Научный поиск: методы тематически-ориентированного поиска научной информации

© Н.В. Авдеева © О.В. Никулина
Российская государственная библиотека, Москва
avdeeva@rsl.ru

© А.С. Хританков © Ю.В. Чехович
ЗАО «Анти-Плагиат», Москва
chehovich@antiplagiat.ru

Аннотация

Статья посвящена проблеме автоматизированного поиска научной информации по заданной теме. В работе описаны возможности тематического поиска с помощью системы «Научный поиск», реализованной в Российской государственной библиотеке (РГБ), с указанием основных принципов работы и возможностей использования. В статье также предложены пути дальнейшего развития системы.

1 Что такое научный поиск?

В настоящее время нет четкого определения понятия «научный поиск». Большинство современных исследователей понимают его как поиск информации по нужной тематике или сфере исследования. При этом зачастую трудно четко ограничить тематику, которая интересует в каждом конкретном случае поиска. Кроме того, чтобы зафиксировать новизну информации, нужно знать, что именно уже известно в данной области на данном этапе. Таким образом, можно сделать вывод, что научный поиск необходим для отбора похожих по тематике научных работ и поиска исследователей, работающих в определенном направлении. При этом для успешного поиска информации исследователю необходимо не только разобраться в тематике исследования и ознакомиться с работами других исследователей, но и являться хорошим специалистом в изучаемой области.

Труды 16-й Всероссийской научной конференции «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» — RCDL-2014, Дубна, Россия, 13-16 октября 2014 г.

2 Методы научного поиска

В век информационных технологий отошел на второй план так называемый «ручной» поиск научной информации, который осуществлялся по каталожным карточкам и картотекам, и его место занял автоматизированный поиск.

Автоматизированные методы поиска научной информации позволяют, используя различные технические средства, осуществить подборку материала по интересующей исследователя научной тематике.

Методы поиска при этом в значительной степени определяются типом поисковых запросов и областью поиска. Как правило, для поиска научной информации используются запросы в виде набора ключевых слов, с поиском как по полным текстам документов, так и по метаданным документов – названиям, аннотациям и т.п. Также используются те или иные виды иерархического поиска по автоматизированным каталогам. Иногда применяются вопросно-ответные сервисы и системы, такие как, например, система Exactus [7].

В информационно-поисковых системах проводится лингвистический анализ запроса с учетом морфологии, поиск среди документов и сайтов, размещенных в свободном доступе в среде Интернет, а результаты предоставляются в ранжированном виде в зависимости от их релевантности. Но зачастую использование информационно-поисковых систем не дает нужного результата при поиске научной информации, т.к. они, в основном, анализируют содержание сайтов на предмет наличия/отсутствия ключевых слов и не производят тематического анализа найденной информации.

Более широкие возможности поиска научной информации предоставляют специализированные базы данных, материалы в которых ограничены одной общей тематикой или распределены по определенным четко обозначенным темам. К данному типу баз данных относятся электронно-

библиотечные системы, научные электронные библиотеки, в которых содержатся полные тексты документов, и электронные каталоги различных библиотек. Зачастую поиск в подобных базах данных можно осуществлять не только по ключевым словам, но и по различным тематическим запросам, при этом результаты поиска формируются с учетом анализа полного текста документа.

Отдельным, быстро набирающим популярность видом поиска является, так называемый «поиск по образцу». Суть поиска по образцу заключается в указании в качестве поискового запроса примера тех объектов, которые нужно найти. Затем поисковая система находит похожие по определенным критериям объекты, ранжирует их по степени похожести на указанный в запросе пример и возвращает в качестве результатов поиска.

Используемые поисковой системой критерии могут быть достаточно сложными и непригодными для «ручного» поиска, но в то же время позволяют эффективно находить похожие объекты.

Примером реализованного тематического поиска по образцу является система eTBLAST, реализующая возможности тематического поиска по базе данных медицинских статей MEDLINE [12].

3 Научный поиск в РГБ

Российская государственная библиотека реализовала пилотную версию программного инструмента «Научный поиск» на сайте http://ss.rsl.ru. «Научный поиск» открыт для использования всем посетителям сайта.

3.1 Информационная составляющая научного поиска

В качестве области поиска используется Электронная библиотека диссертаций Российской государственной библиотеки (ЭБД РГБ).

