МАТЕМАТИЧЕСКИЕ МОДЕЛИ И ВЫЧИСЛИТЕЛЬНЫЕ СТРУКТУРЫ
(Вычислительные системы)
2004 год Выпуск 173
УДК 681.3.06

ИССЛЕДОВАНИЕ ВОЗМОЖНОСТИ ПРИМЕНЕНИЯ ЭЛЕКТРОННЫХ СЛОВАРЕЙ ДЛЯ ПОВЫШЕНИЯ ЭФФЕКТИВНОСТИ ПОИСКОВЫХ СИСТЕМ ИНТЕРНЕТА[1]

А.С. Климов

Введение

В наши дни проблема поиска информации в Интернете является одной из самых острых и актуальных. В Интернете уже содержится свыше 10 миллиардов страниц, обладающих самой различной информацией, и эта цифра продолжает увеличиваться. Нужно также учесть, что Интернет — весьма динамичная система. И кроме независимого обновления информации (например, издание новых Интернет–публикаций, смена товаров в Интернет-магазинах и пр.) существует ярко выраженная зависимость от самого появления новых поисковых систем, изменение в алгоритмах уже существующих систем и пр. Так что, методы поиска информации в подобном носителе отличаются от методов поиска информации в наборе статических текстов и встречают порой очень неожиданные проблемы. Например, одной из основных проблем поисковых систем на данный момент является личный интерес владельцев сайтов в том, чтобы их сайт был

[1]Работа выполнена при финансовой поддержке по проекту 8328 программы Рособразования "Развитие научного потенциала высшей школы".

первым в качестве результата поиска по некоторым ключевым словам. Эти и другие проблемы будут описаны в данной работе, но основная цель — демонстрация построенной поисковой системы, использующей систему Google и словарь морфем русского языка.

Итак, формулировка постановки задачи сводится к следующему.

- Анализ проблем современных поисковых систем. Классификация современных поисковых систем относительно предложенных решений по преодолению этих проблем.
- Создание системы по разрешению проблемы неполноты поиска с помощь вариации выделенных слов запроса их морфологическими формами:
  - создание программы, которая может использовать базы данных и алгоритмы уже существующей поисковой системы;
  - подключение в проект словаря морфологических форм слов русского языка;
  - обработка полученных результатов и представление их пользователю.
- Анализ работы созданной системы.

В таком порядке и будет изложен материал.

Обзор проблем

Основные проблемы поиска в Интернете можно разделить на два пункта:

1) нерелевантность поиска,

2) неполнота поиска.

Рассмотрим более подробно суть этих проблем и причины, по которым они возникают.

1. Нерелевантность поиска — несоответствие результатов поиска ожиданиям пользователя. Эта проблема рождается в первую очередь из-за выборки непредставительных ключевых слов, либо непредставительного набора ключевых слов. Чаще всего это ошибки неопытного пользователя, связанные с низкой культурой использования современных поисковых систем, незнанием предметных областей и неточностью перевода.

ПРИМЕР. *Допустим, пользователь хочет найти информацию о машинном масле в англоязычной части Интернета. Он ищет перевод для слова «масло» и затем вводит в поисковую машину слово «butter». В результате, он получит информацию о коровьем, а не машинном масле.*

Самое радикальное решение этой проблемы — дать возможность пользователю вести диалог с поисковой машиной. Однако отсутствие какой–либо более–менее приемлемой лингвистической модели на данный момент делает эту задачу неразрешимой. Но можно использовать более простые методы: из текста или запроса на естественном языке извлекаются ключевые слова, по которым (возможно после небольшой поправки пользователем) в дальнейшем уже ведется поиск. Очевидно, что такой способ не очень далеко ушел от обычного поиска по ключевым словам.

Существует еще одна причина информационного шума, который приводит к нерелевантным результатам: многозначность. Суть этой проблемы лучше понять на примере. Необходимо лишь указать, что эта проблема лежит в корне и проблемы машинного перевода, и проблемы анализа текста, написанного на естественном языке, и проблемы создания системы поиска высокого качества.

ПРИМЕР. *Предположим, пользователь снова хочет найти информацию о машинном масле. Посмотрим, сколько значений имеет слово «масло» по толковому словарю Ожегова.*

1. Жировое вещество, приготовленное из веществ животного, растительного или минерального происхождения.

Растительное м.

Животное м.

Сливочное м.

Смазочные масла

Фр.: Подлить масла в огонь

М. масляное

Как по маслу

Как маслом по сердцу

2. Такое вещество как пищевой продукт

Фр.: Хлеб с маслом

Жарить на масле

Сбивать м.

Маслом кашу не испортишь

3. Масляные краски, а также картина написанная ими

Др.

Масляные краски

Масляное пятно

Масляная живопись

Масляный выключатель

Как мы можем заметить для слова «масло» можно выделить по меньшей мере четыре смысловых значения: 1) смазка, 2) пищевой продукт, 3) масляная краска, 4) картина, написанная масляной краской. Учтем также все фразеологизмы (например, «как сыр в масле кататься») — они также служат источником шума.

