen K., Liu Y. A document clustering and ranking system for exploring Medline citations // J. Amer. Medical Informatics Association. 2007. V. 14, N 5. P. 651−661.

[Lin, Kaid, 2000] Lin Y., Kaid L. L. Fragmentation of the intellectual structure of political communication study: Some empirical evidence // Scientometrics. 2000. V. 47, iss. 1. P. 143–164.

[Lindsey, 1978] Lindsey D. The scientific publication system in social science. San Francisco: Jossey-Bass, 1978. 169 p.

[Lindsey, 1978]. Lindsey D. The corrected quality ratio: A composite index of scientific contribution to knowledge // Social Studies of Sci. 1978. V. 8, N 3. P. 349–354.

[Lindsey, 1980] Lindsey D. Production and citation measures in the sociology of science: the problem of multiple authorship // Social Studies of Sci. 1980. V. 10, N 2. P. 145–162.

[Line, 1979] Line M. B. The influence of the type of sources used on the result of citation analysis // J. Documentation. 1979. V. 35, iss. 4. P. 265–284.

[Line, Sandison, 1974] Line M. B., Sandison A. ‘Obsolescence’ and changes in the use of literature with time // J. Documentation. 1974. V. 30, iss. 3. P. 283–350.

[Long, et al., 1980] Long J. S., McGinnis R., Allison P. D. The problem of junior-authored papers in constructing citation counts // Social Studies of Sci. 1980. V. 10, N 2. P. 127–143.

[Luukkonen, 1990] Luukkonen T. Citations in the rhetorical, reward and communications systems of science: PhD thesis / Acta Universitatis Tamperensis. ser A. 1990. V. 285. Tampere: Univ. of Tampere. 55 p.

[Luukkonen, et al., 1992] Luukkonen T., Persson O., Sivertsen G. Understanding patterns of international scientific collaboration // Sci., Technol., Human Values. 1992. V. 17, N 1. P. 101–126.

[MacRoberts, MacRoberts, 1986] MacRoberts M. H., MacRoberts B. R. Quantitative measures of communication in science: A study of the formal level // Social Studies of Sci. 1986. V. 16, N 1. P. 151−172.

[MacRoberts, MacRoberts, 1989] MacRoberts M. H., MacRoberts B. R. Problems of citation analysis: A critical review // J. Amer. Soc. Inform. Sci. 1989. V. 40, iss. 5. P. 342–349.

[MacRoberts, MacRoberts, 1996] MacRoberts M. H., MacRoberts B. R. Problems of citation analysis // Scientometrics. 1996. V. 36, iss. 3. P. 435−444.

[Mahlck, Persson, 2000] Mahlck P., Persson O. Socio-bibliometric mapping of intra-departmental networks // Scientometrics. 2000. V. 49, iss. 1. P. 81–91.

[Malin, 1968] Malin M. V. The science citation index: a new concept in indexing // Library Trends. 1968. V. 16, N 3. P. 374−387.

[Martin, Irvine,1983] Martin B. R., Irvine J. Assessing basic research: some partial indicators of scientific progress in radio astronomy // Research Policy. 1983. V. 12. P. 61–90.

[Martyn, 1964] Martyn J. Bibliographic coupling // J. Documentation. 1964. V. 20, iss. 4. P. 236.

[May, 1967] May K. O. Abuses of citation indexing // Science. 1967. V. 156, N 3777. P. 890–892.

[McCain, 1990] McCain K. W. Mapping authors in intellectual space: A technical overview // J. Amer. Soc. Inform. Sci. 1990. V. 41, iss. 6. P. 433–443.

[McCain, 1995] McCain K. W. The structure of biotechnology R & D. // Scientometrics. 1995. V. 32, iss. 2. P. 153–175.

[McCain, 1998] McCain K. W. Neural networks research in context: A longitudinal journal cogitation analysis of an emerging interdisciplinary field // Scientometrics. 1998. V. 41, iss. 3. P. 389–410.

[Meadow, 1974] Meadow A. J. Communication in science. London: Butterworths, 1974. 248 p.

[Merton, 1957] Merton R. K. Priorities in scientific discovery: a chapter in the sociology of science // Amer. Sociological Rev. 1957. V. 22. P. 635–659.

[Merton, 1968] Merton R. K. The Matthew Effect in science: The reward and communication systems of science are considered // Science. 1968. V. 159, N 3810. P. 56–63.

[Merton, 1970] Merton R. K. Science, Technology & Society in Seventeenth-Century England. N.Y.: Howard Fertig, 1970. 279 P. [Electron. resource]. http://books.google.com.

[Merton, 1973] Merton R. K. The sociology of science: Theoretical and empirical investigations. Chicago: Chicago Univ. Press. 1973. 639 P. [Electron. resource]. http://books.google.com.

[Merton, 1977] Merton R. K. The sociology of science: An episodic memoir. The sociology of science in Europe. Carbondale: Southern Illinois Univ. Press, 1977. P. 3−141.

[Merton, 1995] Merton R. K. The Thomas theorem and the Matthew effect // Social Forces.1995. V. 74. P. 379−424.

[Merton, 1996] Merton R. K. The Matthew Effect in Science, II: Cumulative Advantage and the Symbolism of intellectual property. Merton R. K. On Social Structure and Science. Chicago: The Univ. Chicago Press, 1996. P. 318–336.

