

Использование связей цитирования для наукометрических измерений в системе Соционет*

М.Р. Когаловский, Институт проблем рынка РАН

С.И. Паринов, Центральный экономико-математический институт РАН

Соционет: Электронный депонент, 2009.

В работе рассматривается подход к поддержке связей цитирования в электронных публикациях системы Соционет, особенности описания таких связей, а также те новые возможности, которые они могут обеспечить для эконометрических измерений. Предлагаемый подход в настоящее время реализуется в составе программного обеспечения системы.

Введение

В ряде существующих в настоящее время крупных научных электронных библиотек поддерживаются связи цитирования между электронными публикациями, содержащимися в их коллекциях информационных ресурсов. Необходимые для этой цели механизмы имеются, в частности, в электронной библиотеке ACM (Association for Computing Machinery) [1]. Для каждой представленной в этой библиотеке публикации пользователь может запросить содержащийся в ней полный библиографический список использованных источников, а также список тех публикаций в библиотеке, в которых на нее имеются ссылки. Электронная библиотека ACM содержит полные тексты публикаций в издаваемых ACM журналах, текстах докладов, опубликованных в трудах проводимых под эгидой ACM конференций, и другие материалы.

Более развитые возможности предоставляет созданная в 1997 году в NEC Research Institute (Принстон, штат Нью-Джерси, США) на основе технологии Инициативы открытых архивов (Open Archives Initiative) научная электронная библиотека публикаций по информатике CiteSeer [2]. С 2003 года система поддерживается в Колледже Информационных наук и технологий Пенсильванского университета (The Pennsylvania State University's College of Information Sciences and Technology), и в настоящее время в ее коллекциях уже содержится более 750 тыс.

* Работа поддерживается грантом РГНФ 09-02-12117в

публикаций. В этой библиотеке автоматически создается индекс цитирования, который может использоваться для поиска и оценки публикаций, поддерживается статистика цитирования. Важно отметить, что наряду со ссылкой на данный источник в некоторой работе система может выдавать контекст цитирования данной публикации. Это позволяет пользователю получать информацию о том, по какому поводу другие авторы ссылаются на данную публикацию. Система CiteSeer обеспечивает также ряд других полезных возможностей использования ссылок.

Для наукометрических измерений на основе связей цитирования важную роль сыграло создание крупных международных индексов цитирования. Термин «*индекс цитирования*» означает показатель значимости публикации, определяемый на основе статистической оценки множества ссылок на нее в других публикациях [3]. Индексом цитирования иногда называют также информационную систему, которая на основе мониторинга публикаций на заданном корпусе периодических изданий формирует и поддерживает индексы цитирования в указанном выше смысле для публикуемых в них статей.

Наиболее авторитетным международным индексом цитирования статей в периодических изданиях по фундаментальным областям науки является Science Citation Index (SCI) [4]. Он был создан в 1960 году и поддерживался Институтом научной информации (Institute for Scientific Information, ISI) в Филадельфии, США. В настоящее время он принадлежит информационной компании Thomson Reuters. Корпус периодических изданий, библиографию которых поддерживает SCI, включает около 6500 ведущих англоязычных научных журналов. В настоящее время существует веб-версия этого индекса цитирования – Web of Science, которая стала теперь частью проекта Web of Knowledge.

Наряду с индексом цитирования статей, опубликованных в периодических изданиях, Институтом информации в Филадельфии был создан индекс цитирования научных журналов Journal Citation Reports (JCR) [5]. Задачей этого индекса является оценка научной значимости журналов, входящих в корпус анализируемых и оцениваемых в его рамках периодических изданий. Формальной оценкой значимости журнала считается его так называемый *импакт-фактор*. Импакт-фактор журнала вычисляется как отношение числа ссылок в других журналах корпуса на статьи, опубликованные в данном журнале в некоторый период времени, к общему числу опубликованных в этом журнале статей за этот период. Обычно импакт-фактор

пересчитывается с периодом в два года. При формировании значений индекса цитирования каждой статьи в Science Citation Index учитываются импакт-факторы журналов, содержащих ссылки на данную статью.

К сожалению, русскоязычные научные журналы не входят в корпус периодических изданий, обрабатываемых в указанных индексах цитирования. Учитывая это обстоятельство, с 2005 года на основе Научной библиотеки РФФИ (НЭБ РФФИ) eLibrary.ru начал формироваться Российский индекс научного цитирования (РИНЦ) [6].

Система Соционет обладает в настоящее время некоторыми механизмами для наукометрических измерений на корпусе представленных в ней электронных публикаций [7-10]. Разрабатываемые в составе ее программного обеспечения дополнительные средства для исследования связей цитирования представленных в системе электронных публикаций не предназначены для выполнения функций индекса цитирования. Хотя в коллекциях информационных ресурсов системы в настоящее время уже представлено около 900 тыс. различных материалов, для реального использования получаемых наукометрических данных необходим более представительный корпус публикаций. Вместе с тем, возможность наукометрических измерений в поддерживаемом системой информационном пространстве также представляет несомненный интерес. Авторы предполагают использовать систему Соционет также и как полигон для отработки некоторых новых технологий электронной публикации научных документов, а также некоторых оригинальных возможностей наукометрических измерений на множестве связей цитирования, которые рассматриваются далее в этой работе.

