А.С. Злыгостев (руководитель С.И. Родзин)

ИСПОЛЬЗОВАНИЕ КОНТЕКСТОВ КЛЮЧЕВЫХ СЛОВ ДЛЯ СОРТИРОВКИ РЕЗУЛЬТАТОВ В ИНФОРМАЦИОННОЙ ПОЛНОТЕСТОВОЙ ПОИСКОВОЙ СИСТЕМЕ

Введение. В связи с экспоненциальным ростом информации поисковые системы в Интернет сталкиваются с необходимостью обрабатывать всё большее число документов и задачей улучшения качества поиска. Данная статья посвящена улучшению сортировки найденных документов по релевантности[1] в полнотекстовых информационных поисковых системах (ИПС) благодаря анализу взаимоупотребляимости ключевых слов внутри текстов документов.

Информационный поиск – это процесс отыскания в каком-то множестве документов тех, которые посвящены указанной в информационном запросе теме (предмету) или содержат необходимые потребителю факты сведения.[1] Информационный поиск основан на использовании – вместо последовательного просмотра полных текстов документов – поисковых образов этих документов.

Развитие поиска информации в сети Интернет шло от каталогов, составляемых вручную, к полнотекстовым ИПС, выполняющим поиск автоматически. Основные модели поиска были предложены задолго до появления поисковых систем в Интернет. В последние годы в основном добавляются надстройки над уже давно используемыми моделями, позволяющие улучшить результаты поиска.

Строение ИПС. Общая схема функционирования представлена на рис.1.

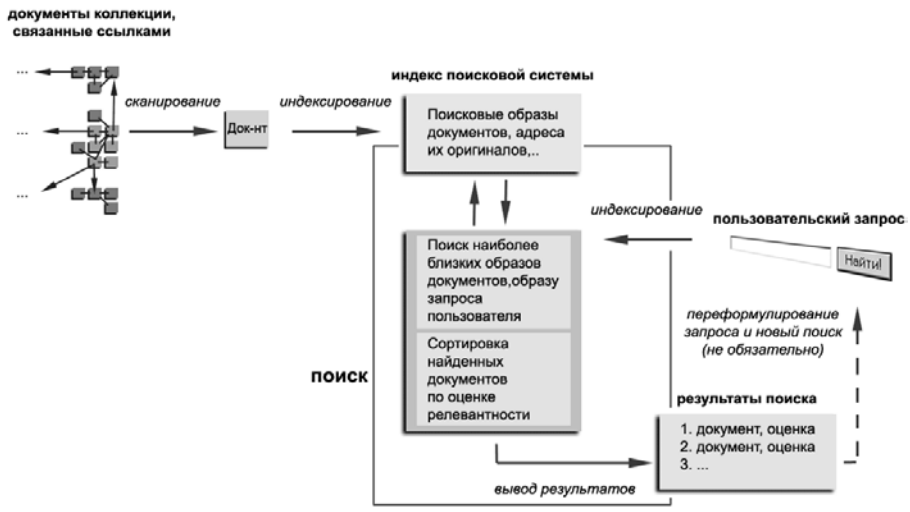

Рис.1. Схема функционирования ИПС

ИПС отрабатывают три основных этапа: сканирование, индексирование и поиск.

На этапе *сканирования* сетевые роботы просматривают общедоступные информационные ресурсы, опубликованные в Интернет, и копирует новые (обновленные) документы, информация о которых отсутствует (устарела) в базе данных ИПС.

[1] Релевантность – (от англ. relevant – “значимый; существенный; важный”) означает степень близости документа поисковому запросу пользователя.

Этап *индексирования* – это сложный процесс «компрессии» текста и перевода его содержания с естественного на информационный язык.[1] При включении нового документа в коллекцию выполняется его индексирование, состоящее из таких шагов как разбор текста документа и запись информации в базу данных. Логическая структура данных для представления индекса поисковой системы зависит от выбранной модели поиска. Основные модели полнотекстового поиска основанного на анализе статистических данных – булева модель, векторная модель, вероятностная модель (относится к многоитерационным видам поиска), латентно семантическое индексирование. Главная задача индекса ИПС это быстрое получение информации о релевантности документа.

На этапе *поиска* – ИПС получает запрос пользователя в виде набора ключевых слов, которые преобразуются в некоторый поисковый образ, после чего через поисковый индекс находятся соответствующие запросу документы. Полученные документы ранжируются[2] по степени близости запросу, и список их адресов, с некоторой дополнительной информацией (например, заголовок, краткое описание документа), возвращается пользователю.

Предлагаемая модель учёта контекстов для сортировки результатов поиска будет задействована на этапе индексирования (при формировании индекса поисковой системы) и в процессе поиска (при оценке релевантности документа пользовательскому запросу).

