Онтологическое аннотирование библиографических ссылок в научных публикациях и его использование в наукометрии

М.Р. Когаловский
Институт проблем рынка РАН
Москва, Россия

Информационные ресурсы России. 2013. № 5. С. 5-13.

Ключевые слова: библиографическая ссылка, научная публикация, наукометрия, онтология ссылок, онтологическая аннотация

Аннотация

Библиографические ссылки на цитируемые источники в научных публикациях не несут какой-либо информации о мотивах их цитирования. В таком виде они используются в сложившихся технологиях индексов цитирования. В статье предлагается подход, предусматривающий семантическое аннотирование вхождений традиционных «немых» ссылок в текстах публикаций на основе специально разработанных онтологий. При этом учитывается, что на один и тот же источник в публикации может быть несколько ссылок, мотивированных различным образом и обладающих в связи с этим различной семантикой. Поэтому каждая ссылка на один и тот же источник должна аннотироваться независимо. Коллекции текстов публикаций, обогащенных семантическими аннотациями ссылок, являются существенно более информативными источниками для оценки индексов цитирования и для других наукометрических исследований по сравнению с традиционно используемыми.

1. Введение

Сложившиеся в научном сообществе этические нормы обуславливают необходимость ссылок в публикуемых работах на цитируемые в них источники. Библиографические ссылки представляют выраженные явным образом ориентированные бинарные связи между цитирующими и цитируемыми публикациями (*связи цитирования*), и их коллекции являются *источником информации* для наукометрических измерений. Такие измерения осуществляются *индексами цитирования,* а также средствами различных научных электронных библиотек. Заметим, что термин *индекс цитирования* имеет два значения. С одной стороны, это библиографическая информационная система, предназначенная для наукометрических измерений, а с другой – одна из разновидностей статистических оценок цитирования научных публикации (традиционное простое количество цитирований, индекс Хирша и др.).

Актуальность измерений различных видов индексов цитирования для ученых – авторов публикаций в последние годы существенно возросла. В настоящее время индексы цитирования стали рассматриваться в качестве ключевого критерия оценки вклада ученого в развитие науки. На основе статистики цитирования измеряются и импакт-факторы периодических изданий, которые рассматриваются как характеристики их научного авторитета, хотя в научном сообществе начинает формироваться критическое отношение к такому подходу (см., например, *The San Francisco Declaration on Research Assessment* [26]).

Для измерений индексов цитирования создан ряд широко признанных международных систем индексов цитирования - Web of Science, SCOPUS, Web of Knowledge и др. Каждая из них базируется на весьма представительном корпусе журнальных публикаций и поддерживает его в актуальном состоянии. К сожалению, в них

совсем не представлены или слабо представлены публикации на русском языке. В нашей стране на основе электронной научной библиотеки eLibrary, созданной при поддержке Российского фонда фундаментальных исследований, создан и развивается отечественный индекс цитирования РИНЦ. Индексы цитирования оцениваются также и в созданной в последние годы системе Google Scholar на основе публикаций, доступных в Веб.

Хотя измеряемые сложившимися в настоящее время методами индексы цитирования признаются как количественные критерии значимости научных публикаций, они обладают, однако, существенным недостатком. Действительно, могут возникать парадоксальные ситуации, когда высоким индексом цитирования (а значит, в соответствии со сложившейся практикой, и высокой степенью научной значимости) обладает статья, содержащая грубые ошибки и/или принципиальные заблуждения, касающиеся злободневной проблемы, и в связи с этим вызывающая активный отклик научного сообщества.

Эта проблема возникает в связи с тем, что сложившиеся методы наукометрии опираются на библиографические ссылки, которые сами по себе не несут какой-либо информации, характеризующей цели цитирования, мотивы, побудившие автора цитирующей работы использовать такие источники, его мнение о них и т.п. Поэтому такие ссылки правомерно называть «немыми» [5]. Они характеризуют лишь факт существования связи цитирования между рассматриваемыми публикациями.

Следует заметить, что принято называть используемые источники *цитируемыми*. Тем не менее, совсем не обязательно в использующей их публикации действительно содержатся цитаты, буквально воспроизводящие или пересказывающие фрагменты текста используемого источника. В тексте «цитирующей» публикации может, например, упоминаться используемый источник в контексте, характеризующем семантику ссылки на него - оценивающем его в целом или конкретный высказанный в нем тезис, указывающий на использование рассмотренных в нем результатов в данной публикации и т.д.

Если учитывать семантику библиографических ссылок в наукометрических измерениях, можно избежать указанных выше ситуаций, формируя более дифференцированную статистику, характеризующую цели и мотивы «цитирования» тех или иных публикаций. Такая статистика будет характеризовать реальную степень их значимости и их роль в развитии научных знаний.