[Miller, 2006] Miller C. W. Superiority of the h-index over the impact factor for physics. [Electron. resource]. arXiv:physics/0608183.

[Miroiu, 2013] Miroiu A. Axiomatizing of the Hirsch index: quantity and quality disjoined // J. Informetrics. 2013. V. 7, iss.1. P. 10–15.

[Moed, 1989] Moed H. F. Bibliometric measurement of research performance and price's theory of differences among the sciences // Scientometrics. 1989. V. 15, iss. 5−6. P. 473–483.

[Moed, 2005] Moed H. F. Citation analysis in resource evaluation // Ser.: Information Science and Knowledge Management. Springer. 2005. V. 9. 346 p.

[Moed, et al., 1985a] Moed H. F., Burger W. J. M., Frankfort J. G., Van Raan A. F. J. The use of the bibliometric data fir the measurement of university resource performance // Res. Policy. 1985. V. 14. P. 131–149.

[Moed, et al., 1985b] Moed H. F., Burger W. J. M., Frankfort J. G., Van Raan A. F. J. A comparative study of bibliometric past performance analysis and peer judgement // Scientometrics. 1985. V. 8, iss. 3−4. P. 149–159.

[Moed, Garfield, 2003] Moed H. F., Garfield E. Basic scientists cite proportionally fewer "authoritative" references as their bibliographies become shorter // Proc. of the 9 th Intern. Conf. on Scientometrics and Informetrics. Dalian: Dalian Univ. Tech. Press, 2003. P. 190−196.

[Moed, Vriens, 1989] Moed H. F., Vriens M. Possible inaccuracies in citation analyses // J. of Inform. Sci. 1989. V. 15, iss. 2. P. 95−107.

[Moravcsik, Murugesan, 1975] Moravcsik M. J., Murugesan. P. Some results on the function and quality of citations // Social Studies of Sci. 1975. V. 5, N 1. P. 86–92.

[Moya-Anegón et al., 2007] Moya-Anegón F., Vargas-Quesada B., Chinchilla-Rodríguez Z., et al. Visualizing the marrow of science // J. Amer. Soc. Inform. Sci. Tech. V. 58, iss. 14. P. 2167–2179.

[Mulkay, 1974] Mulkay M. J. Methodology in the sociology of science: Some reflections on the study of radio astronomy // Social Sci. Information. 1974. V. 13, N 2. P. 107–119.

[Myers, 1970] Myers C. R. Journal citations and scientific eminence in contemporary psychology // Amer. Psychologist. 1970. V. 25. P. 1041–1048.

[Narin, 1976] Narin F. Evaluative Bibliometrics: The use of publication and citation analysis in the evaluation of scientific activity. N. J.: Cherry Hill, Computer Horizons Inc. 1976. 466 p.

[Nederhof, Van Raan, 1987] Nederhof A. J., Van Raan A. F. J. Citation theory and the Ortega hypothesis // Scientometrics. 1987. V. 12, iss. 5−6. P. 325–328.

[Nickles, 1981] Nickles T. What is a problem that we may solve it? // Synthese. 1981. V. 47. P. 85−118.

[Nicolaisen, 2002] Nicolaisen J. The J-shaped distribution of citedness // J. Documentation. 2002. V. 58, iss. 4. P. 383−395.

[Nicolaisen, 2003] Nicolaisen J. The social act of citing: Towards new horizons in citation theory // J. Amer. Soc. Inform. Sci. Tech. 2003. V. 40, iss.1. P. 12−20.

[Nicolaisen, 2004] Nicolaisen J. Social behavior and scientific practice: Missing pieces of the citation puzzle: PhD thesis. Copenhagen:

Royal School of Library and Information Science, 2004. 214 p. [Electron. resource]. http://books.google.com.

[Nicolaisen, 2007] Citation analysis // Annual Review of Information Science and Technology. 2007. V. 41. P. 609–641.

[Nieuwenhuysen, Rousseau, 1988] Nieuwenhuysen P., Rousseau R. A quick and easy method to estimate the random effect on citation measures // Scientometrics. 1988. V. 13, iss. 1−2. P. 45–52.

[Noyons, et al., 1999a] Noyons E. C. M., Moed H. F., Van Raan A. F. J. Integrating research performance analysis and science mapping // Scientometrics. 1999. V.46, iss. 3. P. 591–604.

[Noyons, 2001] Noyons E. Bibliometric mapping of science in a science policy context // Scientometrics. 2001. V. 50, iss. 1. P. 83–98.

[Noyons, et al., 1999] Noyons E. C. M., Moed H.F., Luwel M. Combining mapping and citation analysis for evaluative bibliometric purposes: A bibliometric study // J. Amer. Soc. Inform. Sci. 1999. V. 50, iss. 2. P. 115–131.

[Noyons, VanRaan, 1998] Noyons E. C. M., Van Raan A. F. J. Advanced mapping of science and technology // Scientometrics. 1998. V. 41, iss.1−2. P. 61–67.

[parsCit] [Электрон. ресурс]. http://wing.comp.nus.edu.sg/parsCit/.

[Pearson, 1901] Pearson K. On lines and planes of closets fit to systems of points in space // Philosophical Magazine. 1901. S. 2. P. 559–572.