Модели поиска. Рассмотрим основные модели поиска.

Булева модель является старейшей и наиболее широко используемой моделью информационного поиска. Её распространение связано в первую очередь с простотой реализации, позволяющей индексировать и выполнять поиск в коллекциях документов большого объёма. Строение индекса представлено на рис.2.

документ / терм	index1.htm	index2.htm	index3.htm	index4.htm	index5.htm	index6.htm	index7.htm	...
абрикос	0	0	1	0	0	0	1	
август	0	0	0	0	0	0	0	
австрия	1	0	0	0	0	0	0	
агротехника	0	0	1	0	1	0	0	
адрес	1	0	0	0	0	0	0	
азия	0	1	0	0	1	0	0	
азотистый	0	1	0	0	0	0	0	
айва	0	0	1	1	0	1	0	
акация	0	1	0	0	0	0	1	
...								

Рис.2. Поисковый индекс для булевского поиска в табличном виде. В строке содержится информация о включенности терма в документы коллекции (0 - терм не встречается в документе; 1 – терм встречается)

Индекс поисковой системы для булевской модели поиска можно представить в виде таблицы, в ячейках которой содержится информация о встречаемости терма[3] внутри документа.

Булева модель позволяет обрабатывать запросы пользователя, составленные из ключевых слов, соединённых логическими операциями И ИЛИ НЕ.

[2] Ранжирование – (от англ. range – “располагать в порядке”) сортировка найденных документов в порядке релевантности запросу пользователя.

[3] Терм (от англ. term – “термин”) используется для обозначения элементов поискового образа документа, таких как слова, основы слов, фразы и т.п.

К недостаткам модели относят отсутствие определения степени релевантности документа запросу пользователя. Для сортировки используют эвристики: повторяемость ключевых слов в документе и их присутствие в заголовках, присутствие ключевых слов в текстах ссылок на данный документ, индекс цитирования документа, социальная сеть, рейтинги посещаемости страницы.

Векторная модель поиска была предложена Г. Салтоном в конце 80-х годов. Каждому документу сопоставляется вектор в базисе слов. Пользовательский запрос также преобразуется в вектор. В качестве результата поиска выдаются документы, углы, между векторами которых и вектором запроса, имеют минимальное значение. Строение индекса векторной модели представлено на рис.3.

терм / документ	абрикос	август	австрия	адрес	азия	азотистый	айва	акация	...
index1.htm	0	0	0,0075	0	0	0	0,021	0	
index2.htm	0	0	0	0	0	0	0	0,033	
index3.htm	0,023	0	0	0	0	0	0	0	
index4.htm	0	0	0,117	0	0,0034	0	0	0	
index5.htm	0,0761	0	0	0	0	0	0	0	
index6.htm	0	0,144	0	0	0,0023	0	0	0	
index7.htm	0	0,093	0	0	0	0	0	0	
index8.htm	0	0	0,0567	0,0112	0	0,019	0	0	
...									

Рис.3. Поисковый индекс векторной модели поиска в табличном виде. В строке содержится вектор документа в пространстве термов. В ячейках содержится оценка веса терма для документа

Формула для вычисления веса терма t для документа d по методу tf*idf выглядеть следующим образом: w(d,t)=tf(d,t)*log(|C|/df(f)), где tf - частота встречаемости термина в документе, df - число документов коллекции, содержащих данный терм. Вектора документов в пространстве термов нормализуется (приводятся к единичной длине).

Векторная модель позволяет получить оценку релевантности документа запросу пользователя. При этом оценка учитывает пропорциональность использования терминов из запроса – размеру документа. Один из главных недостатков векторной модели – это большая трудоёмкость, связанная с необходимостью перемножить вектор запроса на вектора всех документов коллекции.

Вероятностная модель воплощает идею многоитерационного поиска. На первой итерации поиска пользователю выдаются все документы, содержащие термы из запроса пользователя. Пользователь выбирает несколько подходящих документов. Происходит расширение запроса пользователя за счёт термов, содержащихся в выбранных пользователем документах. Далее процесс поиска повторяется с уже расширенным запросом поиска.

На каждой новой итерации происходит пересчёт релевантности документов пользовательскому запросу. Велика трудоёмкость метода из-за необходимости пересчёта оценок для всех документов коллекции на каждой итерации поиска.

При помощи *латентного семантического индексирования* (далее LSI) представляется возможным организовать поиск, учитывающий скрытые взаимосвязи элементов текста. Идея этого метода заключается в том, что поиск осуществляется не в индексе документов, как это делается в булевской или векторной моделях поиска, а в пространстве латентных (т.е. скрытых) факторов. Поиск при использовании LSI осуществляется с учётом схожести документов по используемым в них словам, и с учётом схожести слов по значимости (употребляемости) их в документах.

