5. *Kleinberg, J.* The small-world phenomenon: An algorithmic perspective // Proc 32nd Symposium on Theory of Computing, 2000.
6. *Reka, A., Jeong, H., and Barabsi, A.* Diameter of the World Wide Web // *Nature, Vol.* 401, 1999. Pp. 130–131.

Е.Е. Краснощеков

ПОИСК РЕЛЕВАНТНЫХ ДОКУМЕНТОВ В ЭКОНОМИЧЕСКИХ СИСТЕМАХ ПО ЗАПРОСУ, РАСШИРЕННОМУ ПАРАДИГМАТИЧЕСКИМИ ОТНОШЕНИЯМИ

В статье рассмотрены технология адаптивного распознавания образов APRP и общие подходы к организации оптимального поиска с использованием нечеткого сравнения. Проанализированы существующие подходы к решению задач поиска и предложено их улучшение. Рассмотрено применение при поиске автоматически построенных ассоциативных отношений вместо парадигматических отношений, построенных вручную. Проведена экспериментальная оценка метода поиска по расширенному запросу.

Введение

Объем электронных документальных баз данных в экономических системах за последние годы значительно возрос. В настоящее время значительная часть экономических документов доступна в глобальной сети Интернет, которая предоставляет не только почти неограниченные возможности для размещения данных, но и возможности для доступа к документам из любой точки земного шара при условии наличия компьютера, подключенного к «всемирной паутине».

Количество электронных документов в сети Интернет настолько велико, что обнаружить необходимый ресурс без «путеводителя» почти невозможно. Сегодня роль путеводителя и справочника в Интернете играют поисковые системы.

В подобной ситуации резко возросла потребность в системах поиска и анализа данных. Именно поисковая система определяет, превратятся ли в знания многочисленные разрозненные данные, поступающие по различным каналам связи и накапливаемые в разнообразных государственных, ведомственных, частных и прочих электронных архивах.

Поиск документов можно отнести к наиболее важным задачам содержательной обработки текстовой информации, что, в частности, обусловлено потребностью поиска информации в сети Интернет.

Разработка методов текстового поиска имеет давнюю историю, насчитывающую более сорока лет [1]. За это время информационно-поисковые системы эволюционировали от систем формально-логического типа [2] к системам нечеткого поиска, основными особенностями которых являются следующие [3]:

- запрос задается на естественном языке, а не в виде формального выражения булево-контекстного типа;
- некоторые или даже все найденные документы содержат только часть информативных слов запроса;
- найденные документы выдаются в ранжированном виде, т. е. в порядке убывания их соответствия запросу.

Нечеткий поиск используется во многих поисковых машинах Интернет (AltaVista, Excite, Lycos и др.).

Применение нечеткого подхода повышает эффективность поиска, но недостаточно. На малых объемах текстов по-прежнему часто находится мало релевантных документов. При больших объемах текстовой базы релевантных документов более чем достаточно, однако качество ранжирования может оказаться неудовлетворительным.

Одним из подходов к устранению упомянутых недостатков является добавление к запросу слов, связанных парадигматическими отношениями со словами запроса, и последующий поиск по расширенному запросу. Этот метод расширения запроса рассматривается в настоящей статье.

Механизм адаптивного распознавания образов APRP

Механизм адаптивного распознавания образов APRP (Adaptive Pattern Recognition Processing) зародился в процессе исследований в области моделирования сложных биологических систем. Механизм, в котором для обработки информации используются нейронные сети, действует как самоорганизующаяся система, автоматически формирующая и индексирующая двоичные образы документов. Данная технология обеспечивает поддержку нечеткого поиска информации, его высокую точность и полноту, языковую независимость и малый объем индексных файлов.

Вот основные преимущества технологии APRP для контекстного поиска текста:

- нечеткий поиск;
- автоматическая индексация;
- достоверность при сложной индексации;
- высокая скорость поиска информации.

В технологии APRP под нечетким поиском понимается возможность найти достаточно близкое приближение к запрошенному термину или фразе.

APRP работает не с ключевыми словами, а с образами, две–три ошибочные буквы в слове или фразе не могут существенно изменить базовую картину текста. Таким образом, автоматически становится допустимой ошибка как во входных данных, так и в терминах запроса. APRP всегда в состоянии найти ближайшее приближение к терминам и фразам, заданным в качестве объектов поиска.

Нечеткий поиск особенно полезен в ситуациях, когда ввод данных осуществляется с помощью оптического распознавания символов, так как процесс OCR, как уже говорилось, не является на 100 % точным даже при очень высоком качестве печати. Например, если на данной странице с помощью OCR не удалось абсолютно правильно считать ни одного слова, практически никакая система четкого поиска не имеет шансов добиться успеха при поиске этой страницы.