[Peters, Van Raan, 1993] Peters H. P. F., Van Raan A. F. J. Co-word-based science maps of chemical engineering. P. 1: Representations by direct multidimensional scaling // Research Policy. 1993. V. 22, iss. 1. P. 23–45.

[Pinski, Narin, 1976] Pinski G., Narin F. Citation influence for journal aggregates of scientific publications: theory with application to the

literature of physics // Inform. Proc. Management. 1976. V. 12, iss. 5. P. 297–312.

[Polanco, et al., 2001] Polanco X., Francois C., Lamirel J. C. Using artificial neural networks for mapping of science and technology: A multi-self-organizing-maps approach // Scientometrics. 2001. V. 51, iss. 1. P. 267–292.

[Porter, 1977] Porter A. L. Citation analysis: queries and caveats // Social Studies of Sci. 1977. V. 7, N 2. P. 257–267.

[Price, 1961] Price D. Science since Babylon. New Haven: Yale Univ. Press., 1961. 149 p.

[Price, 1963] Price D. Little science, big science. N. Y.: Columbia Univ. Press., 1961. 119 p.

[Price, 1965] Price D. Networks of scientific papers // Science. 1965. V. 149, N 3683. P. 510–515.

[Price, 1966] Price D. The Science of scientists // Medical Opinion & Review. 1966. V. 1, N 1. P. 88–97.

[Price, 1969] Price D. Ethics of citations // Special Libraries. 1969. V. 60, iss. 7. P. 468.

[Price, 1970] Price D. Citation measures of hard science, soft science, technology and non-science. Lexington, Mass: D.C. Heath & Co., 1970. P. 3–22.

[Price, 1979] Price D. The citation cycle // Essays of an Information Scientist. 1979-1980. V. 4. P. 621−633.

[Price, 1980] Price D. Towards a comprehensive system of science indicators. Conference on Evaluation in Science and Technology – Theory and Practice, Dubrovnik, July, 1980. Scientia Yugoslavica. 1980. V. 6. P. 45−65.

[Price, 1981]. Price D. The analysis of scientometric matrices for policy implication // Scientometrics. 1981. V. 3, iss. 1. P. 47−53.

[Pudovkin, Garfield, 2009] Pudovkin A. I., Garfield E. Percentile rank and author superiority indexes for evaluating individual journal articles and the author's overall citation performance // Proc. of the 5th Intern. conf. WIS and 10th COLLNET Meet., Dalian (China), 2009. [Electron. resource]. http://garfield.library.upenn.edu/papers/pudovkingarfielddalianchina2009.pdf.

[Quesada, 2009] Quesada A. Monotonicity and the Hirsch index // J. Informetrics. 2009. V. 3, iss. 2. P. 158–160.

[Quesada, 2010] Quesada A. More axiomatics for the Hirsch index // Scientometrics. 2010. V. 82, iss. 2. P. 413–418.

[Raghupathi, Nerrur, 1999] Raghupathi W., Nerur S. P. Research themes and trends in artificial intelligence: An author co-citation analysis // Intelligence. 1999. V. 10, iss. 2. P. 18–23.

[Randić, 2009] Randić M. Citations versus limitations of citations: beyond Hirsch index // Scientometrics. 2009. V. 80, iss. 3. P. 809–818.

[Ravetz, 1973] Ravetz J. R. Scientific knowledge and social problems. Oxford, UK: Clarendon Press. 1973. 449 p. [Electron. resource]. http://books.google.com.

[Rogers, Kincaid, 1981] Rogers E. M., Kincaid D. L. Communication networks: toward a new paradigm for research. N. Y.: The Free Press, 1981. 386 p.

[Rosvall, Bergstrom, 2008] Rosvall M., Bergstrom C. T. Maps of random walks on complex networks reveal community structure // Proc. of the National Acad. Sci. USA, 2008. V. 105, N 4. P. 1118–1123.

[Rudd, 1977] Rudd E. The effect of alphabetical order of author listing on the career of scientists // Social Studies of Sci. 1977. V. 7, N 2. P. 268–269.

[Salton, 1963] Salton G. Associative document retrieval techniques using bibliographic information // J. of the ACM. 1963. V. 10, iss. 4. P. 440−457.

[Salton, MacGill, 1983] Salton G., MacGill M. J. Introduction to modern information retrieval. N. Y.: McGraw-Hill, 1983. 448 p.

[Salton, Wong, Yang, 1975] Salton G., Wong A., Yang C. S. A vector space model for automatic indexing // Communications of the ACM. 1975. V. 18, N 11. P. 613–620.

[Schneider, et al., 2009] Schneider J. W., Larsen B., Ingwersen P. A comparative study of first and all-author co-citation counting and two different matrix generation approaches applied for author cocitation analyses // Scientometrics. 2009. V. 80, iss. 1. P. 105−132.

[Schreiber, et al., 2012] Schreiber M., Malesios C. C., Psarakis S. Exploratory factor analysis for the Hirsch index, 17 h-type variants, and some traditional bibliometric indicators // J. Informetrics. 2012. V. 6, iss. 3. P. 347–358.

[Schubert, Glänzel, 2007] Schubert A., Glänzel W. A systematic analysis of Hirsch-type indices for journals // J. Informetrics. 2007. V. 1, iss. 3. P. 179–184.

[Schvaneveldt, 1990] Schvaneveldt R. W. Pathfinder associative networks: Studies in knowledge organization. Norwood: Ablex Publ., 1990. 315 p.