Семантика ссылки часто характеризуется автором цитирующей публикации в ее контексте и, таким образом, представлена в текстовом виде. Использование методов смыслового анализа текста для выявления семантики ссылок связано с известными проблемами обработки текстов на естественных языках и не позволяет получить приемлемые результаты. Вероятно, для этой цели можно пытаться использовать лишь результаты активно развиваемых в последние годы исследований, посвященных анализу эмоциональной окраски и тональности текстов [6, 13, 14, 18]. В зарубежной литературе это направление исследований получило название *Sentiment analysis* (или *Opinion mining*). Задача анализа заключается в определении мнения автора анализируемого текста относительно обсуждаемой сущности на основе высказанных им эмоциональных характеристик. Применяя методы *Sentiment analysis* к контексту (окрестности) ссылки, можно выявить, позитивно либо негативно относится автор к цитируемой работе. Однако, к сожалению, ценность достигнутых в этой области результатов в приложении к решению рассматриваемой проблемы весьма ограничена. В лучшем случае могут выявляться только ссылки оценочного характера, отражающие различные степени позитивного или негативного отношения автора анализируемой публикации к цитируемой работе. К тому же, нельзя гарантировать абсолютную точность получаемых результатов.

На наш взгляд, более продуктивным для практического применения в настоящее время является иной подход, на который мы и ориентируемся в этой работе. При его использовании автор работы или эксперт осуществляет семантическое аннотирование структурированными данными ссылок на используемые источники в тексте публикации

на основе контекстной информации и некоторой *онтологии*, которая определяет представляющие интерес классы отношений (связей) цитирования между научными публикациями. Так как для семантического аннотирования ссылок при этом используется онтология, его можно рассматривать как *онтологическое*.

Поскольку при таком подходе аннотация ссылок представляется с помощью структурированных данных, семантика ссылок будет однозначно интерпретироваться программными средствами. При компьютерной обработке текста публикации с аннотированными ссылками из него будут извлекаться данные о связываемых ссылкой публикациях, а также о семантике связи. Эти данные могут подходящим образом коллекционироваться наряду с аналогичными данными о других публикациях, полученными иным образом, и служить источником для наукометрических исследований. Мы будем далее называть ссылки (связи), снабженные онтологическими аннотациями, *семантическими ссылками* (*связями*).

К сожалению, авторы некоторых публикаций включают какие-либо работы в список использованных источников, но не упоминают их в своем тексте. Такие ссылки не учитываются в рассматриваемом подходе. В подобных случаях возможностью определения семантики таких ссылок располагает только автор содержащих их публикаций. Напротив, в случаях, когда цитируемая работа упоминается в тексте публикации, семантику ссылки может установить эксперт на основе анализа контекста, в котором эта работа упоминается.

В работах [3-5, 9, 19, 20] предложены подход и технология для спецификации и использования семантики широкого многообразия связей, отражающих отношения различного рода, существующие между информационными объектами научных электронных библиотек и других репозиториев цифровых информационных ресурсов. При этом охватываются и связи цитирования. Подход, обсуждаемый в данной работе, позволяет исчерпывающим образом выявлять все множество связей цитирования, имеющихся в обрабатываемой публикации благодаря анализу полного ее текста и явным образом описывать их семантику с помощью онтологического аннотирования. Коллекции таких публикаций могут использоваться как новый источник данных для систем, обладающих средствами наукометрических исследований.

В остальной части данной работы более подробно обсуждается предлагаемый подход к онтологическому аннотированию связей цитирования и его реализации. Приводится краткая характеристика существа семантического и, в частности, онтологического аннотирования. Обсуждаются принятые в издательской деятельности способы представления ссылок на используемые источники в научных публикациях. Далее уточняются объекты аннотирования и рассматриваются средства спецификации их семантики, обсуждаются способы выполнения и примеры использования онтологического аннотирования. В заключении подводятся итоги обсуждения рассматриваемого в работе подхода.

2. Что такое семантическое аннотирование?

Текстовые документы и другие информационные объекты, которыми оперируют электронные библиотеки и другие информационные системы, должны снабжаться явным образом представленными описаниями их свойств. Такие описания являются специальным видом данных и называются *метаданными* [2]. В последнее десятилетие активизировались стимулированные деятельностью консорциума W3C в области Семантического Веба разработки систем, имеющих дело со смысловым содержанием (семантикой) информационных объектов. Эти системы обеспечивают семантический поиск текстовых документов, смысловой анализ текстов, поддержку семантических связей между текстовыми документами, анализ и обработку семантической структуры контента системы.

Для описания семантики информационных объектов в таких системах используется особый вид метаданных, которые называются *семантическими* и могут представлять собой неструктурированные или структурированные данные. Примерами неструктурированных семантических метаданных являются аннотации или названия текстовых публикаций. Неструктурированные метаданные предназначены, главным образом, для использования человеком. Для компьютерной обработки, как правило, используются структурированные семантические метаданные благодаря возможности однозначной интерпретации их содержания при компьютерной обработке.