При использовании APRP вы можете проиндексировать все свои данные, не указывая ключевых слов или полей базы данных, не привлекая администратора базы данных и не прибегая к экспертам для определения значимости того или иного слова или фразы по сравнению с другими словами или фразами.

Так как индексируемые образы не задаются заранее пользователем или программно, а выбираются самой нейронной сетью, каждая нейронная сеть становится ассоциативным запоминающим устройством, оптимизированным для актуального текста в базе данных. Введя документ, вы переключаете систему в режим «изучения». В этом режиме она просматривает двоичные образы и определяет, где они находятся, как в виртуальной, так и в физической памяти, с использованием алгоритмов на нейронных сетях. «Изучив» документы в процессе индексации,

система осуществляет поиск информации очень быстро, используя для этого процедуры нечеткой логики.

Нейронную сеть можно оптимизировать как для языковых образов (английский, французский, японский и т. п.), так и для профессиональной лексики (юридической, социологической, химической и т. п.). Система APRP динамически определяет и выделяет образы, которые могут представлять собой все, что угодно – от простой текстовой строки, например, сочетания «акция» (найденное как отдельное слово или фрагмент слова), до сложных фраз типа «принимая во внимание согласие между фрагментом первой части и фрагментом второй части...».

Гибкость методологии поиска APRP позволяет улучшить параметры процесса поиска данных, позволяя пользователю самому определять степень совпадения найденной информации с запросом. Можно сформулировать эффективный запрос без знания правильного написания слов или фраз. Получив запрос найти какой-либо документ, система просматривает образы и составляет список «ближайших приближений» к тому, что было описано в запросе. Затем система упорядочивает содержимое этой области по степени вероятности того, что тот или иной найденный на этом этапе документ является истинной целью поиска. Можно установить «ближайшую десятку», «ближайшую сотню» и т. д. Это потенциально создает среду поиска, в которой пользователь может проводить поиск в интерактивном режиме, чтобы найти ответ, предварительно не определив точно, что же является ответом.

Программные системы, базирующиеся на методологии APRP, имеют возможность динамически использовать ресурсы и архитектуру компьютера для получения более быстрого и точного доступа к информации. Поскольку индекс занимает минимальный объем, его можно мгновенно загрузить в память любого компьютера в сети и работать с ним со скоростью двоичных логических операций.

Такое свойство самооптимизации предполагает использование всех доступных ресурсов системы – память, диски и т. д. На большинстве рабочих станций APRP позволяет осуществлять поиск в объеме 200 000 страниц информации не более чем за десять секунд. Важнейшим преимуществом такого подхода является динамическая природа технологии оптимизации как конкретных данных, так и используемых аппаратных средств. По мере внедрения таких технологий, как параллельная обработка, повышающих мощность и совершенствующих архитектуру компьютеров, система APRP получит возможность функционировать на таких новых аппаратных платформах и автоматически использовать их ресурсы, значительно сокращая время реакции.

Информация любой природы представляется в компьютере одинаково – с помощью нулей и единиц. Это означает, что технология APRP может быть применена таким же образом для индексации и нечеткого поиска изображений, видео- и звукозаписей, сигналов, речи и всего разнообразия мультимедийной информации.

Компания Excalibur Technologies разработала библиотеки, реализующие нечеткий поиск информации различной природы:

- библиотека TRS (Text Recognition Software) предназначена для индексации и нечеткого поиска текстовой информации;
- библиотека SRC (Signal/Sound Recognition Software) предназначена для распознавания (индексации и нечеткого поиска) голосовой, звуковой и сигнальной информации;
- библиотека VRS (Visual Recognition Software) предназначена для индексации и нечеткого поиска изображений (например, поиск по фотографиям, отпечаткам пальцев и т. д.).

Описание метода расширения запроса

Парадигматические отношения (синонимия, родовидовые и пр.) между терминами, как правило, устанавливаются вручную. Информация об этих отношениях фиксируется в тезаурусе и используется в дальнейшем при поиске по запросам, что позволяет повысить полноту (найти больше релевантных документов).

Существуют автоматические методы построения парадигматических отношений. При этом выделенные парадигматические отношения не разбиваются на типы (например, не различаются синонимия и отношение «целое–часть»), и сами отношения обычно называются ассоциативными, а не парадигматическими.

Построение ассоциативных отношений основывается на данных о совместной встречаемости терминов в документах [4]. Автоматически построенные ассоциативные отношения могут применяться при поиске вместо вручную построенных парадигматических отношений либо в дополнение к ним. При этом эффективность поиска повышается только при использовании специальных алгоритмов ранжирования. В работе [6] показано, что максимальный эффект достигается в том случае, если к исходному запросу добавляются те ассоциативные термины, которые имеют высокое соответствие по отношению ко всему запросу, а не к его отдельным словам.