Поддержка связей цитирования в системе Соционет

Мы не рассматриваем здесь организацию информационных ресурсов системы Соционет. Она достаточно подробно описана в работах [9-11]. Напомним лишь, что для каждой электронной публикации в системе создается ее описатель (карточка), содержащий метаданные, состав которых зависит от типа публикации. Для поддержки связей цитирования между представленными в системе электронными публикациями потребовалось дополнить множество используемых в системе типов информационных объектов и, соответственно, состав используемых метаданных.

Некоторые основные моменты реализации в Соционет поддержки связей цитирования (включая семантические атрибуты связей) заимствованы из спецификации европейской модели научных данных CERIF [12].

Форма, в которой выполняется традиционное научное цитирование, при ее перенесении в процедуры электронного депонирования требует пересмотра и развития с целью приведения ее в соответствие, с одной стороны, с новыми возможностями, а с другой – с требованиями развития процедур оценивания результатов исследований. В связи с данной ситуацией возникает следующий вопрос: *Как в процедуре электронного депонирования нужно изменить схему научного цитирования, чтобы получить максимально полные и точные данные об использовании результатов исследований?*

На наш взгляд, решение этой задачи обеспечивается при удовлетворении следующих требований:

- Материалы, представляющие результаты исследований, должны иметь электронную форму представления, упрощающую прямое и однозначное электронное цитирование каждого содержащегося в них результата в отдельности.

- Цитирование должно означать установление прямой и однозначной связи между представленными в электронной форме результатами (или материалами в целом), один из которых является используемым результатом (материалом), а второй – результатом (материалом), использующим первый для получения и представления нового знания.

- Схема электронного цитирования должна позволять автору цитирующего (второго из названных выше) результата (материала) уточнить, в каком именно качестве он использовал цитируемый (первый из названных выше) результат (материал), например, включив в связи цитирования определенные качественные атрибуты, которые, в свою очередь, должны быть сконструированы так, чтобы допускать их автоматическое распознавание и обработку для выявления различных характеристик использования.

Для каждой электронной публикации, представляемой в системе Соционет, в составе ее метаданных предусмотрено поле для списка использованных источников. Интерфейсы Соционет позволяют авторам статей задать из этого поля прямую связь как с документом источником, так и непосредственно с цитируемым фрагментом его текста (см. ниже). Кроме этого для каждой такой связи автор может указать набор

семантических атрибутов, определяющих в виде структурированных данных характер использования автором цитируемого им результата чужих исследований.

В целях обеспечения авторов научных статей удобными средствами формирования связей цитирования в Соционет реализуется подход, названный "*гранулированием*" научных статей. Гранулирование представляет собой выделение из текста статьи смысловых элементов, которые цитируются или могут цитироваться, и их оформление как самостоятельных информационных объектов типа "*цитата*" (citation).

Эта операция может выполняться в процессе цитирования чужих результатов или заранее авторами статей и научных результатов.

Операция гранулирования упрощает организацию процесса многократного и многопланового использования (цитирования) результатов исследований. Она также облегчает автоматический мониторинг за процессами использования результатов исследований.

Предусматривается следующая структура для объектов типа "цитата":

- смысловой элемент статьи, выделенный для обеспечения возможности повторного использования (цитирования);
- заголовок для данного элемента;
- ссылка на статью-источник;
- информация для потенциальных пользователей цитат, например, рубрикаторы областей возможного применения и рекомендации по применению.

Предполагается размещение коллекций информационных объектов типа "цитата" в научном информационном пространстве в открытом доступе, что создает возможности их многократного повторного использования всеми желающим. Контекст связей цитирования и значение атрибутов этих связей могут быть доступны авторам целевых публикаций ссылок, как и другим пользователям системы, по запросам к специальному системному сервису.

Виды ссылок, поддерживаемые в системе

Как отмечалось выше, предполагается, что ссылки цитирования могут снабжаться авторами содержащих их публикаций значениями специального атрибута, определяющего характер ссылки. Допустимое множество значений этого атрибута зависит от реализации механизмов инфраструктуры системы Соционет, обеспечивающей обработку и анализ связей цитирования. В предпринимаемой в

настоящее время реализации прототипа такой инфраструктуры принято следующее множество возможных значений атрибута вида ссылки:

1. Цитируемая публикация служит основой получения результатов, обсуждаемых в данной публикации.
2. Данная публикация подтверждает результаты, излагаемые в цитируемой работе.
3. Цитируемый результат переоткрывает результаты, ранее полученные автором данной публикации.
4. Результаты данной публикации опровергают результаты, изложенные в цитируемой работе.
5. Обсуждаемый в данной публикации результат обобщает результаты цитируемой публикации.
6. Обсуждаемый результат является частным случаем результата, излагаемого в цитируемой публикации и т.д.