Процесс создания семантических метаданных информационного ресурса называется *семантическим аннотированием,* а создаваемые при этом семантические метаданные называются *семантической аннотацией* этого ресурса [17]. Семантическое аннотирование, осуществляемое на основе использования онтологии, называется *онтологическим аннотированием.*

Для семантического аннотирования могут использоваться различные средства – от естественных до формальных искусственных языков. Существуют многочисленные простейшие разновидности структурированных семантических метаданных, Примерами могут служить коды рубрик российского рубрикатора научно-технической информации ГРНТИ, классов отечественной классификационной системы ББК, международных классификаторов УДК и JEL (Journal of Economic Literature Classification System) [15], классификатора международной научно-технической организации ACM (Association for Computing Machinery) [25], а также формируемые в процессе фолксономии теги статей в Википедии и других социальных сетях.

Иным более многоаспектным и широко принятым в настоящее время средством описания семантики информационных объектов является набор элементов метаданных Дублинского ядра (Dublin Core, DC). Для представления семантических метаданных информационных источников и составляющих их информационных объектов могут использоваться онтологии и таксономии. В настоящее время для описания онтологий чаще всего применяются разработанные консорциумом W3C стандарты языков Семантического Веба - RDF, RDFS, OWL, OWL2 и его профили.

Процесс семантического аннотирования текстовых информационных ресурсов может быть полностью или частично автоматизирован, в частности, с помощью технологий извлечения информации из текста (см. например [27]), или должен выполняться экспертом. Сложности полной автоматизации этого процесса связаны с трудностями решения проблемы однозначной интерпретации смыслового содержания текстовой информации.

Семантическое аннотирование может осуществляться на различных уровнях *гранулирования* информационного ресурса: аннотация может относиться, например, к полной публикации, к ее разделу, к внутритекстовой библиографической ссылке, к странице Веба или отдельному ее фрагменту.

Возможны два способа ассоциирования семантической аннотации с аннотируемым информационным объектом или его фрагментом. Семантическая аннотация может быть *встроенной* в описываемый ресурс (*внутренняя аннотация*) или *автономной* по отношению к нему (*внешняя аннотация*). Например, для создания внутренней аннотации фрагментов веб-страниц необходима специальная их разметка, дополняющая стандартную HTML-разметку. Так, дополнительные теги разметки веб-страниц, называемые *микроформатами* [16], позволяют выделять в ней сведения о персонах, организациях, новостях, событиях, почтовых адресах и т.п. Некоторые из них поддерживаются популярными поисковыми машинами Веба и используются при обработке пользовательских поисковых запросов. Яндекс, например, поддерживает микроформаты: hCard - для разметки контактной информации (адресов, телефонов и т. д.); hRecipe - для описания кулинарных рецептов; hReview - для разметки рецензий, отзывов; hProduct – для разметки описаний товаров.

В предлагаемом в данной работе подходе для онтологического аннотирования семантических связей цитирования предполагается использовать встроенные аннотации, представляемые структурированными данными.

Рассмотрим теперь, какие разновидности библиографических ссылок встречаются в публикациях.

3. Разновидности библиографических ссылок

Как известно, *библиографической ссылкой* называют указание внутри текста издания, адресующее читателя к другому изданию [7]. Сложившаяся практика издательского дела предусматривает несколько разновидностей библиографических ссылок на цитируемые источники и возможных способов их представления в текстах публикаций. В нашей стране эти разновидности регламентируются системой стандартов по информации, библиотечному и издательскому делу. В частности, стандарт [1] определяет следующие виды библиографических ссылок: *внутритекстовые*, *подстрочные* и *затекстовые*. Последние содержатся либо в списке используемых источников, либо аналогичны по формату подстрочным, но приводятся в выноске [7], т.е. вынесены за основной текст.

Внутритекстовые библиографические ссылки содержатся непосредственно в тексте публикации и обычно заключаются в круглые скобки. Обычно такие ссылки используются в случаях, когда цитируемая публикация лишь однократно упоминается в тексте цитирующей публикации. Все остальные ссылки выносятся из основного текста и связываются с местом в нем, где цитируемые публикации должны упоминаться, с помощью вхождений их *идентификаторов*, уникальных в данной публикации.

Для связи *подстрочных* библиографических ссылок с текстом документа идентификатором ссылки служит *знак сноски*, в качестве которого используют порядковый номер ссылки или какую-либо литеру алфавита (например, звездочку), помещаемую в нужном месте текста, а также перед ссылкой, в надстрочной позиции шрифта. Таким же образом с помощью *знака выноски* связываются с текстом *затекстовые* библиографические ссылки, приводимые в выноске. Знаки выноски представляются точно так же, как и знаки сноски. Знаки сноски и знаки выноски могут группироваться. В группе они разделяются запятыми и обозначают в тексте группы соответствующих им подстрочных или затекстовых ссылок цитирования. Эти два вида ссылок одновременно не используются в одной и той же публикации.