В его основе лежит известная в курсе линейной алгебры процедура сингулярного разложения матрицы, производимая над матрицей A размерности LxK зависимости слов и документов (L- число термов в словаре коллекции, а K - число документов, L>K). Результатом разложения становятся матрицы U, S и V, обладающие радом свойств, определяющим из которых является следующие: $A=USV^T$.

При получении запроса q от пользователя его переводят в пространстве латентных факторов: $q'=q^TSU^{-1}$.

Теперь мера близости запроса q и документа d оценивается величиной скалярного произведения векторов q' и $V^T[d]$. Здесь $V^T[d]$ обозначает d–столбец матрицы V^T.

Метод латентно-семантического индексирования непригоден для обработки больших коллекций документов из-за высоких требований к оперативной памяти и вычислительным ресурсам. Метод требует подготовленной таблицы с весами термов в документах, что для больших коллекций может быть составлено только автоматически и привносит ошибку в результаты поиска.

Дополнительные критерии оценки термов и документов. При создании первых информационно-поисковых систем в основе поиска лежал только текст. Предполагалось, что именно его следует анализировать, чтобы отыскать нужные пользователю документы. Современные поисковые системы значительно расширили критерии, используемые при ранжировании найденных документов.

Отметим некоторые приёмы, используемые для оценки релевантности документов в современных поисковых системах:

1. *Положение ключевых слов в документе*. Вес ключевого слова в документе может быть увеличен, если оно встречается в заглавии гипертекстового документа, входит в метатеги списка ключевых слов или описания документа, находится в заголовках документа, выделено жирным шрифтом. Данный подход в частности использует поисковая система Aport.ru

2. *Оценка авторитетности документа*. Поисковик Вебальта при сортировке результатов поиска использует показатель «Уровень доверия», Яндекс понятие «Индекс цитирования», а Google понятие «PageRank» (что на английском означает “категория страницы” или “класс страницы”). Индекс подсчитывается поисковыми системами и зависит от количества ссылающихся на сайт ресурсов сети, от цитируемости ссылающихся ресурсов, времени регистрации доменного имени, репутации компаний, на чьих серверах физически размещёна страница и др. Чем выше показатель страницы, тем выше его место в результатах поиска. Алгоритмы подсчёта таких показателей довольно запутаны и поисковые системы не очень стремятся их разъяснять.

3. *Оценка популярности документа в поисковике рейтинге*. Один из способов оценки популярности страниц – ведение статистики её посещаемости. Например, Rambler Top 100 предлагает установить на страницы сайта счётчик посетителей. В результате при запросе поиска система Rambler выдаёт на первых позициях страницы соответствующие данному запросу и лидирующие в рейтинге.

4. *Оценка популярности документа при помощи социальной сети*. Попытку использования «социальной сети» в поисковой системе предпринял сайт Eurekster (http://www.eurekster.com). Он сразу объявил себя поисковой системой с элементами так называемой «социальной сети». Если в Google самыми важными считаются те страницы, к которым ведёт больше всего ссылок, то в Eurekster первыми выдаются страницы, которые наиболее популярны среди посетителей.

Определение релевантности по нахождению слов в контексте. Автором была разработана надстройка над булевской моделью поиска, позволяющей сорти-

ровать результаты поиска с учётом оценки употребляемости слов из пользовательского запроса внутри собственных контекстов в найденных документах.

Для тематики поиска строится матрица взаимоупотребляемости термов внутри тематической коллекции документов. В данную матрицу не входят стоп-слова[4]. При подсчёте взаимоупотребляемости суммируются оценки, даваемые за каждую встречу двух слов внутри абзаца, текста под заголовком, внутри одного документа. Таким образом, чем ближе два слова в тексте, тем больше оценку взаимоупотребляемости они получат.

В качестве примера контекста для терма «*сахар*» в наборе тематических документов по виноделию (использовались материалы сайта wine.historic.ru) был найден контекст из слов: *вино, вода, количество, спирт, сусло, брожение, кисло, ягода, необходимо, вкус, плод, виноделие, добавляем, вещество*. А для терма «*вино*» был найден контекст из слов: *сахар, вода, сусло, бутылка, количество, брожение, спирт, вкус, необходимо, приготовление, день, ягода, температура, виноградный, кислое, виноград, дрожжи, осадок, способ, столовое, винодел, сорт, десертное, красное, качество, час, сладкое, плод, сухое, добавляем, крепкое, условие* (слова из контекста приведены в порядке уменьшения оценки значимости слова для контекста).

После подсчёта оценок взаимоупотребляемости обнуляются шумовые элементы матрицы. Шумовыми взаимоупотребляемостями считаем все с оценкой со значением ниже константы, выбираемой в зависимости от размера словаря термов внутри анализируемой тематической коллекции документов.

