СОВЕРШЕНСТВОВАНИЕ МЕТОДОВ ПОИСКА НАУЧНОГО ЗНАНИЯ КАК ФАКТОР РАЗВИТИЯ СОВРЕМЕННОЙ НАУКИ

Иванов Владимир Константинович
канд. техн. наук, доцент кафедры «Информационные системы» Тверского государственного технического университета, г. Тверь
E-mail: mtivk@mail.ru
Борисов Сергей Юрьевич
аспирант кафедры «Информационные системы» Тверского государственного технического университета, г. Тверь
E-mail: delije-cz@yandex.ru

THE IMPROVING OF METHODS OF SCIENTIFIC KNOWLEDGE RETRIEVAL AS A FACTOR OF DEVELOPMENT OF MODERN SCIENCE

Vladimir Ivanov
ph.D., Associate Professor of the Department "Information Systems", Tver State Technical University, Tver
Sergei Borisov
postgraduate of the Department "Information Systems", Tver State Technical University, Tver

АННОТАЦИЯ

В настоящей статье представлены промежуточные результаты патентных исследований в рамках проекта РФФИ «Интеллектуальная распределенная система информационной поддержки инноваций в науке и образовании» (договор № НК13-07-00342\13, руководитель проекта — Иванов В.К.).

В начале статьи приводится подтверждение актуальности предмета исследования — совершенствования методов поиска и синтеза научного знания, далее делается обзор основных результатов, достигнутых, на данный момент, коллективом исследователей. В конце приводятся основные насущные задачи и перспективы дальнейшего развития исследований.

ABSTRACT

This article presents the interim results of patent research in the framework of RFBR "Intelligent Distributed Information Management System for Innovation in Science and Education" (contract number NK13-07-00342\13, project manager — Ivanov V.K.).

In the beginning of the article is a confirmation of the relevance of the research subject — improving the methods of search and synthesis of scientific knowledge, then reviews the main results achieved by a team of researchers for the moment. At the end there are main pressing challenges and prospects for the further development of research.

Ключевые слова: алгоритм; векторная модель; качество поиска; мера близости; метрика; поиск решений; ранжирование; релевантность; семантика документа; фильтрация.

Keywords: algorithm, vector model, the quality of the search, a measure of intimacy, metrics, search for solutions, ranking, relevancy, the semantics of the document; filtering.

Создание новых эффективных методов поиска и синтеза научного знания (в частности, прорывных технологий и инновационных идей) является, на сегодняшний день, одной из наиболее приоритетных задач исследований и разработок, способствующих развитию как отдельно взятого научного направления — информационного поиска, так и современной науки в целом. Разнообразные исследования касаются особенностей поиска инноваций в экономике, науке, образовании [1, с. 1—8; 3, с. 37—42].

Причиной этому является то, что, на сегодняшний день, в различных учреждениях науки и образования (научно-исследовательских институтах, высших учебных заведениях и др.) накоплен громадный объём информации, значительную часть из которой составляют электронные документы в виде текстов на естественном языке, находящиеся в документных базах данных и электронно-библиотечных системах. Особое место среди данной накопленной информации занимают инновации — научные новшества, являющиеся конечным результатом деятельности учёного или исследователя, обеспечивающие качественный рост показателей эффективности тех или иных процессов или улучшение свойств объектов. Важным условием существования

инновации является то, что данное новшество должно быть внедрено и зафиксировано на каком-либо носителе, как правило, таким носителем является патент на изобретение или научную разработку. Большие объёмы информации в документных базах данных (в том числе, базах патентов) привели к необходимости поиска новых эффективных методов создания и наполнения электронных коллекций новейших идей и технологий, содержащих не просто их описания, а специальным образом отобранные, классифицированные и ассоциированные данные.

На решение данной острой проблемы направлен проект «Интеллектуальная распределенная система информационной поддержки инноваций в науке и образовании». Суть данного проекта заключается в разработке программной системы с функциями семантического поиска и интеллектуального анализа данных для предложения инновационных решений, реализующей новый подход к поиску информации об инновациях.

Естественным первоочередным действием пользователя, которому необходимо получить максимально исчерпывающую информацию о возможных инновационных решениях задачи в какой-либо предметной области, является выполнение одного или нескольких поисковых запросов для поиска научно-технической информации в:

- ресурсах Интернет (издания общероссийских и отраслевых институтов информации, справочники, статьи и обзоры, материалы конференций, ГОСТы, технические регламенты, нормативно-техническая документация, отчеты о НИР/ОКР, рекламные материалы, статистические данные, экспертные оценки).
- специализированных базах данных (патентных, описаний изобретений и полезных моделей, промышленных образцов, реферативной и/или библиографической информации, товарных знаков).

В результате у пользователя в распоряжении будет большое количество данных, в той или иной степени релевантных соответствующим запросам [6, с. 1—9]. Следовательно, основной прикладной задачей исследований, реализуемых в рамках данного проекта, является уточнение

результатов информационного поиска инноваций с помощью средств и методов интеллектуального анализа данных. Одним из таких методов является общеизвестный и широко применяемый метод кластерного анализа электронных текстовых документов на естественном языке.