[Schvaneveldt, et al., 1989] Schvaneveldt R. W., Durso F. T., Dearholt D. W. Network structures in proximity data // The psychology of learning and motivation: Advances in research and theory. V. 24. N. Y.: Acad. Press, 1989. P. 249–284.

[Sen, Gan, 1983] Sen S. K., Gan S. K. A mathematical extension of the idea of bibliographic coupling and its applications // Annals of Library Sci. Document. 1983. V. 30. P. 78−82.

[Shapin, 1995] Shapin S. Here and everywhere: Sociology of scientific knowledge // Annual Rev. Sociology. 1995. V. 21. P. 289−321.

[Sharabchiev, 1989] Sharabchiev J. T. Cluster analysis of bibliographic references as scientometric method // Scientometrics. 1989. V. 15, iss. 1−2. P. 127–137.

[Shaw, 1985] Shaw W. M. Critical thresholds in co-citation graph // J. Amer. Soc. Inform. Sci. 1985. V. 36, iss. 1. P. 38–43.

[Sidiropuolos, et al., 2007] Sidiropoulos A., Katsaros D., Manolopoulos Y. Generalized Hirsch h-index for disclosing latent facts in citation networks // Scientometrics. 2007. V. 72, iss. 2. P. 253–280.

[Simkin, Roychowdhury, 2003] Simkin M. V., Roychowdhury V. P. Read before you cite! // Complex Systems. 2003. V. 14. P. 269 – 274.

[Simkin, Roychowdhury, 2007a] Simkin M. V., Roychowdhury V. P. A mathematical theory of citing, 2007, arXiv:physics/0504094v3.

[Simkin, Roychowdhury, 2007b] Simkin M. V., Roychowdhury V. P. An introduction to the theory of citing, 2007, arXiv:math/0701086v1.

[Small, 1973] Small H. Co-citation in the scientific literature: A new measure of the relationship between two documents // J. Amer. Soc. Inform. Sci. 1973. V. 24, iss. 4. P. 265–269.

[Small, 1974] Small H. Multiple citation patterns in scientific literature: the circle and the hill models // Information Storage and Retrieval. 1974. V. 10. P. 393−402.

[Small, 1977] Small H. G. Co-citation model of a scientific specialty: – a longitudinal study of collagen research // Social Studies of Sci. 1977. V. 7, N 2. P. 139–166.

[Small, 1978] Small H. G. Cited documents as concept symbols // Social Studies of Sci. 1978. V. 8, N 3. P. 327−340.

[Small, 1986] Small H. The synthesis of specialty narratives from co-citation clusters // J. Amer. Soc. Inform. Sci. 1986. V. 37, iss. 3. P. 97–110.

[Small, 1987] Small H. The significance of bibliographic references // Scientometrics. 1987. V. 12, iss. 5-6. P. 339–341.

[Small, 1994] Small H. A SCI-map case study: building a map of AIDA research // Scientometrics. 1994. V. 30, iss. 1. P. 229–241.

[Small, 1995] Small H. Navigating the citation network // Proc. of the 58th Annual Meet. Amer. Soc. for Information Science, Chicago (USA), 1995. V. 32. P. 118–126.

[Small, 1997] Small H. Update on science mapping: creating large document space // Scientometrics. 1997. V. 38, iss. 2. P. 275–293.

[Small, 1998] Small H. A general framework for creating large-scale maps of science in two or three dimensions: The SciViz system // Scientometrics. 1998. V. 41, iss. 1−2. P. 125–133.

[Small, 1999] Small H. Visualizing science by citation mapping // J. Amer. Soc. Inform. Sci. 1999. V. 50, iss. 9. P. 799–813.

[Small, 1999a] Small H. A passage through science: Crossing disciplinary boundaries // Library Trends. 1999. V. 48, N 1. P. 72–108.

[Small, 2000] Small H. Charting pathways through science: Exploring Garfield's vision of a unified index to science / In: The web of knowledge: a Festschrift in honor of Eugene Garfield. Medford, N. J.: Information Today, 2000. P. 449−473.

[Small, Garfield, 1985] Small H., Garfield E. The geography of science: disciplinary and national mappings // J. Inform. Sci. 1985. V. 11, iss. 4. P. 147–59.

[Small, Griffith, 1974] Small H., Griffith B. C. The structure of scientific literatures. 1: Identifying and graphing specialties // Science Studies. 1974. V. 4. P. 17–40.

[Small, Sweeney, 1985] Small H, Sweeney E. Clustering the science citation index using co-citations. 1. A comparison of methods // Scientometrics. 1985. V. 7, iss. 3−6. P. 301−409.

[Small, Sweeney, Greenlee, 1985] Small H., Sweeney E., Greenlee E. Clustering the science citation index using co-citations. 2. Mapping Science // Scientometrics. 1985. V. 8, iss. 5-6. P. 321–340.

[Smith, 1981] Smith L. C. Citation analysis // Library Trends. V. 30, N 1. 1981. P. 83–106.

[Sokal, Sneath, 1963] R. R. Sokal, P. H. A. Sneath. Principles of numerical taxonomy. San Francisco: Freeman, 1963. 359 p.

[Soper, 1976] Soper M.E. Characteristics and use of personal collections // Library Quarterly. 1975. V. 46, N 4. P. 397–415.