Российская государственная библиотека располагает уникальным фондом подлинников кандидатских и докторских диссертаций, защищенных в стране по всем специальностям, кроме медицины и фармации. Для решения основных проблем: сохранения такого огромного фонда, а главное – обеспечения доступа к нему одновременно большого количества читателей, – в начале XXI века РГБ было принято решение о создании ЭБД РГБ на основе современных информационных технологий. В 2003 году был оцифрован стартовый пакет диссертаций по наиболее востребованным специальностям: экономические, юридические, педагогические, психологические и философские науки (всего около 28 000 полных текстов). Начиная с 2004 года, состав ЭБД РГБ пополнялся объемом диссертаций по всем специальностям (кроме медицины и фармации), что составляет около 30 000 – включая 20 000 кандидатских и 10 000 докторских – диссертаций в год. В рамках проекта ретроконверсии в 2006 году были оцифрованы все диссертации за 1985 год. А с 2007 года в ЭБД РГБ поступают диссертации по всем дисциплинам, включая работы по медицине и фармации [1].

На сегодняшний день в ЭБД РГБ содержится 810 383 полных текста: 382 277 диссертаций и 428 106 авторефератов [2], по всем специальностям Высшей аттестационной комиссии при Министерстве образования и науки Российской Федерации (ВАК при Минобрнауки России): физико-математические науки, химические науки, биологические науки, искусствоведение, технические науки, сельскохозяйственные науки, исторические науки, экономические науки, философские науки, филологические науки, географические науки, юридические науки, педагогические науки, медицинские науки, фармацевтические науки, ветеринарные науки, геолого-минералогические науки, архитектура, психологические науки, военные науки, социологические науки, политические науки, культурология, а также науки о земле.

Каталог ЭБД РГБ находится в свободном доступе для всех пользователей сети Интернет. Полные тексты диссертаций и авторефератов представлены в формате PDF.

3.2 Программная составляющая научного поиска

Программная реализация системы «Научный поиск» на сайте http://ss.rsl.ru выполнена компанией ЗАО «Анти-Плагиат».

Предварительная подготовка для организации поиска в системе проводится в два этапа. Она включает построение тематической модели и индексирование документов коллекции.

Построение тематической модели начинается с предварительной обработки текстов. Выполняется удаление колонтитулов, номеров страниц, устраняются знаки переноса, исправляются проблемы с кодировкой текстов.

После этого производится лемматизация (приведение слов в различной грамматической форме к лемме – основной форме слова), с помощью библиотеки Apache Lucene, тексты документов переводятся в частотное представление по выделенным словам.

Частотное представление текстов используется для удаления стоп-слов по списку, фильтрации по частоте вхождения слов в документы коллекции, в каждый из документов. Формируется общий словарь терминов текстов коллекции как совокупность использованных в текстах слов после фильтрации [8]. Документам сопоставляются векторы терминов из словаря с указанием числа вхождений.

Построение тематической модели заключается в формировании определенного количества взвешенных наборов терминов, называемых темами. Моделируемая коллекция документов рассматривается как один из возможных результатов выполнения так называемой

вероятностной порождающей модели LDA [11] в модели мешка слов (bag-of-words). Для оценки параметров распределения вероятности этой порождающей модели и выявления тем используется алгоритм тематического моделирования CVB0 [10], реализованный в библиотеке Mahout 0.7 в инфраструктуре Hadoop (использована реализация по версии MapR 1.2.9). Алгоритм представляет собой итеративную процедуру, оптимизирующую отклонение приближенного распределения от неизвестного истинного распределения в порождающей модели. Каждая тема может не иметь четкого смыслового наполнения и служит формальным центром определенного подмножества документов.

Полученные распределения терминов по темам используются для расчета распределения тем по документам.

После построения тематической модели осуществляется индексирование документов коллекции и построение обратного индекса (inverse file index), которое позволяет для любого вектора терминов найти наиболее релевантные вектора терминов среди проиндексированных документов коллекции и соответствующие этим векторам документы. При сравнении векторов и определении релевантности используется косинусная мера близости (cosine similarity).

Построенная тематическая модель позволяет относительно быстро добавлять в индекс новые документы. Затратной процедурой с точки зрения вычислительных ресурсов является перестройка тематической модели, которую необходимо осуществлять лишь в случае серьезных пополнений или изменений в коллекции документов.