Для затекстовых библиографических ссылок из списка использованных источников идентификаторами ссылок служат так называемые *знаки отсылки*, помещаемые, как и в предыдущих случаях, в соответствующем месте текста, а также и непосредственно перед библиографической ссылкой. Знак отсылки является для данной публикации уникальным идентификатором ссылки. Он представляет собой, в соответствии со стандартом [1], ее порядковый номер в списке использованных источников, сопровождаемый в случае, если цитируется фрагмент публикации, указанием ее раздела или составляющих его страниц. В зарубежных публикациях в качестве знака отсылки часто используется указание фамилий авторов или группа литер, образованная из этих фамилий, вместе с годом издания и с другими характеристиками, например, номерами томов, номерами страниц и т.п. Для обеспечения уникальности при необходимости к такой конструкции дополняются буквы латинского алфавита. Знаки отсылки – порядковые номера - заключаются в прямые скобки, а знаки отсылки остальных видов - в круглые скобки. Знаки отсылки могут группироваться, как и знаки сноски и выноски. Такие знаки – порядковые номера разделяются в группе запятыми и каждый отдельно либо вся группа заключаются в прямые скобки. Знаки отсылки другого вида разделяются в группе точкой с запятой, и каждый отдельно либо вся их группа заключаются в круглые скобки. Группа знаков отсылки обозначает в тексте публикации соответствующую группу библиографических ссылок из списка использованных источников.

Таким образом, каждая библиографическая ссылка внутри текста публикации представляется либо непосредственно, либо вхождением в текст публикации ее идентификатора, уникального в рамках данной публикации. В качестве таких идентификаторов используются знаки сноски, выноски, а также отсылки. Как уже отмечалось, возможны, однако, исключения. Некоторые авторы, вопреки сложившимся нормам, включают библиографические ссылки на некоторые публикации в списки используемых источников, но не связывают их с основным текстом своей работы. Ссылки на такие публикации оказываются в результате не представленными внутри текста «цитирующей» публикации.

4. Объекты онтологического аннотирования

Существующие в настоящее время системы – индексы цитирования генерируют данные для наукометрических измерений на основе анализа затекстовых библиографических ссылок, содержащихся в списках используемых источников обрабатываемых публикаций. В результате формируется структура информационного пространства такая, что каждая представленная в ней публикация находится в бинарной *связи* (отношении) *цитирования* с каждой публикацией, упоминаемой в ее списке использованных источников. Таким образом, в сгенерированной и пополняемой при обработке новых публикаций структуре информационного пространства между двумя публикациями может существовать *единственная* связь. Эта связь, как уже отмечалось, не несет какой-либо информации о мотивах цитирования. Поэтому мы называем такие связи «немыми» [5]. Если «озвучить» эти немые связи, явным образом специфицируя их семантику, можно получать более тонкие, дифференцированные по мотивам цитирования наукометрические показатели. Кроме того, построенная таким образом семантическая структура может использоваться как полигон для более широких наукометрических исследований.

Важное обстоятельство, не учитываемое традиционной наукометрией, заключается в том, что один и тот же источник может многократно цитироваться в научной публикации. Поэтому между цитирующей и цитируемой публикациями может существовать *не одна, а множество связей*, вообще говоря, с различной семантикой.

Учитывая рассмотренное выше разнообразие способов представления ссылок цитирования, можно утверждать, что каждая внутритекстовая библиографическая ссылка, а также каждый из знаков сноски, выноски и отсылки, связывающих с текстом некоторую библиографическую ссылку, представляет свой экземпляр связи между цитирующей и цитируемой публикациями. Иначе говоря, связи между данной публикацией и используемыми в ней источниками представляются в ее тексте внутритекстовыми ссылками и вхождениями идентификаторов библиографических ссылок других видов.

Таким образом, в качестве объектов онтологического аннотирования следует использовать содержащиеся в тексте публикации внутритекстовые библиографические ссылки, а также вхождения идентификаторов других ссылок, но не сами эти ссылки. Каждая из таких конструкций представлена в тексте публикации в рамках некоторого контекста, который правомерно называть *контекстом связи*. Чаще всего этот контекст характеризует мотивы цитирования и может использоваться экспертом, осуществляющим аннотирование, для определения семантики связи.

Следует заметить, что в случае групповой ссылки все знаки сноски, выноски и отсылки, входящие в одну группу, представляются в тексте публикации в рамках одного и того же контекста. Следовательно, они одинаково мотивированы и обладают одной и той же семантикой.

5. Спецификация семантики связей цитирования

Как отмечалось, в предлагаемом подходе для спецификации семантики связей цитирования в процессе их семантического аннотирования предусматривается использование встроенного их аннотирование с использованием специально разработанных онтологий связей, отражающих научные отношения между публикациями и используемыми в них источниками. Точнее, предполагается использовать *таксономию связей цитирования,* определяемую иерархией классов связей принятой онтологии. При этом каждая внутритекстовая библиографическая ссылка и каждое вхождение идентификаторов других библиографических ссылок на используемые источники ассоциируется с некоторым классом таксономии, который и характеризует определяемую семантику данной ссылки.