В англо-русском политехническом словаре мы встречаем следующие пояснения к слову «масло»:

масло 1. тех. oil

с.сочетания:

варить м. boil an oil

вводить загустителъ в м. thicken an oil

вводить присадки в м. dope an oil

м. вспениваётся the oil churns (foams)

загущать м. give (more) body to an oil

м. застывает the oil solidifies

м. коксуется the oil has carbon-forming prpperties

компаундировать м. blend (the) oil

обесцвечивать м. decolourize oil

осветлять м. clarify oil

отбеливать м. bleach oil

отжимать м. isolate oil by pressing (by expression)

очищать м. refine oil

продувать м. (для окисления примесей) blow the oil

прокачивать м. circulate oil (through a system)

уплотнять м. body oil

2. (коровье) butter

(растительное) oil

м. горкнет butter becomes rancid

пахтать (сбивать) м. churn butter

...

и там же далее свыше 100 устойчивых сочетаний этого слова:

авиационное м. aviation oil
автотракторное м. motor oil
арахисовое м. peanut (groun-nut) oil
ацетоновое м. acetone oil
белое м. white oil
вазелиновое м. petrolatum, petroleum jelly
веретенное м. spindle oil
всесезонное м. (авто) multigrade oil
высыхающее м. drying oil

...

Более того, в этом же словаре есть еще 34 термина, образованные от корня «масло»:

Маслобак oil tank
Маслобензостойкость oil-and-petrol resistance
Маслобойка butter chorn

...

Обычно эту проблему можно разрешить с помощью тематического указания, задающего направление поиска. Но при способе поиска через ключевые слова этот способ не годится — здесь тематику задают сами ключевые слова. И нам бы понадобилось едва ли не столько же тематических разделов, сколько существует значений у ключевых слов.

Этот момент может сработать при организации поиска по образцу текста.

Выделим также третью проблему нерелевантности поиска: система на данный момент не может автоматически отслеживать конечную цель поиска пользователя. Приведем очередной пример, чтобы пояснить эти слова.

ПРИМЕР.

Пользователь по-прежнему ищет информацию о машинном масле. Однако:

• *его не интересует информация о том, где это масло можно купить или продать,*

• *его не интересует информация о том, как это масло производить,*

• *его интересует информация о методах исследования морозостойкости масла.*

Современные системы на данный момент не обладают функциональностью, чтобы решить эту проблему. Пользователю приходится либо просматривать массу ненужной информации, либо подыскивать нужные сочетания ключевых слов («масло, морозостойкость», «масло, свойства, исследование» и т.п.). При этом не дается никаких гарантий, что это приведет к удовлетворительному результату.

2. Обсудим теперь проблему неполноты поиска. Проблему можно описать следующим образом: информация, интересующая пользователя, в базе данных есть, она имеет косвенное отношение к запросу, но найти этот текст система не может.

Самая очевидная причина этого явления — синонимия. Однако одни синонимичные связи не объясняют эту проблему. Тут в дело вступают тонкие семантические связи, которые автоматическая система на данный момент отыскать не в силах.

Также к этой же проблеме можно, например, отнести неспособность поисковой системы Google воспринимать морфемы одного и того же русского слова как единую логическую единицу поиска. А это становится критичным при довольно длинных запросах, состоящих из четырех–пяти и более слов (опытный пользователь использует примерно такой диапазон для сужения предметной области и получения более релевантных результатов).

В качестве более-менее приемлемого решения здесь возможно послужат тезаурусы-словари.

«Спам» — дополнительная проблема поиска в Интернете

Спамом в нашем случае будем называть информацию, которая негативно влияет на работу поисковой системы. Причина наличия подобной информации банальна — владелец сайта, которому удалось «обмануть» поисковую систему имеет шанс безосновательно подняться в результатах поиска по некоторым ключевым словам, что обеспечит хорошую рекламу данному сайту.

Данная проблема проявила себя на заре возникновения поисковых систем и тогда получила название — «спамдексинг» (буквально — спам индексов поисковых систем).

Опишем основные приемы «спамдексинга» и то, как с ними боролись.

1. *Использование «чужих» слов.* Для обеспечения работы поисковых систем на самой заре их появления владелец сайта указывал набор ключевых слов, индексируя которые поисковые системы «понимали» какого рода информация представлена на данной странице. Это делается с помощью тэга meta keywords, который включается в HTML-код страницы и который содержит набор ключевых слов. Изначально поисковые системы искали соответствие между этими словами и словами запроса пользователя. Очевидно, чем чаще пользователи употребляют некоторое слово, тем чаще будут появляться ссылки на страницы, содержащие это слово в тэге meta keywords. Таким образом, используя ряд наиболее употребимых слов можно было легко обмануть поисковую систему и подняться в рейтинге поисковых машин.

Таким образом, поисковые системы сделали шаг в своем развитии — сейчас тэг meta keywords не играет большой роли в индексации страницы. Появилась система весов слов, суть которой заключается в следующем: анализируется не только содержимое тэга meta keywords, но и самого содержимого страницы. При этом проверяется, какое отношение имеет количество вхождений представленного ключевого слова в сам текст к количеству слов в тексте (за исключением предлогов и прочих не несущих смысла частей речи). Это отношение и носит название «вес слова в тексте».

2. *«Ожирение» у ключевых слов.* Развитие проблемы: если теперь важен вес слова — его нужно увеличить. Очевидно, что если наводнить сам текст повторяющимися словами, то текст, скорее всего, просто потеряет смысл. Поэтому дальнейшие усилия спамеров были направлены на то, чтобы системы видели и ранжировали один текст, а пользователь видел другой.