[Steinbach, et al., 2000] Steinbach M., Karypis G., Kumar V. A comparison of document clustering techniques // KDD Workshop on Text Mining, Boston (USA), Aug., 2000. N. Y.: ACM Press, 2000. P. 109−110.

[Steinhaus, 1956] Steinhaus H. Sur la division des corps matériels en parties // Bull. Acad. Polon. Sci. 1956. V. 4. P. 801−804.

[Sullivan, et al., 1977] Sullivan D., White D. H., Barboni E. J. Co-citation analyses of science: an evaluation // Social Studies in Sci. 1977. V. 7, N 2. P. 223−240.

[Tagliacozzo, 1977] Tagliacozzo S. Self-citation in scientific literature // J. Documentation. 1977. V. 33, iss. 4. P. 251–265.

[Tarjan, 1983] Tarjan R. E. Data structures and network algorithms. Philadelphia: Soc. Industrial and Appl. Mathematics, 1983. 131 p.

[Thorne, 1977] Thorne F. C. The citation index: another case of spurious validity // J. Clinical Psychol. 1977. V. 33. P. 1157–1161.

[Tol, 2009] Tol R. S. J. The h-index and its alternatives: An application to the 100 most prolific economists // Scientometrics. 2009. V. 80, iss. 2. P. 317–324.

[Tryon, 1939] Tryon R. C. Cluster analysis. London: Ann Arbor Edwards Bros. 139 p.

[Turner, et al., 1988] Turner W., Chartron G., Laville F., Michehet B. Packaging information for peer review: new co-word analysis techniques / In: Handbook of quantitative studies of science and technology. Amsterdam: North Holland, 1988. P. 291−323.

[Van der Veer Martens, 2001] Van der Veer Martens. B. Do citation systems represent theories of truth? // Information Research. 2001. V. 6, N 2. [Electron. resource]. http://informationr.net/ir/6-2/paper92.html.

[Van Eck, et al., 2008] Van Eck N. J., Waltman L., Noyons E. C. M., Buter R. K. Automatic term identification for bibliometric mapping. Techn. Rep. ERS-2008–081–LIS, Erasmus Univ. Rotterdam, Erasmus Research Institute of Management, 2008.

[Van Eck, et al., 2006] Van Eck N. J., Waltman L., Van den Berg J., Kaymak U. Visualizing the computational intelligence field // IEEE Comput. Intelligence Magazine. 2006. V. 1, N 4. P. 6–10.

[Van Eck, et al., 2010] Van Eck N. J., Waltman L., Dekker R., Van den Berg J. A comparison of two techniques for bibliometric mapping: Multidimensional scaling and VOS // J. Amer. Soc. Inform. Sci. Tech. 2010. V. 61, iss. 12. P. 2405–2416.

[Van Eck, Waltman, 2007] Van Eck N. J., Waltman L. Bibliometric mapping of the computational intelligence field // Intern. J. Uncertainty, Fuzziness and Knowledge-Based Systems. 2007. V. 15, N 5. P. 625–645.

[Van Eck, Waltman, 2007a] Van Eck N. J., Waltman L. VOS: A new method for visualizing similarities between objects // Proc. of the 30th Annual conf. of the German Classification Society, Studies in Classification, Data Analysis, and Knowledge Organization. Springer, 2007. P. 299–306.

[Van Eck, Waltman, 2008] Van Eck N. J., Waltman L. Appropriate similarity measures for author co-citation analysis // J. Amer. Soc. Inform. Sci. Tech. 2008. V. 59, iss. 10. P. 1653–1661.

[Van Eck, Waltman, 2009] Van Eck N. J., Waltman L. How to normalize cooccurrence data? An analysis of some well-known similarity measures // J. Amer. Soc. Inform. Sci. Tech. 2009. V. 60, iss. 8. P. 1635−1651.

[Van Eck, Waltman, 2010] Van Eck N. J., Waltman L. Software survey: VOSviewer, a computer program for bibliometric mapping // Scientometrics. 2010. V. 84, iss. 2. P. 523–538.

[Van Raan, 1990] Van Raan A. F. J. Fractal dimension of cocitations // Nature. 1990. V. 347. P. 626. [Electron. resource]. http://garfield.library.upenn.edu/histcomp/van-raan_fractals-nature/index-tl.html.

[Van Raan, 1991] Van Raan A. F. J. Fractal geometry of information space as represented by co-citation clustering // Scientometrics. 1991. V. 20, iss. 3. P. 439–449.

[Van Raan, 1998] Van Raan A. F. J. In matters of quantitative studies of science the fault of theorists is offering too little and asking too much // Scientometrics. 1998. V. 43, iss. 1. P. 129−139.

[Van Raan, 2006] Van Raan A. F. J. Comparison of the Hirsch-index with standard bibliometric indicators and with peer judgment for 147 chemistry research groups // Scientometrics. 2006. V. 67, iss. 3. P. 491–502.

[Vehlo, 1986] Vehlo L. The “meaning” of citation in context of scientific peripheral country // Scientometrics. 1986. V. 9, iss. 1−2. P. 71–89.

[Vehlo, 1987] Vehlo L. The author and the beholder: how paradigm commitments can influence the interpretation of Resource results // Scientometrics. 1987. V. 11, iss. 1−2. P. 59–70.