Сервисы для обработки связей цитирования

Для обработки и анализа связей цитирования в системе Соционет, а также для исполнения пользовательских запросов, связанных с использованием связей цитирования, в настоящее время создается пилотный образец соответствующих компонентов программного обеспечения системы. Среди них важное место занимает специальный поисковый сервис, который позволит по запросам автора представленной в системе публикации, ее читателей, авторов публикаций, связанных ссылками цитирования с данной работой, получать сведения, связанные с характером поддерживаемых в системе связей цитирования.

Предполагается, что такой поисковый сервис будет обрабатывать набор запросов, позволяющих получать следующие сведения, которые при необходимости могут быть ассоциированы с каким-либо периодом времени:

1. Набор связей цитирования других публикаций библиотеки с данной публикацией заданных (или любых) видов, которые были установлены в любом или каком-либо указанном периоде времени (до, после данного момента, в заданном промежутке времени).
2. Аналогичный запрос относительно связей цитирования, исходящих из данной публикации.

3. Количество связей цитирования данной публикации в других публикациях библиотеки заданных (или любых) видов, которые были установлены в любом или каком-либо указанном периоде времени (до, после данного момента, в заданном промежутке времени).

4. Количество связей цитирования других публикаций в библиотеке в данной публикации заданных (или любых) видов, которые были установлены в любом или каком-либо указанном периоде времени (до, после данного момента, в заданном промежутке времени).

Предусматривается также разработка ряда программных компонентов, в том числе, и онлайн-сервисов, необходимых для мониторинга и поддержки динамики на множестве связей цитирования в системе, необходимых для поддержки технологии «живых» документов. Эта тема подробно обсуждается в работах [13, 14].

Заключение

Создание в системе Соционет механизмов поддержки метаданных, описывающих связи цитирования между представленными в системе электронными публикациями, а также функциональных компонентов для их анализа и обработки, обеспечивает новые возможности для развития функциональных возможностей средств наукометрических измерений в системе. Представляется важным включение в состав метаданных, которые описывают связи цитирования, наряду с контекстом их вхождения в публикацию также и сведений о семантике ссылки в форме атрибута, определяющего вид ссылки. Использование такого атрибута позволяет получить дополнительные возможности для автоматизации статистического анализа ссылок. Авторы системы CiteSeer предусмотрели лишь возможность выдачи для данной ссылки только контекста ее вхождения в цитирующую публикацию. Этот контекст должен анализировать пользователь с тем, чтобы определить характер ссылки – является ли она позитивной, негативной и т. п. В нашем случае поддержка атрибута вида ссылки обеспечивает статистическую обработку ссылок по указанным их свойствам.

Литература

[1] ACM Digital Library.

<http://portal.acm.org/dl.cfm?coll=portal&dl=ACM&CFID=65459678&CFTOKEN=64972349>

[2] CiteSeer.IST Scientific Literature Digital Library. <http://citeseer.ist.psu.edu/>

- [3] Википедия. <http://www.wikipedia.org>
- [4] Web of Science.
http://thomsonreuters.com/products_services/science/science_products/a-z/web_of_science
- [5] Journal Citation Reports.
http://thomsonreuters.com/products_services/science/science_products/a-z/journal_citation_reports
- [6] Российский индекс научного цитирования. <http://www.ipu.ru/period/pu/rinz.htm>
- [7] Паринов С.И. e-Science - онлайнное будущее науки //Информационные технологии. - 2007. - № 9. Приложение.
- [8] Паринов С.И. Новый подход к оценке результатов научно-исследовательской деятельности. Соционет: электронный депонент, 2008.
<http://socionet.ru/publication.xml?h=repec:rus:mqijxk:20>
- [9] Когаловский М.Р., Паринов С.И. Метрики онлайнных информационных пространств //Экономика и математические методы, 2008, т. 44, №2, с. 108-120.
- [10] Когаловский М.Р., Паринов С.И. Информационные ресурсы, наукометрические показатели и показатели качества метаданных системы Соционет. Труды девятой Всероссийской научной конференции «Электронные библиотеки: перспективные методы и технологии, электронные коллекции, Переславль, 2007». – Переславль-Залесский: Изд. «Университет города Переславля», 2007.
- [11] Паринов С.И., Ляпунов В.М., Пузырев Р.Л. Система Соционет как платформа для разработки научных информационных ресурсов и онлайнных сервисов //Электронные библиотеки. – 2003. - Том 6. - Выпуск 1.
- [12] CERIF 2008 1.0 - Full Data Model, <http://www.eurocris.org/cerif/cerif-releases/cerif-2008/>
- [13] Паринов С.И., Когаловский М.Р. Технология поддержки электронных научных публикаций как «живых» документов. Труды XI Всероссийской научной конференции «Электронные библиотеки: перспективные методы и технологии, электронные коллекции», Петрозаводск, 17-21 сентября 2009 г. – Петрозаводск: Карельский научный центр РАН, 2009.
- [14] Паринов С.И., Когаловский М.Р. «Живые» документы в электронных библиотеках //Прикладная информатика. – 2009. - № 6 (в печати).