Здесь было «выработано» три способа: а) скрытие текста; б) перенаправление посетителей (переадресация); в) замена текста.

Рассмотрим каждый из этих способов.

а) *Скрытие текста.* Трудно прочитать белый текст на белом фоне, однако, он вполне доступен для поискового робота. Таким образом, посетители видели нормальный текст, а роботу для индексирования предоставлялся текст накачанный неверными ключевыми словами. Поисковики ответили на это фильтрами, которые научились различать цвет текста и цвет фона, и в случае совпадения — страница исключалась из базы данных.

Более разработанный вариант этого приема: просто очень мелкий текст (зачастую на маскирующей цветовой гамме: светло-серым по белому, различные «полосатые» фоны и т.п.).

На сегодняшний день появилась технология CSS (каскадные таблицы стилей). С помощью этой технологии можно, в частности, управлять расположением какого угодно объекта на экране. Таким образом, можно обойти любой автофильтр и разместить сколько угодно скрытого текста на экране. На помощь поисковикам пришли сами спамеры — наличие на сайте текста, не предназначенного для посетителей, автоматически трактуется как «спамдексинг». И спамеры просто-напросто проверяют тех, кто вырывается вперед с помощью данного метода, и в случае нарушения информируют поисковую систему. В результате штрафник просто получает «бан» в поисковой системе — его навсегда исключают из базы данных, так что теперь он просто не сможет быть там найден.

б) *Переадресация.* Технология входных страниц или дорвеев (doorway). Одна из самых неприятных для поисковых систем. Следуя этому принципу, в поисковых системах продвигается специальная страница — дорвей. А когда, следуя по ссылке из поисковика, прийдет пользователь — он будет перенаправлен на целевую (рекламную) страницу.

Понятно, что для целевой страницы в Интернете можно создать и независимо продвигать неограниченное число дорвеев. В результате поисковые системы попадают в ловушку — они вынуждены разбирать и индексировать просто кучу мусора, которая не имеет ничего общего с реальными текстами. Ловушка здесь в самой системе автоматического поиска — технология дорвеев заставляет поисковые системы искать в специально

созданном для них мусоре. Как наполняется этот мусор — очевидно.

Кроме того, дорвей, как правило, располагается не на основном (рекламируемом) сайте, поэтому наказать спамера и удалить целевой сайт из индекса модератор не может. Он может только «забанить» страницу дорвея. Но на смену им приходят все новые. Более того — несложно написать программу, которая будет генерировать эти дорвеи десятками тысяч.

Бороться с таким приемом сложно. Через автоматический фильтр можно «поймать» только некоторые варианты переадресации. Однако и таким фильтрам спамеры противопоставили простейший прием: вместо автоматической переадресации пользователю показывается страница с единственной надписью «Вход на сайт», так что он волен выбрать сам — перейти по ссылке или уйти. Часто идут дальше по ссылке, так что метод действует.

Последний вариант уже гораздо лучше — пользователь по крайней мере проинформирован, что скорее всего там он не найдет нужного текста. Однако, тем не менее, все равно поиск дал отрицательный результат — релевантный документ найден не был. Эта проблема актуальна и по сей день — очередной найденный метод переадресации вызывает массовое появление дорвеев. Радикального пресечения данного метода не найдено.

в) *Замена текста.* Следующие два метода, носящие название «клоакинг» и «своп» — это техника замены содержимого проиндексированной страницы.

Клоакинг (cloaking — маскировка, сокрытие) — методика распознавания визита поискового робота на страницу. Так что, когда на страницу приходит робот, ему показывается страница, отличная от той, что предоставляется пользователю. Необходимо понять, что отличие, например, от метода сокрытия текста на экране, заключается в том, что здесь подменяется сам HTML-код. Так что теперь поисковой системе не могут помочь даже конкуренты-спамеры — они не могут увидеть, что на самом деле ранжируется поисковиком.

Своп (swap — замена, обмен) — это более простая техника. Иное ее название bait-and-switch (можно перевести как «наживи

и подсекай»). Смысл ее заключается в том, что робот посещает страницу и отслеживает изменения раз в несколько недель. Прием заключается в том, что содержимое страницы подменяется на ожидаемое время посещения поискового робота, затем меняется обратно на корректное. Чаще всего это делается автоматически, так что у конкурентов не так много шансов «поймать» этого нарушителя и способ себя окупает.

Классификация существующих поисковых систем

На данный момент поисковые системы по методу осуществления поиска можно разделить на:

- поиск среди тематических каталогов;
- поиск по ключевым словам;
- поиск по вопросу на естественном языке;
- поиск по образцу текста;
- поиск с использованием тезауруса;
- поиск по сценарию;
- смешанные механизмы поиска.

Итак, более подробно о каждом пункте.

Поиск среди тематических каталогов.

Примеры: Yahoo, Rambler.

Основные категории пользователей: начинающие пользователи; пользователи с типичными информационными потребностями.

Плюсы: легкость выбора тем посредством меню; хорошая релевантность, так как человек (модератор) следит за качеством информации и человек распределяет информацию по тематическим каталогам; простота реализации.

Минусы: жесткая схема выбора; неполнота, как следствие предыдущего недостатка.