Сервис «Научный поиск» позволяет набрать/вставить в соответствующее поле текст (длиной не менее 1024 символов), загрузить документ в одном из распространенных текстовых форматов или указать ссылку на доступную страницу в сети Интернет. Затем документ обрабатывается на сервере с использованием методики «поиска по образцу» и сравнением выделенного системой набора тем в загруженном документе с тематической моделью коллекции, в результате чего формируется список наиболее релевантных запросу документов. Чем в большей степени наборы тематик похожи, тем выше в результатах поиска будут отображаться найденные документы коллекции. Для представления результатов поиска пользователям используется метаинформация и текстовые описания документов из ЭБД РГБ. Список работ отображается в виде библиографических записей со ссылкой на полный текст документа [9].

Благодаря данной технической реализации любой пользователь может найти перечень диссертаций и авторефератов из ЭБД РГБ, соответствующих тематике его исследования, и ознакомиться с полными текстами работ.

На сегодняшний день сервис научного поиска позволяет осуществлять поиск близких по тематике документов среди более 800 тысяч проиндексированных диссертаций и авторефератов [3,5]. Построенная тематическая модель использует 400 автоматически выделенных тем. В терминологическом словаре содержится около 280 тысяч терминов [6].

3.3 Практическое использование

Для получения и изучения результатов поиска пользователю необходимо лишь войти на сайт сервиса «Научный поиск» http://ss.rsl.ru, находящийся в свободном доступе в сети Интернет, загрузить текст документа одним из предложенных системой способов и запустить процесс научного поиска. Результаты поиска представляются в виде списка диссертаций и авторефератов из ЭБД РГБ, тематически наиболее схожих с загруженным текстом, с указанием релевантности по каждому из источников. Документы пронумерованы и расположены по мере убывания степени релевантности. Дополнительно пользователю предоставляется возможность просмотра полного текста каждого документа из перечня найденных документов или поиска работ, схожих с уже найденными.

Доступ к полному тексту выбранной диссертации или автореферата осуществляется через сайт ЭБД РГБ http://diss.rsl.ru, куда пользователь автоматически перенаправляется системой «Научный поиск» при нажатии на название выбранной работы. Для открытия полного текста работы пользователю на выбор предоставляются разные виды программного обеспечения: 2 системы, web-интерфейс для on-line просмотра и Acrobat Reader, для открытия произведений, находящихся в свободном доступе; и 2 системы, DefView (Defence Viewer) и DVS (Documents View System), которые позволяют открывать документы, находящиеся как в свободном, так и в ограниченном доступе, во исполнение 4 части Гражданского кодекса РФ, защищая при этом произведения от несанкционированного копирования. Система DVS является усовершенствованной разработкой и не требует установки дополнительного программного обеспечения, потому что реализована как web-приложение.

Интерфейс программного обеспечения позволяет пользователю производить ряд операций над документом: просматривать текст документа, осуществлять переход между его страницами, изменять масштаб, поворачивать страницы или использовать режим инверсии цветов. Очень важной является функция полнотекстового поиска, которая позволяет найти интересующие слова и/или словосочетания в тексте, выводит на экран перечень страниц документа, где встречаются искомые слова и/или словосочетания, а также подсвечивает

найденные фрагменты на выбранной странице документа.

Доступ к полным текстам диссертаций и авторефератов из ЭБД РГБ предоставляется, в основном, на территории Виртуальных читальных залов РГБ, т.е. на рабочих местах виртуальных читателей, оборудованных персональными компьютерами, установленными исключительно на территории библиотек различных организаций. Количество открываемых Виртуальных читальных залов с каждым годом только увеличивается, и на сегодняшний день уже создано 578 ВЧЗ РГБ почти во всех регионах России, в 9 странах СНГ и на территориях Республики Иран, Грузии, Монголии и Финляндии [2].

4 Возможности дальнейшего развития

С целью дальнейшего развития проекта планируется перестройка тематической модели с расширенным составом тем для улучшения качества поиска. Кроме того, планируется тематическое моделирование рубрикатора (с учетом классификации научных специальностей), что позволит пользователю задавать поисковый запрос и как набор рубрик.

Кроме этого планируется дополнить выдачу тематического поиска результатами проверки текста с помощью системы «Антиплагиат.РГБ», что позволит находить документы, релевантность которых запросу обусловлена использованием общих цитат [3, 4].

5 Выводы

Создан инструмент, обеспечивающий возможности тематического поиска среди полных текстов коллекции диссертаций и авторефератов ЭБД РГБ. Инструмент полезен при проведении исследований в областях, которые не всегда легко описать набором ключевых слов, но по которым есть примеры в виде научных текстов – статей, диссертаций, монографий или их фрагментов.