Строки, соответствующие контексту терма, нормализуются, приводятся к единичной сумме оценок взаимоупотребляемости внутри одного контекста. Пример матрицы контекстов приведён на рис.4.

терм / контекст терма	абрикос	август	австрия	адрес	азия	азотистый	айва	акация	…
абрикос	0	0	0,0075	0	0	0	0,021	0	
август	0	0	0	0	0	0	0	0,033	
австрия	0,023	0	0	0	0	0	0	0	
адрес	**0**	**0**	**0,117**	**0**	**0,0034**	**0**	**0**	**0**	
азия	0,0761	0	0	0	0	0	0	0	
азотистый	0	0,144	0	0	0,0023	0	0	0	
айва	0	0,093	0	0	0	0	0	0	
акация	0	0	0,0567	0,0112	0	0,019	0	0	
…									

Рис.4. Матрица контекстов. Для терма «адрес» мы видим, что была взаимоупотребляемость с термами «Австрия» и «Азия»

Процедура поиска происходит в три этапа. Первый этап это ввод пользователем набора ключевых слов, по которым осуществляется поиск. Второй этап осуществляет булевский поиск по коллекции документов, т.е. выделяется множество документов из коллекции, которые включают слова из пользовательского запроса. Третий этап осуществляется при помощи матрицы совместной встречаемости термов и определяет степень нахождения терма внутри своего контекста. Релевантность будет сформирована как сумма оценок нахождения термов в своем контексте, перемноженных на коэффициенты важности слов.

Оценки нахождения слов внутри своего контекста будут находиться при помощи поискового индекса, сформированного для булевского поиска простой про-

[4] Стоп-слова – это наиболее употребительные в данном языке слова, удаление которых не повлияет на качество поиска, более того, может его улучшить (предлоги, союзы, местоимения и иные служебные слова)

веркой на встречаемость слов из контекста в найденном документе. Таким образом, трудозатраты предположительно возрастут в линейном порядке по сравнению с булевским поиском.

Заключение. К достоинствам метода анализа контекстов для оценки релевантности можно отнести направленность на анализ текста документов, а не на внешние факторы, например, на цитируемость страницы или авторитетность вебсайта, который включает данный документ, т. к. данные методы в последнее время дискредитировали себя из-за возможностей извне влиять на результаты поиска. Метод предназначен для тематического поиска. Единожды построив матрицу взаимоупотребляемостей для тематики, можно использовать её для разных коллекций документов схожей тематики. Также матрица контекстов может быть использована для расширения запроса пользователя, для поиска синонимов и омонимов (за счёт анализа контекстов для термов из контекста анализируемого терма).

БИБЛИОГРАФИЧЕСКИЙ СПИСОК

1. *Черный А.И.* Введение в теорию информационного поиска. – М.: Наука, 1975.
2. *Рыбаков Ф.И., Руднев Е.А., Петухов В.А.* Автоматическое индексирование на естественном языке. – М.: Энергия, 1980. – 160 с.
3. *Ландэ Д.В.* Поиск знаний в INTERNET. – М., Спб, Киев: Диалектика, 2005. – 272 с.
4. *Злыгостев А.С., Сяйлев И.А.* Применение методов линейной алгебры в поисковых системах на примере латентного семантического индексирования //Сборник трудов IV Всероссийской научной конференции молодых ученых, аспирантов и студентов "Информационные технологии, системный анализ и управление", Таганрог, ТРТУ, 17 ноября 2006г. - Таганрог: Изд-во ТРТУ, 2006. – С. 68-69.
5. http://www.aport.ru/.
6. http://www.google.com/.
7. http://www.webalta.net/ru/about_webmaster_trust.html.

Ю.И. Рогозов, А.Н. Самойлов

ОПРЕДЕЛЕНИЕ НАПРАВЛЕНИЙ РАЗРАБОТКИ АВТОМАТИЗИРОВАННЫХ СИСТЕМ ИЗМЕРЕНИЯ ОБЪЕМА КРУГЛОГО ЛЕСОМАТЕРИАЛА

Такая отрасль народного хозяйства как деревообрабатывающая в настоящее время переживает бурный подъем. Вопрос точного учета сырья и производимой продукции является одним из важнейших в условиях рыночных отношений и постоянной борьбы за минимизацию издержек производства.

Актуальной задачей является разработка автоматизированной системы измерения объема круглого лесоматериала. В данной статье классифицируем все методы измерения объема лесоматериалов, наиболее часто используемые в лесной промышленности с целью определения направлений разработки автоматизированной системы.

В литературе [1] применяются два классификационных признака: первый – по количеству измеряемых лесоматериалов за один цикл измерения, второй – по принципу измерения. Эти классификации отображены на рис.1,2.