Перспективы: автоматическое наращивание тематических каталогов за счет привязки их к тезаурусу; автоматическая классификация; возможности задания пользователем своих рубрик (настройка на пользователя; настройка может осуществляться по образцу текста/сайта, по результатам предварительных запросов по списку ключевых слов/тем, сужением темы и т.д.).

Поиск по ключевым словам

Примеры: Гугл, Яндекс.

Основные категории пользователей: продвинутые пользователи; пользователи с разнообразными информационными потребностями.

Плюсы: гибкость; возможность индексации текстов по ключевым словам и словосочетаниям; сравнительная простота реализации.

Минусы: проблемы многозначности и полноты: требуется какая-то минимальная подготовка пользователя для выбора представительных ключевых слов, и, тем более, словосочетаний; трудность в решении задачи распознавания конечной цели пользователя по ключевым словам.

Перспективы: совмещение с другими механизмами (тематические каталоги, тезаурусы, сценарии); включение механизмов разрешения многозначности; включение механизмов разрешения неполноты (тезаурусы); морфологический анализ запроса; возможности настройки на пользователя (фиксирование тематических и коммуникативных потребностей и задание их по умолчанию).

(Примечание. Дополнительные возможности включают: И/ИЛИ/НЕ-запросы; повторный запрос в результатах предыдущего поиска; запрос по устойчивым словосочетаниям; запрос по однокоренным словам.)

Поиск по вопросу на естественном языке

Примеры: примеры реализации поиска к текстовым базам данных по ЕЯ-запросу на английском языке можно найти в поисковых машинах Ask Jeeve, Start, в энциклопедии Britannica.

Основные категории пользователей: начинающие пользователи; пользователи со сложными или специфичными информационными потребностями.

Плюсы: создается иллюзия диалога на естественном языке; пользователь определяет конечные цели поиска в явном виде.

Минусы: необходимо реализовать анализ ЕЯ-текста, частичный или полный; сложности обработки косвенных вопросов; значительные трудности, возникающие при попытках организации уточняющего диалога.

Перспективы: развитие этого способа можно представить в виде последовательных этапов:

1 этап: выявление ключевых слов и переход к запросам по ключевым словам,

2 этап: выявление ключевых слов и переход к сценариям,

3 этап: выявление ключевых слов и конечных целей и переход к сценариям,

4 этап: выявление ключевых слов и конечных целей и переход к уточняющему диалогу.

Поиск по образцу текста.

Примеры: поиск по образцу применяется в поисковых машинах Ramble, Infoseek.

Основные категории пользователей: начинающие пользователи; пользователи с плохо формулируемыми информационными потребностями.

Плюсы: более полный охват информационных потребностей пользователя, чем в запросе по ключевым словам или по запросу на естественном языке; больше средств тематической фиксации.

Минусы: трудности выявления доминантной потребности при их многообразии; те же проблемы многозначности и полноты; те же проблемы формулирования конечных целей; новые проблемы синтаксического анализа и понимания текста.

Перспективы: совмещение с другими механизмами (тематические каталоги, тезаурусы, сценарии); включение механизмов разрешения многозначности; включение механизмов разрешения неполноты (тезаурусы); морфологический анализ; возможности настройки на пользователя (фиксирование тематических потребностей и конечных целей и задание их по умолчанию); синтаксический анализ, выявление коммуникативной направленности текста (какой главный вопрос?)

Поиск с использование тезауруса.

Примеры: с переменным успехом был в разное время реализован на нескольких поисковых машинах (Infossek, AltaVista, Yahoo). Как разновидность этого поиска можно рассматривать ограничение запроса по ключевым словам тематическим каталогом.

Основные категории пользователей: пользователи со сложными или специфичными информационными потребностями.

Плюсы: в зависимости от заданного диапазона можно использовать универсально:

1) как средство увеличения полноты поиска,

2) как средство разрешения многозначности.

Минусы: требует дополнительной эрудиции пользователя; не всегда эффективен по времени и качеству.

Перспективы: автоматическое формирование тезаурусов, тематических каталогов; более жесткий отбор сайтов/страниц на основе фильтрации семантических связей.

(Примечание. Практически не применяется в самостоятельном режиме. Используется как дополнительная поддержка при поиске по ключевым словам для расширения фронта поиска или разрешения многозначности.)

Поиск по сценарию.

Примеры: в поисковых машинах используется сравнительно мало, чаще всего при организации ввода и выбора фактографических данных.

Основные категории пользователей: начинающие пользователи; пользователи со сложными или специфичными информационными потребностями.

Плюсы: можно очень точно выявить информационные потребности пользователя; можно хорошо настроиться на информационные потребности пользователя, создать его профиль; замедление работы компенсируется качеством поиска.

Минусы: необходимо привыкание пользователей к интерфейсу; кажущаяся громоздкость; замедление работы.

Перспективы: гибкие сценарии, обеспечивающие быстрое переключение между режимами; необходимо повысить эффективность поиска настолько, что это полностью компенсирует замедление работы; совмещение всех механизмов (рубрикаторы, тезаурусы, сценарии); включение механизмов разрешения многозначности; включение механизмов разрешения неполноты (тезаурусы, морфологический анализ); возможности настройки на пользователя (фиксирование тематических потребностей и конечных целей и задание их по умолчанию).

Смешанные механизмы поиска.