[Vinkler, 1986] Vinkler P. Evaluation of some methods for relative assessment of scientific publications // Scientometrics. 1986. V. 10, iss. 3−4. P. 157–177.

[Vinkler, 1987] Vinkler P. A quasi-quantitative citation model // Scientometrics. 1987. V. 12, iss. 1−2. P. 47–72.

[Vinkler, 2009] Vinkler P. The π-index: A new indicator for assessing scientific impact // J. Inform. Sci. 2009. V. 35, iss. 5. P. 602–612.

[Virgo, 1977] Virgo J. A. A statistical procedure for evaluating the importance of scientific papers // Library Quarterly. 1977. V. 47, N 4. P. 415–430.

[Waltman, Van Eck, 2007] Waltman L., Van Eck N. J. Some comments on the question whether co-occurrence data should be normalized // J. Amer. Soc. Inform. Sci. Technol. 2007. V. 58, iss. 11. P. 1701–1703.

[Webb, 1966] Webb E. Unobtrusive measures: nonreactive resource in the social sciences. Chicago: Rand McNally, 1966. 225 p.

[Weinberg, 1974] Weinberg B. H. Bibliographic coupling: A review // Information Storage and Retrieval. 1974. V. 10. P. 189–196.

[Weiner, 1977] Weiner J. The footnote fetish // Telos. 1977. N 31. P 172–177.

[Weinstock, 1971] Weinstock M. Citation indexes / In: Encyclopedia of Library & Information Sci. N. Y.: Marcel Dekker, 1971. V. 5. P. 16–40.

[White, 1990] White H. D. Author co-citation analysis: Overview and defense / In: Scholarly Communication and Bibliometrics. Newbury Park: Sage, 1990. P. 84−106.

[White, 2004] White H. D. Reward, persuasion, and the Sokal hoax: A study in citation identities // Scientometrics. 2004. V. 60, iss. 1. P. 93−120.

[White, Griffith, 1981] White H. D., Griffith B. C. Author cocitation: A literature measure of intellectual structure // J. Amer. Soc. Inform. Sci. 1981. V. 32, iss. 3. P. 163–171.

[White, Griffith, 1983] White H. D., Griffith B. C. Authors as makers of intellectual space: co-citation in studies of science, technology and society // J. of Documentation. 1982. V. 38, iss. 4. P. 252–272.

[White, McCain, 1997] White H. D., McCain K. W. Visualization of literatures // Annual Rev. Inform. Sci. Technol. 1997. V. 32. P. 99–168.

[White, McCain, 1998] White H. D., McCain K. W. Visualizing a discipline. An author co-citation analysis of information science, 1972−1995 // J. Amer. Soc. Inform. Sci. 1998. V. 49, iss. 4. P. 327−355.

[White, Wang, 1997] White M. D., Wang P. A qualitative study of citing behavior: Contributions, criteria, and metalevel documentation concerns // Library Quarterly. 1997. V. 67, N 2. P. 122−154.

[Whitley, 1984] Whitley R. The intellectual and social organization of the sciences. Toronto: Oxford Univ. Press, 1984. 319 p.

[Wikipedia] Wikipedia, free encyclopedia. [Electron. resource]. http://en.wikipedia.org/wiki/Main_Page.

[Woeginger, 2008] Woeginger G. J. An axiomatic characterization of the Hirsch-index // Mathematical Social Sci. 2008. V. 56. P. 224–232.

[Woolgar, 1976] Woolgar S. Writing an intellectual history of scientific development: The use of discovery accounts // Social Studies of Sci. 1976. V. 6. P. 395−422.

[Wouters, 1998] Wouters P. The signs of science // Scientometrics. 1998. V. 41, iss. 1−2. P. 225−241.

[Wouters, 1999a] Wouters P. Beyond the holy grail: From citation theory to indicator theory // Scientometrics. 1999. V. 44, iss. 3. P. 561−580.

[Wouters, 1999b] Wouters P. The Citation Culture: PhD thesis. University of Amsterdam. 1999. 278 P. [Electron. resource]. http://garfield.library.upenn.edu/wouters/wouters.pdf.

[Wu, 2010] Wu Q. The w-index: A measure to assess scientific impact by focusing on widely cited papers // J. Amer. Soc. Inform. Sci. Tech. 2010. V. 61, iss. 3. P. 609–614.

[Ye, 2009] Ye F. Y. An investigation on mathematical models of the h-index // Scientometrics. 2009. V. 81, iss. 2. P. 493−498.

[Zhang, 2009] Zhang C. The e-index, complementing the h-index for excess citations // PLoS ONE. 2009. V. 4, iss. 5:e5429.

[Zitt, et al., 2000] Zitt M., Bassecoulard E., Okubo Y. Shadows of the past in international cooperation: Collaboration profiles of the top five producers of science // Scientometrics. 2000. V. 47, iss. 3. P. 627–657.

[Zuckerman, 1968] Zuckerman H. Patterns of name-ordering among authors of scientific papers: a study of social symbolism and its ambiguity //Amer. J. Sociol. 1968. V. 74. P. 276–291.

[Zuckerman, 1987] Zuckerman H. Citation analysis and the complex problem of intellectual influence // Scientometrics. 1987. V. 12, iss. 5−6. P. 329−338.