В настоящее время создана серьезная основа для создания таксономий семантических связей. Специалистами в области биомедицины из Оксфордского и Болонского университетов разработан модульный онтологический комплекс SPAR (*the Semantic Publishing and Referencing Ontologies*) [22, 23]. Этот комплекс состоит из восьми независимых повторно используемых детализированных онтологий, позволяющих описывать семантику библиографических объектов и их отношений. Первые четыре из них (FaBiO, CiTO, BiRO and C4O) позволяют описывать библиографические объекты, библиографические записи и источники в списках литературы в публикациях, связи цитирования, контексты цитирования и их связи с релевантными разделами цитируемых публикаций. Остальные онтологии рассматриваемого комплекса (DoCO, PRO, PSO and PWO) служат для создания управляемых словарей компонентов документов, ролей публикаций, состояний публикаций и потоков работ в издательских процессах.

В Главном госпитале в Массачусетсе и Медицинской школе в Гарварде разработана онтология SWAN (*Semantic Web Applications in Neuromedicine*) [21]. Как и SPAR, она состоит из набора онтологий-модулей. Цель ее разработки – обеспечение в рамках Семантического Веба комфортной среды для создания и сохранения семантического контекста научных коммуникаций, обеспечения доступа к нему, его интеграции, а также обмена неструктурированной или слабоструктурированной цифровой научной информацией.

Свой вклад в создание онтологий связей внес также консорциум W3C, который разработал рекомендацию SKOS (*Simple Knowledge Organization System*) [24], предназначенную для поддержки использования систем организации знаний, таких как тезаурусы, схемы классификации, таксономии и рубрикаторы (*Subject Heading Systems*) в среде Семантического Веба.

Следует, наконец, упомянуть работы в рассматриваемой области международной организации euroCRIS, предложившей в рамках проекта CERIF комплекс спецификаций [11, 12], которые также могут быть использованы для обсуждаемых здесь целей.

Итак, таксономию семантических связей цитирования можно сформировать на основе подходящей известной онтологии или группы онтологий как совокупность классов связей, образующих некоторую иерархию. Построенную таксономию можно представить для удобства использования в виде набора семантических управляемых словарей. Каждый словарь соответствует какому-либо классу связей, например, *оценочные связи* (подтверждение полученных в цитируемой работе результатов, опровержение их, оценка их как достижения мирового уровня и т.п.), *связи использования* научных результатов (использование метода или результатов эксперимента, изложенных в цитируемой работе, обобщение изложенного результата и т.п.). Отдельные значения в контролируемом словаре рассматриваются как имена подклассов класса таксономии, соответствующего данному словарю.

Одна из версий онтологий семантических связей цитирования и связей, представляющих ряд других классов отношений между информационными объектами контента разработана, поддерживается и используется в системе Соционет [4, 5, 9]. В системе имеются средства для формирования на основе этой онтологии многослойной

семантической структуры семантических связей между информационными объектами ее контента, в том числе и связей цитирования. Каждый ее слой соответствует одному их классов связей, определяемых онтологией. Семантическая структура контента системы является источником данных для ряда сервисов, позволяющих обрабатывать пользовательские запросы. К их числу относятся и запросы наукометрической статистики.

6. Процесс аннотирования ссылок цитирования

Онтологическое аннотирование ссылок цитирования в тексте публикации может осуществляться «вручную» ее автором либо экспертом в процессе просмотра текста публикации на компьютере. При этом может использоваться вспомогательная программа, которая сканирует текст публикации, обнаруживает внутритекстовые библиографические ссылки или вхождения идентификаторов других ссылок, которые подлежат аннотированию. При этом на экране компьютера показывается контекст обнаруженной внутритекстовой ссылки или идентификатора ссылки, на основе анализа которого лицо, осуществляющее аннотирование, может выбрать нужный класс ссылки в контролируемых словарях, представляющих таксономию ссылок цитирования, и далее инициировать генерацию аннотации ссылки и «имплантацию» ее в текст публикации. Идентификаторы одиночных и групповых ссылок аннотируются одинаковым образом.

Использование таксономий для аннотирования связей цитирования в тексте публикации позволяет представить такие аннотации в виде простейших структурированных данных, описывающих классы таксономии, к которым относятся соответствующие ссылки цитирования. Такие аннотации легко обнаруживаются в тексте, и они однозначно интерпретируются как человеком, так и обрабатывающей такой текст программой.

7. Использование онтологически аннотированных ссылок

Итак, библиографические ссылки цитирования, вообще говоря, имеют многократные вхождения в текст публикации либо непосредственно (внутритекстовые ссылки), либо опосредованным образом через вхождения их идентификаторов. Как уже отмечалось, могут использоваться вхождения групповых ссылок, имеющих одинаковую семантику и тем самым единую аннотацию.