Примеры: примеры применения смешанных механизмов поиска можно найти на сайтах поисковых машин Infoseek, Altavista. В основном это сочетание поиска по ключевым словам с поддержкой поиска по тезаурусу.

Основные категории пользователей: все пользователи.

Плюсы: максимальный охват всех категорий пользователя, информационных и коммуникативных потребностей; гибкость.

Минусы: необходимо привыкание к интерфейсу.

Перспективы: см. по категориям.

(Примечание. Смешанные механизмы поиска можно рассматривать как специфичную форму сценария. В этом случае ввод первоначального запроса по ключевым словам активизирует набор альтернатив, которые можно рассматривать как ветви сценария.)

Описание созданной системы

Как уже было сказано выше после анализа текущих проблем поисковых систем было решено попытаться преодолеть проблему неполноты поиска, варьируя некоторые выделенные слова в запросе.

В данный момент поисковой системой Гугл поддерживается веб–сервис, позволяющий делать запросы к данной системе автоматически. Поэтому мета–поисковик строился именно для этой системы.

В качестве словаря морфем был выбран COM–объект представляющий интерфейс, который позволяет автоматически получать морфемы некоторого слова русского языка (http://www.aot.ru/download/RusMorph.zip).

Теперь стоит задача разбора строки запроса пользователя и вариации выделенного слова по различным морфемам. А затем формирование ряда запросов с вставкой этих морфем вместо выделенного начального слова.

Понятно, что существуют слова, для которых морфологические формы совпадают.

ПРИМЕР. *Нормальная форма слова «кошка». В винительном падеже множественного числа морфема будет «кошек», в родительном множественного числа морфема будет такой же.*

Так как для системы Гугл нет возможности указать, какая именно форма имеется ввиду, то подобные повторения были просто исключены из запроса.

Следующим шагом был разбор полученных результатов. После того, как были сделаны запросы с каждой уникальной морфемой выделенного слова, программа имеет массив результатов, которые можно вывести на экран. Но учтем, что в среднем каждое слово русского языка имеет по 8–12 морфем, а для однословных запросов Гугл имеет гораздо больше десяти результатов для вывода. Так что система в среднем имеет по 80–120 результатов, пригодных для вывода. Это довольно много и хотелось бы как–то ранжировать подобный вывод, чтобы более релевантные результаты отображались в первую очередь.

Эта проблема была решена следующим образом.Вполне возможно, что в некотором тексте могут встречаться различные морфемы одного и того же слова, и ссылка на этот текст может быть продублирована в запросе. Также, очевидно, что подобные продублированные ссылки должны быть расположены выше остальных результатов. Но нужно также учесть, какое положение эти ссылки занимали в результатах Гугл.

То есть, рассмотрим следующий случай. Некоторая страница была на первом месте в рейтинге Гугл по данному запросу, но выбелась один раз, а также есть еще одна страница, которая появилась в результатах два раза, (она «откликнулась» на две различные морфемы), но оба раза она расположена на десятом месте по рейтингу Гугл. Вопрос: какая страница должна отображаться выше уже в рейтинге созданной системы? И как ранжировать множество всевозможных других вариантов? Для подобных расчетов очень бы пригодилось реальное значение рейтинга Гугл, чтобы оценка была более объективной. То есть, даже если использовать формулу ранжирования подобных результатов, то все равно сравниваются результаты похожих запросов, но все же не одних и тех же. И на один запрос система суммарно получает одно количество страниц, на другой — другое. И при этом каждая страница, попадая в десятку, «выигрывает» разную конкуренцию относительно других странице (т.е. на один запрос на первом месте может оказаться страница, у которой

рейтинг Гугла ниже, чем на 10 месте у другого запроса). Однако подобной информацией система обладать не может (Гугл, конечно же, скрывает, какой рейтинг у него имеет по данным ключевым словам та или иная страница). Кроме того, на данный момент нет алгоритма, который был бы в состоянии определить количественно, что важнее при сравнении страниц из различных запросов: то, как они были отсортированы системой Гугл, либо повторения в различных видах запроса.

Вернемся к формуле ранжирования результатов, из которой будет видно, как ей были учтены, либо проигнорированы перечисленные факторы. Но прежде чем эта формула будет представлена, необходимо указать, что решение о правильности назначения веса различным факторам может быть принято при большой выборке с оценкой количества релевантных результатов для каждой системы весов.

Но, если говорить о релевантности, то необходимо еще сделать и проверку релевантности ссылок. Однако это тот момент, когда нужно преодолеть три пункта: 1) выборка должна быть действительно большой; 2) оценка не должна быть субъективной; 3) авторитет эксперта должен быть подтвержден.

Так что оценка релевантности не могла быть сделана автором из-за первого пункта по причине ограниченности ресурсов. Большой выборки можно достигнуть благодаря автоматизации оценки релевантности, но даже если забыть то, что фактически это проблема семантического анализа текста на естественном языке, то все равно нужны серьезные доказательства, что мнение данной автоматической системы объективны. Таким образом, коэффициенты получены из эмпирических соображений.

