Разработка мультиаспектной методики поиска документальной информации

© Е.И. Болотин

Национальный исследовательский ядерный университет «МИФИ»
eugenebolotin@gmail.com

Аннотация

Описаны проблемы, с которыми сталкиваются пользователи поисковых систем при осуществлении поиска документальной информации. В зависимости от уровня компетенции пользователей в предметной области рассматриваются различные стратегии поиска и их результативность. В работе предлагается подход к созданию методики, рекомендующей пользователям стратегию поиска, которая позволяет повысить качество поиска для различных типов пользователей.

1 Введение

В настоящее время наряду с популярными поисковыми системами интернета развиваются системы документального поиска, предоставляющие множество различных механизмов поиска. Поисковые машины интернета обычно предлагают один единственный механизм поиска, и основное развитие такого типа поисковых систем связано с разработкой алгоритмов оптимизации поиска, разработки методов ранжирования результирующих выборок, учета гиперссылок. Использование единственного механизма поиска позволяет использовать поисковые системы интернета неподготовленными пользователями, однако они являются менее управляемыми, чем документальные системы поиска. Документальные поисковые системы предлагают различные механизмы поиска, однако требуют подготовки от конечного пользователя. Процесс поиска в данном случае становится многоступенчатым, на каждом этапе возможно использование различных механизмов поиска. В настоящее время остается открытым вопрос о стратегии поиска в документальных системах поиска.

Механизмы поиска можно разделить на четкие и нечеткие. Четкие методы поиска позволяют искать документы, соответствующие строгому запросу пользователя – булевому выражению, содержащему дескрипторы. Нечеткие методы позволяют осуществлять автоматизированный поиск документов с использованием статистических характеристик и информации о релевантных документах, выбранных пользователем.

При использовании четких методов поиска пользователь полностью контролирует процесс поиска, включая в поиск новые дескрипторы, тем самым расширяя запрос. Однако дескрипторы лишь приблизительно описывают тематическое содержание документов и запросов. Поэтому обычно выдача в ответ на тематический запрос не бывает полной и точной. Одновременно с этим формулировка четкого запроса требует от пользователя понимания того, что и как он ищет. Это зависит от компетентности пользователя в предметной области поиска. Также пользователям свойственно включение в запрос контекстно-значимых терминов, статистические связи между терминами в заданной выборке неочевидны для пользователя.

С помощью нечетких методов возможно осуществление автоматического расширения выдачи с включением статистически-значимых терминов. Система предлагает пользователю документы, похожие на ранее выбранные им и отмеченные как релевантные. Управление осуществляется полностью системой, а не пользователем. С помощью нечетких методов могут быть выявлены статистические связи, которые не очевидны при осуществлении поиска по дескрипторам. Данный метод прост для пользователя, так как требует только указания группы релевантных документов. Таким образом, пользователь выбирает между методами точными, но требующими компетенции механизмами, и нечеткими, но простыми в использовании.

В связи с этой проблемой возникает задача объединения различных механизмов поиска в комплексные методики для упрощения работы пользователей с системой и повышения качества поиска. Также не исследована зависимость полноты поиска при различных стратегиях поиска.

Основными проблемами при решении данной задачи являются определение критериев использования того или иного механизма поиска, а также комплексирование результатов поиска.

Труды 12[й] Всероссийской научной конференции «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» – RCDL'2010, Казань, Россия, 2010

2 Стратегии использования механизмов поиска документальных данных

Схемы и механизмы поиска автоматизированной информационно-поисковой системы (АИПС) строятся в предположении, что любая нетривиальная реальная информационная потребность не может быть удовлетворена одним или несколькими сразу найденными документами, а требует проведения серии поисков и выделения полезных фрагментов информации на каждой стадии развития запроса. Это осуществляется следующими путями:

- переформулировкой и дополнением выражения запроса, в том числе использованием других терминологических систем;
- применением различных поисковых механизмов с разной степенью жесткости критерия отбора;
- использованием процедур поиска, основанных на технологии обратной связи по релевантности, обеспечивающих последовательное расширение пространства терминов и документов.

Поисковые системы должны обеспечить возможность использования различных механизмов поиска для реализации всех типов поисковых задач. Механизмы поиска по логическому выражению из терминов (вербальные) и поиска документов-аналогов (эвристические) образуют полную группу механизмов поиска [1], т. е. любой возможный механизм поиска документальных данных будет основываться на механизмах из данной группы.