Каким образом можно использовать текст публикации с аннотированными ссылками цитирования? Прежде всего такой текст может быть подвергнут автоматизированной обработке. При первом его просмотре из него извлекается массив ссылок цитирования всех видов с их идентификаторами (для ссылок, отличных от внутритекстовых, которые сами себя представляют).

Далее производится повторный просмотр текста публикации, в процессе которого обнаруживаются аннотированные ссылки или их идентификаторы. Если обнаружен идентификатор групповой ссылки, то она разгруппировывается.

Для каждой обнаруженной внутритекстовой ссылки цитирования или для обнаруженного идентификатора ссылки генерируется описание экземпляра связи цитирования, которое включает библиографические описания данной и цитируемой публикации, а также класс связи, указанный в аннотации вхождения ссылки. Библиографическое описание ссылки, представленной в тексте ее идентификатором обнаруживается в сгенерированном ранее массиве ссылок по ее идентификатору.

Сформированный набор описаний связей цитирования может содержать дубликаты связей одного и того же класса с одним и тем же цитируемым источником, которые должны быть удалены. Полученный набор описаний связей цитирования может быть импортирован в системы индексов цитирования, в научные электронные библиотеки или в иные цифровые репозитории информационных ресурсов, обладающие сервисами для наукометрических исследований и способные обрабатывать информацию такого рода.

Примером таких систем может служить Соционет [8, 10] – крупное научно-образовательное информационное пространство, интегрирующее коллекции и открытые архивы публикаций сотрудников нескольких десятков академических институтов, вузов и других организаций в различных областях знаний. Система динамично развивается как по объему поддерживаемых информационных ресурсов, так и по функциональным возможностям. Эта система является вместе с тем исследовательским полигоном для разработки новых подходов и функциональных механизмов, которые могут быть использованы в развитых научных электронных библиотеках.

В Соционет поддерживаются семантические связи между представленными в ней информационными объектами, в частности, и связи цитирования. Эксперты-пользователи системы имеют возможность в онлайновом режиме учреждать семантические связи с использованием контролируемых словарей, представляющих развитую таксономию связей. Эта таксономия разработана как некоторое расширение таксономии, синтезированной из компонентов рассмотренных выше онтологий. Имеются и разрабатываются новые сервисы для создания, поддержки, развития и использования комплекса управляемых словарей семантических связей, а также для генерации ряда новых более дифференцированных по сравнению с традиционными наукометрических показателей.

Явное представление в системе семантики связей между информационными объектами обеспечивает для пользователя визуализацию многослойной семантической структуры ее контента, каждый слой которой соответствует некоторому классу связей. Предусмотрены возможности визуальной семантической навигации по слоям этой структуры.

Семантические связи создаются в системе как информационные объекты, описываемые собственными метаданными. Благодаря этому их коллекции могут поддерживаться в виде самостоятельного репозитория, интегрироваться с другими аналогичными репозиториями, образуя глобальные полигоны для наукометрических исследований. Использование репозитория семантических связей позволяет не только генерировать различные виды семантических окрашенных индексов цитирования, но и исследовать топологию структуры семантических связей с целью получения различных содержательных характеристик охваченного им корпуса научных знаний, важных для поддержки развития научных исследований.

Проиллюстрируем возможности использования семантических связей в системе Соционет несколькими примерами применительно к связям цитирования. Прежде всего, могут обрабатываться статистические запросы пользователей. Характер генерируемой по запросам статистической информации может быть весьма разнообразным. Например, в применении к связям цитирования можно запросить для конкретной публикации количество входящих или исходящих из ее связей (т.е. связей, в которых данная публикация участвует как целевая или, соответственно, как исходная), относящихся к некоторому классу связей цитирования, включенных в заданную таксономию связей. В частности, получаемая статистика может показывать сколько имеется позитивных или негативных ссылок на данную работу, в каком количестве работ используются методы, предложенные в данной статье, или содержащиеся в ней научные данные и т.д.

Наряду с обработкой статистических запросов предусматривается возможность получения перечня конкретных публикаций, представленных в системе и связанных с заданной публикацией как с исходной или с целевой в связях цитирования заданных классов. Например, можно выяснить, на результаты каких публикаций опирается данная конкретная работа или, наоборот, в каких публикациях получены результаты, основанные на данной работе.

Укажем еще одну важную группу возможностей - операции над полным графом связей цитирования. Исследуя этот граф, можно решать различные задачи, связанные как с анализом топологии графа и вычленением подграфов с дугами разных классов

таксономии связей цитирования, так и с визуализацией подграфов. Например, можно вычленить и визуализировать из многослойной структуры полного графа слой, соответствующий классу связей, указывающему на использование одних публикаций как основополагающих для других публикаций. Этот слой будет в общем случае представлять множество графов, каждый из которых соответствует одной из публикаций, на результатах которой базируются какие-либо иные публикации. Можно также запросить подграф, образованный связями цитирования, характеризующими цитирующие публикации как развивающие те результаты, которые обсуждаются в цитируемой публикации и принадлежат ее автору. Полученный подграф будет характеризовать логику развития данной области науки, конечно, если в системе достаточно основательно представлены публикации, относящиеся к рассматриваемой области.