Итак, внутри созданной системы каждой уникальной ссылке присваивается следующий вес:

$$\textbf{PagePound} = \textbf{Summ of Pound(i)},$$

где **Pound(i)** — вес ссылки при запросе с i-й уникальной морфемой, который рассчитывается следующим образом:

$$\textbf{Pound(i)} = \textbf{CountImpact} + \frac{(\textbf{perPageResult} - \textbf{position} + 1)}{\textbf{PositionImpact}},$$

где

PoundImpact — коэффициент, от которого зависит влияние позиции в десятке от Гугл. Изменяя его, можно влиять на порядок результатов в нашем выводе. В созданной системе он равен 20;

PositionImpact — коэффициент влияния повторов в результатах, равен 1;

perPageResult — количество результатов для каждого запроса, равен 10.

После разбора результатов и присвоения им различных весов (для этого создается класс, который хранит структуру внутреннего представления) массив оцененных результатов сортируется с помощью пузырькового метода сортировки. Надо отметить, что на таком объеме реализовывать более сложный (но и более быстрый) алгоритм сортировки автору не представляется необходимым. Если в дальнейшем удастся повысить объем результатов, необходимых для сортировки (не очень большое количество результатов связано с ограничениями на использование web-service'а и об этих ограничениях будет упомянуто), или система должна будет удовлетворять запросам более чем одного исследователя (так что станет критичным вопрос загрузки ресурсов), то этот метод сортировки может быть с легкостью заменен на любой другой.

Итак, теперь система может вывести результат.

В выводе участвуют:

- все запросы, которые были отправлены системе Гугл,
- суммарное количество страниц по каждому запросу, которые Гугл посчитал релевантными данному запросу,
- суммарное количество релевантных страниц по всем запросам, отправленным к системе Гугл,
- количество повторов данной ссылки в необработанных результатах,
- вес, который данная ссылка получила в созданной системе,
- в поле Snippet записывается значение snippet для каждого повторения ссылки в первоначальном результате,
- в конце вывода результатов выводится количество уникальных ссылок в конечном результате работы системы.

Кроме того, справа от данных результатов выводятся результаты работы системы Яндекс. Зачем это нужно — в данной системе морфемы одного и того же русского слова воспринимаются как единая логическая единица поиска. Так что подобное сравнение работы созданной системы и поисковой системы Яндекс было просто необходимо.

(Кратко об этом выводе. У Яндекса тоже создана система похожая на web-service. Этой системе можно задавать описанный Яндексом XML-запрос и получать XML-ответ. Этот ответ обрабатывается и выводится на экран с помощью шаблонов преобразований, расположенных в XSLT-файле.)

Анализ результатов и выводы

Прежде всего, автор хотел бы привести результат, из которого можно сделать вывод, что поисковая система Гугл действительно воспринимает и индексирует различные морфемы одного и того же слова русского языка как совершенно разные слова.

ПРИМЕР. В строку поиска вводится все то же слово «кошка». «Кошка» имеет 10 уникальных морфем: кошка, кошки, кошке,кошку, кошкой, кошкою, кошек, кошкам, кошками, кошках. Таким образом, системе Гугл посылается 10 запросов и система получает 100 ответов (данное слово не настолько редкое, чтобы хоть по одной форме мы получили меньше десяти ответов, т.е. в системе Гугл нашлось бы менее десяти результатов по какой-либо морфеме). Система информирует, что из ста ссылок, которые система получила на десять запросов, уникальными оказались 98. Что в свою очередь означает, что для Гугл десять морфем слова «кошка» являются самостоятельными независимыми словами и именно поэтому мы получили такой большой процент уникальных ссылок.

Подберем слова (см. приведенную ниже таблицу) из различных предметных областей, чтобы показать, что этот пример не является вырожденным.

Посмотрим на результаты при увеличении количества ключевых слов.

Слово	Кол-во уникальных морфем	Кол-во уникальных ссылок
молоко	5	50
компьютер	10	99
солнце	8	80
деревня	10	100

Рассмотрим запрос «молоко матерь». Выделенным ключевым словом, подлежащим варьированию по морфемам, будет «матерь». У данного слова семь морфем: матерь, матери, матерью, матерей, матерям, матерями, матерях. Для данного вида запросов мы все еще получаем большое количество уникальных ссылок. В частности, для этого запроса из 70 полученных ссылок уникальными оказались 68. В целом для некоторых слов цифры оказываются иными (это будет пояснено в дальнейшем на примере более длинных запросов). Однако отношение количества уникальных ссылок к количеству полученных в ответ на ряд запросов является близким к отношению, полученному в данном примере.

Необходимо отметить, что уже на запросах из двух слов можно заметить, что мы получаем более полные результаты на введенный запрос, таким образом, избавляя пользователя от перебора всех морфем данного редкого слова. А, следовательно, решена одна из проблем из постановки задачи — решение неполноты поиска с точки зрения вариации выделенных ключевых слов по морфемам.

Однако автор предлагает ознакомиться с результатами, полученными при дальнейшем увеличении количества ключевых слов.

В случае запроса из трех слов «молоко дети матерь» (ключевым по прежнему является слово «матерь») система выдает следующие результаты: на первом месте значится ссылка с количеством повторов 6 и внутренним весом в системе 8,05; на втором месте ссылка со значениями 5 и 6,75; на третьем — 3 и 3,75.

Видно, что веса стали увеличиваться в результате того, что в результатах стали появляться повторы. И страницы, ссылки на которые повторялись, поднялись вверх в результатах.