Наиболее фундаментальным трудом, посвящённым вопросам информационного поиска и, в частности, кластерного анализа, является [5, с. 353—402], где осуществлено подробное рассмотрение различных алгоритмов кластеризации текстов, по сей день активно применяющихся в многочисленных информационно-поисковых системах для кластеризации результатов поисковых запросов, в новостных порталах для выделения рубрик новостей и т. п. Вопросам интеллектуального анализа данных, кластеризации текстов также посвящёно множество публикаций в научно-популярных журналах по данной тематике, в Интернете и других источниках. Среди них можно выделить [7, с. 21—32] и многие другие.

В рамках патентных исследований проекта по разработке информационной системы поддержки инноваций предложен эвристический алгоритм фильтрации и семантического ранжирования результатов поиска документов, включающий в себя механизмы кластерного анализа, вычисления меры близости документов, основанные на использовании модели векторного пространства (VSM) документов, которая является фундаментальной для многих задач информационного поиска [2, с. 5].

Существует ряд важных задач, решение которых должно способствовать развитию как некоторых отдельных фрагментов указанного выше алгоритма, так и всего проекта в целом.

Во-первых, это касается автоматизации формирования множества поисковых запросов из описания общего запроса с помощью генетического алгоритма, оптимизирующего суммарную релевантность (или вес) результирующей выборки документов при заданных ограничениях на количество выполняемых операций (глубину эволюционного процесса). Ознакомиться с теоретическими основами генетических алгоритмов можно,

например, здесь [4, с. 180—193]. Отметим лишь, что основными операторами генетического алгоритма являются: скрещивание (операция, при которой две хромосомы обмениваются своими частями) и мутация (случайное изменение одной или нескольких позиций в хромосоме). Под «хромосомой», в данном случае, понимается вектор (последовательность), содержащий набор значений, что является ничем иным, как последовательностью слов, введённых пользователем, осуществляющим информационный поиск (поисковым запросом). Во время операции скрещивания из двух разных поисковых запросов автоматически формируется один запрос, в результате выполнения которого должна повыситься релевантность получаемых результатов поиска. Во время операции мутации поисковый запрос изменяется под действием внешних факторов, например, в результате действий самого пользователем или эксперта.

Кроме того, важнейшей задачей проекта является исследование изменения времени выполнения алгоритмов кластеризации текстовых документов, которые включает в себя единый эвристический алгоритм, описанный в [2, с. 1—10]. Данные патентные исследования проводятся, исходя из предположения о значительном нелинейном увеличении времени работы алгоритма вследствие увеличения объёма обрабатываемой коллекции документов. При решении данной задачи необходимо осуществить выполнение алгоритма в локальной и распределённой вычислительной средах с целью нахождения путей уменьшения времени его работы за счёт высокопроизводительных вычислений на кластере.

Подводя итог, можно сказать, что осуществление предложенных выше мероприятий по совершенствованию методов поиска научных знаний, инновационных предложений должно стать немаловажным фактором развития современной науки, особенно, в условиях огромных объёмов информации, с которыми сегодня приходится работать исследователям.

Работы проводились при финансовой поддержке РФФИ (договор № НК13-07-00342\13).

Список литературы

1. Байгулов Р.М., Рожкова Е.В. Управление промышленным предприятием: специфика поиска инновационных бизнес-идей // Современные проблемы науки и образования. — 2012. — № 2 [Электронный ресурс] — Режим доступа. — URL: www.science-education.ru/102-5896 (дата обращения: 01.10.2013).
2. Иванов В.К., Виноградова Н.В. Эвристический алгоритм фильтрации и семантического ранжирования результатов поиска документов // Вестник Тверского государственного университета: научный журнал: Серия «Прикладная математика». Твер. гос. ун-т. — 2013 (принята к публикации).
3. Куракова Н.Г., Зинов В.Г. Создание прорывных инноваций на основе комбинации научных заделов мирового уровня как компетенция инновационного менеджмента // Инновации. — 2012. — № 10. — С. 37—42.
4. Макаренко С.И. Интеллектуальные информационные системы: учеб. пособие. Ставрополь: СФ МГГУ им. М.А. Шолохова, 2009. — 206 с.: ил.
5. Маннинг К.Д., Рагхаван П., Шютце Х. Введение в информационный поиск: Пер. с англ. М.:ООО «Вильямс», 2011. — 528 с.: ил.
6. Палюх Б.В., Иванов В.К., Сотников А.Н. Архитектура интеллектуальной системы информационной поддержки инноваций в науке и образовании // Программные продукты и системы. — 2013. — № 4 (принята к публикации).
7. Пескова О.В. Методы автоматической классификации электронных текстовых документов без обучения: статья // Всероссийский институт научной и технической информации РАН. — 2006. — № 12. — С. 21—32.