Литература

[1] Авдеева Н.В. Опыт создания и поддержки полнотекстовых баз данных неопубликованных документов // Труды XIV Всероссийской научной конференции «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» – RCDL’2012, Переславль-Залесский, Россия. – Переславль-Залесский: Изд-во «Университет города Переславля», 2012. – С. 294–299

[2] Авдеева Н.В. Электронные ресурсы Российской государственной библиотеки – основа успешного научного и образовательного процесса вуза // Высокие интеллектуальные технологии и инновации в национальных исследовательских университетах: материалы Международной научно-методической конференции, 5–7 июня 2014 г., Санкт-Петербург. – Т. 6. Пленарные доклады. – СПб.: Изд-во Политехн. ун-та, 2014. – С. 58–69. http://www.spbstu.ru/conference/2014/hit_2013_proceeding_plrep.pdf

[3] Авдеева Н.В., Ботов П.Ю., Букаев А.С., Вислый А.И., Груздев И.А., Житлухин Д.А., Романов М.Ю., Чехович Ю.В. Внедрение системы «Антиплагиат» в Российской государственной библиотеке // Материалы Международной конференции «Интеллектуализация обработки информации» (ИОИ-8) – октябрь, 2010. – С. 499–503.

[4] Авдеева Н.В., Ботов П.В., Букаев А.С., Вислый А.И., Груздев И.А., Ивахненко А.А., Никулина О.В., Чехович Ю.В. Система «Антиплагиат.РГБ»: задачи, проблемы, результаты, перспективы // Материалы Международной конференции «Интеллектуализация обработки информации» (ИОИ-9) – сентябрь, 2012. – С. 593–596.

[5] Авдеева Н.В., Ботов П.Ю., Букаев А.С., Вислый А.И., Груздев И.А., Ивахненко А.А., Чехович Ю.В. Система «Антиплагиат»: возможности и перспективы // Материалы международной научной конференции «Новые информационные технологии и менеджмент качества (NIT&QM’2011)» / редкол.: А.Н. Тихонов (пред.) и др., ФГУ ГНИИ ИТТ «Информика». – М.: ООО «Арт-Флэш», 2011. – С. 123–125.

[6] Авдеева Н.В. Научный поиск // Тезисы Восемнадцатой международной конференции «SCIENCE ONLINE: электронные информационные ресурсы для науки и образования». – С. 25. http://elibrary.ru/projects/conference/turkey2014/program.pdf

[7] Осипов Г.С., Тихомиров И.А., Смирнов И.В. Exactus – система интеллектуального метапоиска в сети Интернет. // Труды десятой национальной конференции по искусственному интеллекту с международным участием КИИ-2006. – М.: Физматлит, 2006. – Т. 3. – С. 859–866.

[8] Царьков С.В. Автоматическое выделение ключевых фраз для построения словаря терминов в тематических моделях коллекций текстовых документов // Естественные и технические науки. – № 6 (62). – 2012. – С. 456–464.

[9] Сайт сервиса «Научный поиск». http://ss.rsl.ru/

[10] A. Asuncion, M. Welling, P. Smyth, Y.W. Teh. On smoothing and inference for topic models / In Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence (UAI '09). AUAI Press, Arlington, Virginia, United States. – 2009. – P. 27–34.

[11] D.M. Blei, A.Y. Ng, M.I. Jordan. Latent dirichlet allocation / J. Mach. Learn. Res. 3 (March 2003). – 2003. – P. 993–1022.

[12] James Lewis, Stephan Ossowski, Justin Hicks, Mounir Errami, and Harold R. Garner. 2006. Text similarity: an alternative way to search MEDLINE. Bioinformatics 22, 18 (September 2006). – P. 2298–2304. DOI=10.1093/bioinformatics/btl388. http://dx.doi.org/10.1093/bioinformatics/btl388.

Scientific Search: Methods of Thematically Oriented Search of Scientific Information

N. Avdeeva, O. Nikulina, A. Khritankov, Y. Chekhovich

The article is devoted to the problem of automatic search of scientific information on the theme given. It describes the capabilities of thematic search with the help of the system “Scientific search” realized at the Russian State Library (RSL). The main principles of the system functioning and capabilities of its usage are indicated. The article also offers ways for further development of the system.