Ссылки начали склеиваться потому, что увеличив количество ключевых слов, мы конкретизировали запрос и уточнили предметную область, которая нас интересует. Таким образом, уменьшилось количество страниц, которые система Гугл могла счесть релевантными запросу. И именно поэтому в результатах появилось больше повторов.

Прежде чем перейти к представлению результатов, полученных при большем количестве ключевых слов, автор хочет провести сравнение между поисковыми системами Гугл и Яндекс, так как примерно на этом количестве ключевых слов начинают быть заметны результаты, которые могут быть оценены аналитически.

Однако прежде чем приступить к формулировке неких результатов, автор считает необходимым привести еще раз доводы того, что адекватно оценивать какая поисковая система Интернета дает более релевантные результаты на данный момент невозможно. Человек сталкивается с проблемой того, что анализ должен быть проведен на огромной выборке — т.е. придется разбирать просто груды материала. Кроме того, неочевидны критерии, по которым следует оценивать данную релевантность. Машина же на данный момент не в состоянии делать семантический анализ текста.

Поэтому автор предлагает ознакомиться с количеством уникальных ссылок, которые выдает каждая система на одинаковые запросы.

Итак, на указанный выше запрос «молоко дети матерь» поисковая система Яндекс выдала 7576 уникальных ссылок.

В созданной системе суммарное количество уникальных результатов было 19139. Однако данные цифры ни о чем говорить не могут. Кроме того, данный результат вовсе является бесполезным — в большинстве случаев пользователь не просматривает более десяти первых результатов.

Цифры, которые будут являться неким критерием, получаются при увеличении количества ключевых слов.

Увеличим количество ключевых слов до пяти. В качестве примера рассматривается запрос «"сетевое планирование" время выполнения проекта». (В случае, когда необходимо, чтобы словосочетание воспринималось как одна логическая единица за-

проса — оно заключается в кавычки.) Выделенным ключевым словом будет слово «проекта». В результате ряда запросов получаем 1094 ссылок, доступных для просмотра. Поисковая система Яндекс на тот же самый запрос выдает только три уникальных ссылки.

Приведем результаты запроса, который может пояснить полученный результат. Итак, запрос: «"Институт систем информатики им. А.П.Ершова"». Данный длинный запрос без изменений посылается и системе Гугл, и системе Яндекс.

Результаты: Гугл нашел 547 страниц по этому запросу. Высшая ссылка в рейтинге Гугл ссылается на сайт института. Остальные ссылаются либо на другие страницы сайта института, либо на тексты, где встречается полное название института. Поисковая система Яндекс не нашла по данному запросу ничего. Таким образом, алгоритмы, реализованные в системе Яндекс, на данный момент являются объективно слабее алгоритмов, используемых поисковой системой Гугл. Следовательно, используя Гугл, созданная система может выигрывать по количеству результатов у поисковой системы Яндекс. Таким образом, созданная система отличается большей полнотой поиска и чем система Яндекс, и чем система Гугл.

Теперь вниманию читателя предлагается следующее замеченное свойство созданной системы. При формировании большого запроса (5–6 ключевых слов) опытный пользователь может правильно подобрать слова для вариации по морфемам. Тогда при ранжировании результатов большой вес получают (а соответственно занимают первые строчки в результатах и предоставляются вниманию пользователя в первую очередь) ссылки на тексты, содержащие выделенные слова в различных морфологических формах. На данный момент структура Интернета и обрабатывающая его поисковая система Гугл таковы, что мы получаем в качестве первых результатов ссылки на статьи/доклады, содержащие в большом количестве ключевые слова, которые задают тематику запроса, а также содержащие различные морфемы выделенных ключевых слов. Это обеспечивает не просто более богатое присутствие этих слов в полученном тексте, а также более богатое присутствие вариаций этих слов. Что в свою очередь

позволяет сделать предположение, что данный текст посвящен именно выделенным ключевым словам в тематике, которую задают остальные ключевые слова.

Помимо этого, зачастую в первые строчки поднимаются ссылки, ведущие на обсуждение данной тематики на различных форумах. Зачастую именно на этих форумах можно найти прямые ссылки с весьма высокой долей релевантности конечным целям пользователя, и/или просто задать вопрос пользователям, принимающим участие в обсуждении данной тематической ветки форума.

Таким образом, для достижения подобных результатов автор предлагает следующий подход к построению набора ключевых слов.

Ключевые слова разделяются на слова, задающие тематику поиска — второстепенные ключевые слова. А также на слова, которые являются главными в поиске — выделенные ключевые слова. Выделенные ключевые слова будут варьироваться по морфемам и, с учетом сказанного выше, пользователь зачастую будет получать результаты, которые будут удовлетворять его требованиям.

В завершении разбора результатов автор хочет обратить внимание читателя на следующий факт: при подходе к обработке запроса в созданной системе пользователь получает возможность варьировать ключевые слова в составе устойчивых словосочетаний, выделенных кавычками. При обычном подходе данное словосочетание вообще удерживается от вариаций и иногда это полезно. Однако иногда необходимо различные морфологические формы именно этого словосочетания. Но написать такой запрос современной поисковой системе Интернета невозможно. И как уже было сказано это реально при варьировании выделенных ключевых слов.