Рассмотрим типовые стратегии поиска с использованием описанных выше механизмов:

- от вербального поиска к эвристическому;
- от эвристического к вербальному;

Выбор одной из стратегий поиска зависит от компетентности пользователя в предметной области, то есть от владения лексикой области. В случае, если пользователь компетентен, то вербальный поиск будет обеспечивать получение наилучшего результата. Однако вербальный поиск не может вывести пользователя за пределы лексики, используемой в запросе. Эвристический поиск позволяется обнаруживать смежные области и может вывести пользователя на новую лексику, которую можно использовать для поиска новых релевантных документов. Если пользователь некомпетентен, то эвристический поиск позволит пользователю освоить лексику предметной области и в дальнейшем эффективно использовать вербальный поиск. Указанные стратегии могут повторяться в рамках осуществления одного поиска с целью добиться наибольшего показателя полноты поиска.

Рассмотрим экспериментальные данные, собранные на основе проведения лабораторных работ студентов 4-го курса НИЯУ МИФИ в рамках изучения дисциплины «Информационные системы». Студентам предлагалось осуществить проработку заданной темы с использованием АИПС «Irbis» [2]. Данная система предоставляет большой набор механизмов поиска, средства для оценки релевантности найденных документов, выполнения операций над множествами найденных документов. Студенты использовали следующие механизмы поиска: вербальные – поиск по ключевым словам, по полю «реферат», по всем полям, поиск с использованием автомаскирования (для нормализации лексики), невербальный – эвристический механизм поиска. Эвристический механизм позволяет осуществлять поиск документов, похожих на усредненный документ выбранной совокупности релевантных документов.

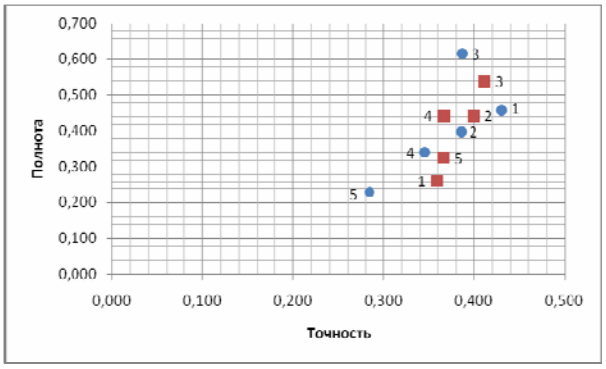
Были обработаны 30 работ студентов. Полученные в результате эксперимента данные – точность и полнота отдельных поисковых механизмов – были кластеризованы на две группы по соотношению показателя полноты эвристического механизма поиска. Рассмотрим усредненные показатели обеих групп. Для первой группы получились следующие показатели:

Группа 1			
		Среднее	**Дисперсия**
Ключевые слова	p	**0,430**	**0,024**
	r	**0,458**	**0,039**
Реферат	p	**0,386**	**0,042**
	r	**0,398**	**0,072**
По всем полям	p	**0,386**	**0,022**
	r	**0,615**	**0,057**
С автомаскированием	p	**0,345**	**0,031**
	r	**0,342**	**0,089**
Эвристический поиск	p	**0,284**	**0,019**
	r	**0,228**	**0,043**

Для второй группы:

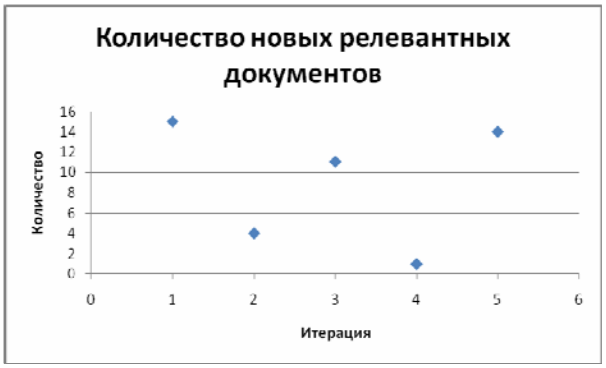
Группа 2			
		Среднее	**Дисперсия**
Ключевые слова	p	**0,359**	**0,018**
	r	**0,261**	**0,024**
Реферат	p	**0,366**	**0,027**
	r	**0,442**	**0,049**
По всем полям	p	**0,411**	**0,048**
	r	**0,539**	**0,063**
С автомаскированием	p	**0,400**	**0,035**
	r	**0,442**	**0,088**
Эвристический поиск	p	**0,366**	**0,047**
	r	**0,326**	**0,060**