Более подробно указанные возможности системы Соционет обсуждаются в работах [3, 4, 5, 9].

Заключение

Подводя итоги обсуждения предлагаемого в данной работе подхода, следует отметить, что онтологическое аннотирование связей цитирования в текстах научных публикаций и использование явным образом определенной с их помощью семантики связей создает качественно новые возможности для наукометрических исследований, позволяет получать новые результаты, полезные для научного сообщества.

Для обеспечения интероперабельности систем, использующих и обрабатывающих семантическую информацию о связях цитирования представляется целесообразным разработать и использовать в деятельности научного сообщества стандарты таксономий библиографических ссылок и форматов представления их онтологических аннотаций.

Практическое использование предлагаемого в данной статье подхода, конечно же, связано со значительными трудностями. Прежде всего, наукометрические исследования семантики связей цитирования имеют смысл лишь на представительном репозитории цифровых публикаций. Однако онтологическое аннотирование библиографических ссылок в большом массиве публикаций крупного репозитория связано с существенными затратами, хотя эта деятельность может осуществляться в научном сообществе в распределенном режиме с соответствующим распределением связанных с нею затрат, в частности, силами не только экспертов, но и самих авторов публикаций, представляемых в подобный семантизированный репозиторий. Но главная трудность заключается в преодолении сложившейся традиции наукометрических измерений, которой следуют организации - создатели и владельцы авторитетных в научном сообществе индексов цитирования и пользователи их продукции во всем мире. Тем не менее, автор надеется, что идеи данной статьи достойны обсуждения и со временем их практическое использование приобретет право на жизнь.

Благодарности

Работа частично поддержана грантом РФФИ, проект 12-07-00518-а.

Литература

1. ГОСТ 7.0.5-2008 Система стандартов по информации, библиотечному и издательскому делу. Библиографическая ссылка. Общие требования и правила составления. Издание официальное. - М.: Стандартинформ, 2009.
2. Когаловский М.Р. Метаданные в компьютерных системах //Программирование. МАИК/Наука «Интерпериодика». - 2013. - № 4.
3. Когаловский М.Р., Паринов С.И. Новый источник данных для наукометрических исследований. Представлена на XV Всероссийскую научную конференцию

«Электронные библиотеки: перспективные методы и технологии, электронные коллекции - RCDL-2013». Ярославль, 14-17 октября 2013 г.
4. Когаловский М.Р., Паринов С.И. Классификация и использование семантических связей между информационными объектами в научных электронных библиотеках //Информатика и ее применения. - 2012. - Т. 6. Вып. 3. - С. 32-42.
5. Когаловский М.Р., Паринов С.И. Семантическое структурирование контента научных электронных библиотек на основе онтологий. В кн.: "Современные технологии интеграции информационных ресурсов: сборник научных трудов». - Санкт-Петербург: Президентская библиотека им. Б.Н. Ельцина, 2011. - С. 26-45.
6. Клековкина М.В., Котельников Е.В. Метод автоматической классификации текстов по тональности, основанный на словаре эмоциональной лексики. Труды XIV Всероссийской научной конференции «Электронные библиотеки: перспективные методы и технологии, электронные коллекции - RCDL-2012», Переславль-Залесский, Россия, 15-18 октября 2012 г. - Переславль-Залесский, Институт программных систем РАН, 2012. – С. 118-123.
7. Мильчин А.Э. Издательский словарь-справочник: [электронное издание]. - 3-е изд., испр. и доп. - М.: ОЛМА-Пресс, 2006. http://slovari.yandex.ru/~книги/Издательский словарь [Дата обращения 25 июня 2013 г.]
8. Онлайновая научная инфраструктура Соционет. http://socionet.ru/ [Дата обращения 25 июня 2013 г.]
9. Паринов С.И., Когаловский М.Р. Технология семантического структурирования контента научных электронных библиотек. Труды XIII Всероссийской научной конференции «Электронные библиотеки: перспективные методы и технологии, электронные коллекции – RCDL-2011». Воронеж, 19-22 октября 2011 г. – г. Воронеж: Воронежский государственный университет, 2011. – С. 94-103.
10. Паринов С.И., Ляпунов В.М., Пузырев Р.Л. Система Соционет как платформа для разработки научных информационных ресурсов и онлайновых сервисов //Российский научный электронный журнал «Электронные библиотеки». – 2003. – Том 6. – Вып. 1. http://www.elbib.ru/index.phtml?page=elbib/rus/journal/2003/part1/PLP [Дата обращения 25 июня 2013 г.]
11. CERIF 1.3 Full Data Model (FDM): Introduction and Specification. euroCRIS, 2012. http://www.eurocris.org/Uploads/Web%20pages/CERIF-1.3/Specifications/CERIF1.3_FDM.pdf [Дата обращения 25 июня 2013 г.]
12. CERIF 1.3 Semantics: Research Vocabulary. CERIF Task Group, euroCRIS, 2012. http://www. eurocris.org/Uploads/Web%20pages/CERIF-1.3/Specifications/CERIF1.3_Semantics.pdf [Дата обращения 25 июня 2013 г.]
13. Feldman R. Techniques and Applications for Sentiment Analysis. Communications of the ACM, April 2013, vol. 56, no. 4, pp. 82-89.
14. Galassini C., Malizia A., and Bellucci A. An approach for developing intelligent systems for sentiment analysis over social networks. Intelligent Systems and Control /742: Computational Bioscience, J.F. Whidborne, P. Willis, G. Montana, Eds. Cambridge, United Kingdom, July 11 – 13, 2011.
15. Journal of Economic Literature (JEL) Classification System. http://www.aeaweb.org/jel/jel_class_system.php [Дата обращения 25 июня 2013 г.]
16. Microformats. http://microformats.org/ [Дата обращения 25 июня 2013 г.]
17. Oren E., Moller K.H., Scerri S., Handschuh S., and Simtek M. What are Semantic Annotations? http://www.siegfried-handschuh.net/pub/2006/whatissemannot2006.pdf [Дата обращения: 25 июня 2013 г.]
18. Pang B., Lee L. Opinion Mining and Sentiment Analysis //Foundations and Trends in Information Retrieval. - 2008. - Volume 2, Issue 1-2. January 2008, pp. 1-135. http://dl.acm.org/citation.cfm?id=1454712 [Дата обращения 25 июня 2013 г.]