Заключение

• Создана система, использующая две из наиболее мощных современных поисковых систем Гугл и Яндекс, а также словарь морфем русского языка. Функциональность созданной системы состоит в следующем:

1) на базе введенного пользователем запроса создается ряд уникальных подзапросов;

2) подзапросы отсылаются в поисковую систему Гугл;

3) полученным результатам с помощью разработанного алгоритма назначаются веса;

4) результаты ранжируются согласно полученным весам;

5) первоначальный запрос пользователя в необработанном виде посылается в поисковую систему Яндекс;

6) результат работы поисковой системы Яндекс приводится в вид, пригодный для вывода на экран;

7) результат работы поисковой машины автора и поисковой системы Яндекс выводятся на экран для сравнения.

• Данная система позволяет осуществлять более полный поиск по русскоязычной части Интернета по сравнению с оригинальной поисковой системой Гугл, а также поисковой системой Яндекс, в которой проблема неполноты, в частности, также решалась с помощью использования морфем русского языка.

• Архитектура системы позволяет расширить функциональность. Например, планируется расширение функциональности до анализа запроса на естественном языке (или запроса на естественном языке специального вида) и формирование ряда запросов поисковой системы Гугл на базе этого анализа.

• Поставлены эксперименты для различных запросов при увеличении количества ключевых слов.

• Сделаны выводы из результатов эксперимента, а также предложен метод поиска с использованием специфичного построения запроса для созданной системы. (Данный пункт более полно раскрыт выше при анализе результатов.)

В качестве дальнейших перспектив развития данной системы автор рассматривает

• Внедрение в систему онтологий. В частности планируется использование онтологий для уточнения запроса и выбора предметной области в самом начале поиска. Кроме того, возможно добавление модуля, позволяющего производить поиск по образцу текста с использованием онтологий. В данный момент работа ведется только с сигнатурами (наборами понятий).

• Развитие использования электронных словарей. Например, анализ структуры запроса сформированном в виде предложения на естественном языке. А затем перевод этого запроса в запрос или серию запросов, понятных поисковой системе Гугл.

• Развитие или разработка собственного морфологического словаря.Так как уже используемый словарь обладает небольшим недостатком: очень редко, но все же этот словарь выдает морфемы, которые не являются верными. (Например, вот набор морфем, который получен при запросе «дети»: дети, детить, дечу, детим, детишь, детите, детит, детят, детил, детила, детило, детили, детив, детивши, детимте, детивший, детившего, детившему, детившим, детившем, детившая, детившей, детившую, детившею, детившее, детившие, детивших, детившими, деченный, деченного, деченному, деченным, деченном, дечен, деченная, деченной, деченную, деченною, дечена, деченное, дечено, деченные, деченных, деченными, дечены.)

• Возможность временного отказа от использования поисковой системы Гугл и создание собственной поисковой системы, основанной на тех же принципах. Созданная система будет анализировать утвержденный экспертами объем текстов (около 100 000), на которых уже можно оценивать количественно релевантность запроса. При получении удовлетворительных результатов можно будет вернуться к использованию поисковой системы Гугл, и все же попытаться оценить релевантность результатов отлаженной системы.

Литература

1. ПАЛЬЧУНОВ Д.Е. Алгебраическое описание смысла высказываний естественного языка //Модели когнитивных процессов. — Новосибирск, 1997. — Вып. 158: Вычислительные системы. — С. 127–148.

2. ПАЛЬЧУНОВ Д.Е. Синтаксическая близость предложений языка первого порядка //Измерение и модели когнитивных процессов. — Новосибирск, 1998. — Вып. 162: Вычислительные системы. — С. 58–80.

3. PAL'CHUNOV D.E. Algebraische Beschreibung der Bedeutung von Aeusserungen der natuerlichen Sprache //Zelger, Josef/

Maier, Martin (1999, Hrsg.): GABEK. Verarbeitung und Darstellung von Wissen. Innsbruck–Wien: STUDIENVerlag. — P. 310–326.

4. PAL'CHUNOV D.E. On a logical analysis of GABEK. // GABEK II. Zur Qualitativen Forschung On Qualitative Research. STUDIENVerlag, Innsbruck–Wien–Munchen. —2000. — P. 185–203.

5. PAL'CHUNOV D.E. Logical Methods of Ontology Generation with the Help of GABEK //IV International GABEK Symposium, 2002. — Innsbruck. — P. 17.

6. [Survey] Survey of the State of the Art in Human Language Technology /Cole, Ronald, et al (eds.) Studies in Natural Language Processing. Cambridge University Press. — 1998. — 533 p.

7. [Reports] Search Engine Status Reports. //http://www.searchenginewatch.com/ reports/index.html. — 1999.

8. ПОЛЯКОВ В.Н., ШОХИН Д.В. Использование лингвистических технологий для сбора и анализа научных данных в компьютерной сети Интернет //Обработка текста и когнитивные технологии. — Вып. 2: — М., Пущино: ОНТИ ПНЦ РАН. — 1999.

9. ПОЛЯКОВ В.Н. Интеллектуальная поисковая машина. Ч.1. - М.,2000. Электронная версия: http://www.geocities.com/ SiliconValley/Campus/7926/ Polykov/IntelSE.htm#z5

10. ДИЛАН ТВИН. Поиск — мое ремесло //Мир персональных компьютеров. — 1997, январь. — С. 114–123.

Поступила в редакцию
19 августа 2004 года