Полученные результаты могут быть представлены в виде точечной диаграммы. Цифрами обозначены механизмы поиска: 1 – поиск по ключевым словам, 2 – поиск по полю «реферат», 3 – поиск по всем полям, 4 – поиск с автомаскированием, 5 – эвристический поиск:

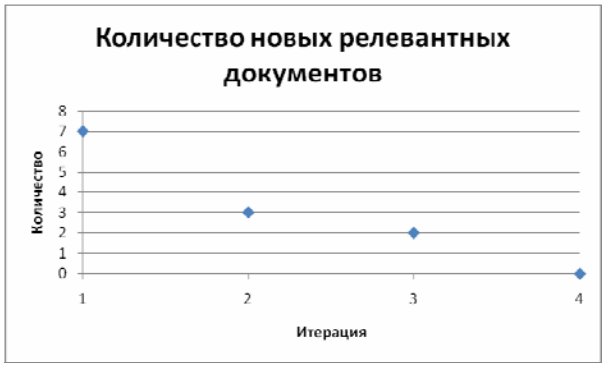


Полученные на графиках результаты отражают разделение пользователей на две группы по признаку компетентности в предметной области поиска. Первая группа показывает наибольшие показатели полноты у вербальных механизмов поиска, эвристический механизм обеспечивает наименьшую полноту из рассмотренных механизмов. По описанным выше типовым стратегиям поиска эта группа отражает результаты компетентных пользователей, которые имели четкое представление о терминологии рассматриваемой области и в меньшей степени полагались на эвристические механизмы. Для второй группы эвристический механизм дает лучшие результаты, чем поиск по ключевым словам. Это показывает, что пользователь изначально не был осведомлен о лексике, используемой в предметной области, и полагался на эвристические методы, выполняющие всю работу в автоматизированном режиме.

Также был проведен профессиональный экспертный поиск по разработке темы «Исследования по использованию тория в реакторах» с привлечением эксперта в области ядерной физики для формирования исходного представления о проблемной области. Для исследования заданной области была использована документальная база данных с материалами, касающимися исследований в сфере ядерной промышленности. Интерфейс к БД также обеспечивала система Irbis. Базируясь на сведениях, предоставленных экспертом, был осуществлен поиск по ключевым словам с расширением исходного запроса терминами, найденными в документах. Поиск по ключевым словам неравномерен по количеству получения новых результатов, т. к. каждое введенное понятие может определять различные подобласти исходной тематики:



На определенной итерации поиска отсутствуют новые понятия, которые возможно включить в запрос для расширения области поиска. Как говорилось выше, это недостаток вербальных механизмов поиска, так как они не могут вывести пользователя за пределы лексики, используемой в запросе. В таком случае целесообразно применение невербальных механизмов поиска. В данной работе применялся эвристический механизм. Производилось несколько итераций эвристического поиска до тех пор, пока он не переставал обнаруживать новые результаты. Рассмотрим усредненную кривую насыщения результативности эвристического механизма поиска:



Данный график показывает, что после определенной итерации эвристический поиск перестает обнаруживать новые результаты. Результатом работы эвристического механизма является обнаруженная новая лексика, которая вновь позволяет производить вербальный поиск. В данном эксперименте в среднем производилось 3 цикла вербального-невербального поиска, прежде чем поиск переставал обнаруживать новые релевантные документы. Это также доказывает необходимости сочетания различных механизмов для обеспечения максимальной полноты поиска.

3 Методика мультиаспектного документального поиска

Как было показано выше, максимальные значения полноты поиска могут быть получены путем сочетания вербальных и невербальных механизмов. Суть методики мультиаспектного поиска состоит в циклическом использовании вербальных и невербальных механизмов с целью расширения исходной лексики предметной области и нахождения документов, соответствующих этой лексике.