Указатель обозначений

Символ	Значение
$+\infty$ $-\infty$	Бесконечность. Символы $+\infty$ и $-\infty$ применяются для определения граничных значений и сходимости
$P(X)$; p_i	Вероятность случайного события X; вероятность того, что объект i удовлетворяет некоторым условиям
$P(A\|B)$	Вероятность условная. $P(A\|B)$ означает вероятность события A, вычисленная в предположении, что событие B наступило
$\prec$	Доминирование. $X \prec Y$ означает “вектор Y доминирует над вектором X”
$\div$	Запись $1 \div r$ означает “от 1 до r”
:=	Знак определения. $x := y$ означает “x по определению равен y”
$\equiv$	Знак тождества. Равенство, выполняемое на всем множестве значений входящих в него переменных
$\Rightarrow$	Импликация. $A \Rightarrow B$ означает “из посылки A следует B”
[a, b]	Интервал закрытый. $\{x \in R \mid a \le x \le b\}$
(a, b)]a, b[	Интервал открытый. $\{x \in R \mid a < x < b\}$
(a, b]]a, b]	Интервал, открытый слева. $\{x \in R \mid a < x \le b\}$
[a, b) [a, b[	Интервал, открытый справа. $\{x \in R \mid a \le x < b\}$
$\forall$	Квантор всеобщности. $\forall x\ V(x)$ означает “$V(x)$ верно для всех значений x”
$\exists$	Квантор существования. $\exists x\ V(x)$ означает “существует хотя бы одно значение x, такое, что верно $V(x)$”
cov (X, Y)	Ковариация случайных величин X и Y
$\sqrt{}$	Корень квадратный $\sqrt{x}$ означает неотрицательное действительное число, которое в квадрате дает x
$\log_a b$	Логарифм b по основанию a
$\vee$	Логическая операция дизъюнкция. Результат $(A \vee B)$ истинен, когда хотя бы один из операндов A или B истинен

$\wedge$	Логическая операция конъюнкция. Результат ($A \wedge B$) истинен тогда и только тогда, когда оба операнда A и B истинны
$\neg$	Логическая операция отрицание. Результат $\neg A$ истинен тогда и только тогда когда операнд A ложен
max	Максимум функции или множества
$E(X)$, $\mu(X)$	Математическое ожидание случайной величины X
A^{T}	Матрица транспонирование. A^{T} – получена путем замены строк на столбцы в матрице A, $A^{\mathrm{T}}[i, j] = A[j, i]$
$[w_{ij}]$	Матрица элементов w_{ij}
Diag (A)	Матрица, главная диагональ которой совпадает с главнай диагональю матрицы A, а значения остальных элементов равны нулю
min	Минимум функции или множества
$\subseteq$ $\subset$	Множество A является подмножеством B. $A \subseteq B$ означает “каждый элемент из A также является элементом из B”; $A \subset B$ означает “каждый элемент из A также является элементом из B, однако B имеет как минимум один элемент не входящий в A”
$\mathbb{R}$	Множество вещественных (или действительных) чисел
$\mathbb{R}^+$	Множество вещественных положительных чисел
$\mathbb{N}$	Множество натуральных чисел без нуля, $\mathbb{N} = \{1, 2, 3, ...\}$
$\mathbb{N}_0$	Множество натуральных чисел с нулем, $\mathbb{N}_0 = \{0, 1, 2, 3, ...\}$
{ }	Множество элементов, например $\{a, b, c\}$
| |	Модуль. Если x переменная, то $\|x\|$ означает абсолютную величину x; если A конечное множество, то $\|A\|$ означает мощность множества A; если v – вектор, то $\|v\|$ означает длину вектора
$\left\| \cdot \right\|$	Норма в нормированном векторном пространстве. Обобщенное понятие длины вектора
$\bigcup$	Объединение множеств. Запись $A \bigcup B$ означает множество элементов, принадлежащих A или B (или обоим сразу)
$\rightarrow$	Отображение. Запись f: $X \rightarrow Y$ означает функцию f с областью определения X и областью значений Y
$\bigcap$	Пересечение множеств. Запись $A \bigcap B$ означает множество элементов, принадлежащих и A и B

$\approx$	Приблизительное равенство. Пример, $e \approx 2{,}718$ с точностью до 10^{-3}
$\in$ $\notin$	Принадлежность множеству. Запись $a \in S$ означает "a является элементом множества S"; запись $a \notin S$ означает "a не является элементом множества S"
$\mathbb{R}^k$	Пространство действительных чисел размерности k, $k \in \mathbb{N}$
$\mathbb{E}^k$	Пространство евклидово размерности k, $k \in \mathbb{N}$
$\varnothing$	Пустое множество. $\varnothing$ не содержит ни одного элемента
$\mathbb{R}\backslash\{0\}$	Разность множеств. Запись $\mathbb{R}\backslash\{0\}$ означает множество вещественных чисел, исключая ноль
$O(n)$	Сложность алгоритма. Абстрактное время исполнения, зависящее от параметра n
C_n^m	Сочетание. Число комбинаций из n элементов по m
$\gg$	Сравнение. Много больше
$\ll$	Сравнение. Много меньше
$\sim$	Сравнение. Одного порядка, эквивалентны
$\leq$ $\geq$	Сравнение. Запись $x \leq y$ означает, что "x меньше или равен y"; $x \geq y$ означает, что "x больше или равен y"
$\overline{x}$	Среднее значение. $\overline{x} = 1/n\ (x_1+x_2+ \ldots +x_n)$.
$\sum$	Сумма $\sum_{k=1}^{n} a_k$ означает "сумма a_k, где $1 \leq k \leq n$"
$\Delta(ABC)$	Угол ABC с вершиной B
{ | }	Условие. Запись $\{x\|P(x)\}$ означает множество всех x, для которых верно $P(x)$
arccos m	Функция арккосинус числа m
cos x	Функция косинус угла x
π	Число Пи $\approx 3{,}14159265$
$(X)_{ij}$	Элемент i , j матрицы X