19. Parinov S. Open Repository of Semantic Linkages. In: Proc. of 11th Intern. Conference on Current Research Information Systems e-Infrastructure for Research and Innovations (CRIS 2012), Prague, 2012. http://socionet.ru/publication.xml?h=repec:rus:mqijxk:29 [Дата обращения 25 июня 2013 г.]
20. Parinov S.I., Kogalovsky M.R. Semantic Linkages in Research Information Systems as a New Data Source for Scientometric Studies. CERN Workshop on Innovations in Scholarly Communication (OAI8). Wednesday 19 June 2013 - Friday 21 June 2013. University of Geneva. Book of Abstracts. http://indico.cern.ch/conferenceDisplay.py /abstractBook?confId=211600 [Дата обращения 27 июня 2013 г.]
21. Semantic Web Applications in Neuromedicine (SWAN) Ontology. W3C Interest Group Note, 20 October 2009. http://www.w3.org/TR/2009/NOTE-hcls-swan-20091020/ [Дата обращения 25 июня 2013 г.]
22. Shotton D. Introduction the Semantic Publishing and Referencing (SPAR) Ontologies. October 14, 2010. http://opencitations.wordpress.com/2010/10/14/introducing-the-semantic-publishing-and-referencing-spar-ontologies/ [Дата обращения 25 июня 2013 г.]
23. Shotton D. and Peroni S. Semantic annotation of publication entities using the SPAR (Semantic Publishing and Referencing) Ontologies /Beyond the PDF Workshop, La Jolla, 19 January 2011. http://imageweb.zoo.ox.ac.uk/pub/2010/Publications/Shotton&Peroni_semantic_annotation_of_publication_entities.pdf [Дата обращения 25 июня 2013 г.]
24. SKOS Simple Knowledge Organization System Reference. W3C Recommendation, 18 Aug. 2009. http://www.w3.org/TR/skos-reference/ [Дата обращения 25 июня 2013 г.]
25. The 2012 ACM Computing Classification System. March 30, 2012. http://www.acm.org/about/class/2012 [Дата обращения 25 июня 2013 г.]
26. The San Francisco Declaration on Research Assessment (DORA). http://am.ascb.org/dora/ [Дата обращения 25 июня 2013 г.]
27. Uren V., Cimiano P., Iria J., Handschuh S., Vargas-Vera M., Motta E., Ciravegna F. Semantic annotation for knowledge management: Requirements and a survey of the state of the art Web Semantics: Science, Services and Agents on the World Wide Web. Volume 4, Issue 1, January 2006, pp. 14-28.

Ontological annotation of bibliographical references in scientific publications and its use in scientometrics

Mikhail R. Kogalovsky

Market Economy Institute of RAS, Moscow

Keywords: bibliographical reference, ontology of references, ontological annotation, scientific publication, scientometrics

Abstract

Bibliographic references to cited sources in scientific publications do not carry any information on motives of their citing. In such kind they are used in the existed technologies of citation indexes. In the article an approach providing semantic annotation of occurrences of traditional "mute" references in texts of publications is offered. The approach is based on specially developed ontology. It takes in account also, that in the given publication there can be several references to the same source. Reference occurrences to the same source can be motivated in different way and therefore they have different semantics. Because of that each occurrence of reference to the source is annotated independently. Collections of publications enriched by semantic references annotations are essentially more informative sources for an estimation of indexes of

citations and for others scientometrics researches in comparison with traditionally used ones.