Любой поиск начинается с формирования исходного запроса в виде множества T_0 дескрипторов, исходя из начальных знаний пользователя о предметной области. Так как формулировка запроса в виде дескрипторов наиболее естественна для пользователя, первым этапом методики осуществляется вербальный поиск. На каждой итерации вербального поиска пользователь выбирает среди найденных документов релевантные R_i и релевантные термины T_i, которые могут расширить исходный запрос. От-

меченные релевантные документы и термины заполняют общее множество релевантных документов и терминов:

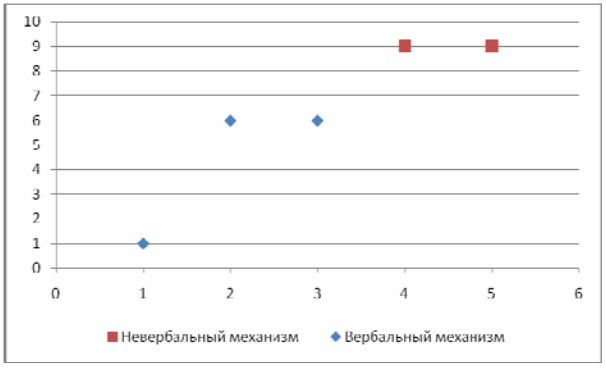
$$R = R \cup R_i, \quad T = T \cup T_i.$$

Вербальный механизм используется до тех пор, пока удается расширять выборку релевантных документов и терминов. Когда в результате работы вербального механизма новые термины не обнаруживаются, используется невербальный поиск для расширения лексики. Итерации невербального алгоритма позволяют пользователю обнаружить новые релевантные документы R_i и выделить среди дескрипторов новые термины T_i предметной области. В случае обнаружения пользователем новых дескрипторов запрос по ключевым словам будет расширен и будет начат новый этап вербального поиска. Если невербальный поиск не обнаруживает новых релевантных документов, то алгоритм завершает работу.

Итак, входными данными алгоритма является исходный вербальный запрос. В результате работы алгоритм дает следующие результаты:

- совокупность R найденных релевантных документов;
- совокупность T найденных терминов (лексики) предметной области.

В рамках эксперимента по разработке темы «Исследования по использованию тория в реакторах» была испльзована данная методика поиска. Были получены следующие результаты насыщения результативности поиска при осуществлении итераций вербального и невербального поиска:



В данном случае потребовался один цикл использования вербального и невербального поиска, так как новых понятий в процессе использования эвристического механизма выявлено не было.

4 Заключение

В данной работе проводилось исследование различных типовых стратегий поиска, используемых пользователями. Было показано, что использование единственного механизма не позволяет достичь той полноты, которая может быть достигнута при использовании нескольких механизмов поиска. В зависимости от типов поисковых задач и типов пользователей выбирается определенная схема применения механизмов. Было обнаружено, что компетентные в предметной области пользователи успешно используют вербальные механизмы, дающие в данном случае наибольшие показатели полноты поиска. Некомпетентные пользователи полагаются на эвристические механизмы, так как не владеют лексикой предметной области.

Предлагается методика мультиаспектного поиска, позволяющая обеспечить максимальную полноту для различных типов пользователей. Суть методики состоит в циклическом применении вербального и невербального механизмов. Вербальный поиск используется для обнаружения документов, соответствующих лексике предметной области, обнаруженной на определенной итерации, и ее расширению за счет новых терминов, выявленных в найденных документах. В случае, когда вербальный механизм не обнаруживает новых терминов, предлагается использование невербальных механизмов для выявления смежных областей и обнаружения новой лексики предметной области.

Дальнейшая работа по развитию предложенной методики заключается в оптимизации методики с точки зрения повышения точности поиска, а также уменьшения общего количества просматриваемых пользователем документов.

Литература

[1] Голицына О.Л., Максимов Н.В., Попов И.И. Информационные системы. – Форум: Инфра-М, 2007, 496 с.

[2] Максимов Н.В. Документальная информационно-аналитическая система xIRBIS: программа для ЭВМ / Максимов Н.В., Васина Е.Н., Голицына О.Л. и др. // Свидетельство о гос. регистрации №2008611511 от 25.03.2008.

Development of a multiaspect method to search inside documents

E.I. Bolotin

This paper describes a number of problems that users of search systems in general face while looking for various document information. Depending on level of the competence of users in a subject domain, strategy and productivity of their search will be various. The multiaspect search technique is offering to raise quality of search for various types of users.