Указатель терминов и сокращений

Индекс научного цитирования российский (РИНЦ)	*RINC*-Index	122
Индекс подобия обобщенный	Generalized similarity index (S_G)	193
Индекс Прайса	Price index (*PI*)	121
Индекс Солтона	Salton index	170
Индекс цитирования	Citation index	12
Индикатор цитирования *PRI*	Percentile rank index (*PRI*)	259
Институт научной информации	Institute for scientific information (*ISI*)	92
Историограф	Historiograph	259
Карта иерархическая	H-map	253
Карта науки	Map of science	219
Карта центрической формы	C-map	253
Карта кольцевидной формы	R-map	253
Кластеризация иерархическая объединяющая / разъединяющая	Agglomerative / divisive hierarchical clustering	205
Кластеризация методом *k*-means	*K*-means clustering	209
Кластеризация методом одиночной связи	Single-linkage clustering	208
Кластеризация с варьированием уровней	Variable level clustering	213
Кластеризация с учетом плотности	Density clustering; Density-based clustering	181
Когнитивная психология (КП)	Cognitive psychology	29
Концепции информации	Information concepts	45
Коцитирование (КЦ)	Co-citation	150
Коэффициент *RC*	Relative Co-Citation	170
Коэффициент ассоциативности	Association strength	188
Коэффициент БиС (КБС)	Coupling bibliographic strength	145
Коэффициент КЦ (ККЦ)	Co-citation frequency	150
Критерий χ^2 (Пирсона)	Chi-square test	166

Отчет о цитировании журналов	Journal citation report (*JCR*)	93
Оценка результатов кластеризации	Cluster validity	179
Передовые исследования (ПИ)	Research fronts	25
Программный продукт *HistCite*	Software *HistCite*	258
Программный продукт *Pajek*	Software *Pajek*	209
Программный продукт *SciViz*	Software *SciViz*	247
Программный продукт *VOSviewer*	Software *VOSviewer*	264
Процесс производства информации	Information production processes (*IPP*)	110
Распределение Парето	Pareto distribution	278
Расстояние Евклидово	Euclidean distance	202
Расстояние линейное	Linear distance	247
Расстояние логарифмическое	Logarithmic distance	247
Расстояние манхэттенское	Manhattan distance	202
Расстояние Минковского	Minkowski distance	202
Расстояние обратное	Inverse distance	247
Расстояние Чебышевское	Chebyshev distance	202
Реферативный журнал (РЖ)	Abstracts (Reviews) journal	91
Самоорганизующиеся карты	Self-organizing maps (*SOM*)	234
Самоцитирование	Self-citation	71
Самоцитирование диахронное	Diachronous self-citation	72
Самоцитирование синхронное	Synchronous self-citation	72
Сеть цитирования	Citation network	130
Силовой алгоритм размещения	Force directed placement (*FDP*)	238
Соавторство	Co-authorship	68
Список “*B*”	Brooks list	59
Список “*C*”	Chubin-Moitra list	56

Список “*M*”	Moravcsik, Murugesan list	55
Список “*T*”	Thorne list	58
Список “*W*”	Weinstock list	57
Ссылка	Reference	15
Триангуляция	Triangulation	236
Указатель научного цитирования	Science citation index (*SCI*)	17
Указатель цитирования “Общественные науки”	Social sciences citation index (*SSCI*)	241
Фактор оперативности	Immediacy factor	118
Факторный анализ	Factor analysis	227
Хирша индекс, *h*-индекс	Hirsch index, *h* index	267
Хирша ядро	Hirsch core	267
Цикл цитирования	Citation cycle	48
Цикл экспертной оценки	Peer review cycle	46
Цитата, цитирование	Citation	15
Цитировать	Cite	11
Частота терма	Term frequency (*tf*)	140
Число цитирований на одну публикацию (среднее)	Average citations per publication (*CPP*)	279
Школа бихевиористская (БШ)	Behavior school	28
Школа гештальтская (ГШ)	Gestalt school	28
h-подобные индексы	*h*-type indices	286
h-последовательность	*h*-sequence	286

ОГЛАВЛЕНИЕ

Сергей Всеволодович Бредихин,
Александр Юрьевич Кузнецов,
Наталья Григорьевна Щербакова

АНАЛИЗ ЦИТИРОВАНИЯ В БИБЛИОМЕТРИИ

Редактор *Л. Н. Ковалева*
Верстка *О. Г. Заварзина*

Подписано в печать 30.04.13. Формат 60×84/16. Печать офсетная.
Усл печ. л. 19,2. Уч.-изд. л. 12,8. Тираж 100 экз. Заказ № 72.

Отпечатано в типографии ООО "Омега Принт", 630090, Новосибирск,
просп. Акад. Лаврентьева, д. 6. Тел. (3832) 335-65-23