Недостатком ассоциативных отношений является относительно невысокая скорость их построения и требуемые при этом большие объемы текстов. Немаловажно также и то обстоятельство, что в разных предметных областях один и тот же термин может иметь разный набор ассоциативных отношений. Поэтому ассоциативные отношения имеет смысл выделять динамически в процессе поиска по запросу. Предположим, что среди найденных по запросу документов есть такие, в которых содержатся все информативные слова запроса, причем в компактном виде. Из таких документов выделяются информативные слова, находящиеся в окрестности появлений запроса. Каждому из выделенных слов присваивается ассоциативный вес. Основной компонентой ассоциативного веса слова является количество появлений запроса, в окрестности которых содержатся эти слова.

Далее выполняется нечеткий поиск по расширенному запросу, составленному из исходного запроса путем добавления к нему ассоциативных слов с максимальными весами. Расширенный запрос содержит все информативные слова исходного запроса и динамические ассоциативные слова, для которых ассоциативный вес больше порога. Количество добавленных к запросу ассоциативных слов зависит от числа информативных слов исходного запроса (например, при 5 информативных словах запроса отбирается не более 3 ассоциативных слов).

Списки документов, найденных по исходному и расширенному запросам, объединяются в единый список, который и считается результатом поиска. Если некоторый документ присутствует в двух списках, то его вес вычисляется по формуле сложения вероятностей.

В качестве результата поиска выдается начальная часть объединенного списка. Количество выдаваемых документов определяется как произведение числа документов, найденных по исходному запросу, на некоторый множитель, который имеет фиксированное значение или же задается в виде таблицы.

Можно ожидать, что расширение запроса ассоциативными словами поможет решить две задачи:

- выдачу дополнительных релевантных документов. Эта задача важна при поиске в небольших текстовых базах данных. Поскольку поиск нечеткий, то по расширенному запросу могут быть найдены документы, в которых интересующая тема выражена по-другому, причем с помощью включенных в запрос ассоциативных слов;

- построение тематического представления результатов поиска. Очень актуально при поиске в больших текстовых базах экономических систем. Если по запросу найдено много документов, то, как правило, они разбиваются на несколько групп тематически однородных документов, причем пользователя интересует только некоторые из этих тематических групп. Поэтому очень полезно разбиение найденных документов по тематикам. Такое разбиение значительно облегчает пользователю отбор нужных ему документов среди множества найденных. Если провести кластеризацию текстовых фрагментов с появлениями запроса, то это можно считать тематическим представлением, однако относительно ограниченного множества документов. Однако, если по каждой из групп фрагментов, полученных в результате кластеризации, построить расширенный запрос, то с помощью такого множества запросов будет реализовано полноценное тематическое представление всего множества найденных документов.

Экспериментальная оценка метода

Экспериментальная оценка поиска по расширенному запросу выполнялась на текстовой базе малого объема. В такой ситуации актуально повышение полноты, т. е. выдача дополнительных релевантных документов. Как показано ниже, расширение запроса ассоциативными отношениями в определенной степени решает данную задачу.

Для эксперимента использовалась программа «Следопыт» [5], разработанная компанией «МедиаЛингва». Эта программа реализует нечеткий поиск текстовой информации по запросу на естественном языке.

Следопыт ищет документы по их содержанию. Запрос на поиск задается в виде фразы на русском, английском или немецком языке. Допускаются и комбинированные запросы, состоящие из смеси русских, английских и немецких слов. Программа сама учитывает все формы слов запроса на основе использования бессловарной машинной морфологии и оценивает компактность их расположения в текстах найденных документов.

Найденные документы программа ранжирует в порядке уменьшения их соответствия теме запроса, то есть наиболее важным в большинстве случаев будет первый по порядку из найденных документов. Хорошее качество ранжирования достигается путем реализации следующих частных механизмов:

- учет только информативных слов запроса (не входящих в словарь неинформативных слов и выражений);
- учет статистики распределений слов запроса по документам, среди которых выполняется поиск;
- учет расстояния между словами запроса в документе;
- статистика полных и частичных вхождений запроса в документ;
- учет количества слов и их взаимной информативности в появлении запроса (полном или частичном);
- приближенный (без использования словарей) морфологический анализ русских и английских текстов с синонимией не только на уровне словоформ, но и на уровне словообразования (одинаковыми считаются слова «море» и «морской»).

Следопыт способен находить документы, в которых тема запроса выражена другими словами. Иначе говоря, в программе реализован не логический (на полное соответствие запросу), а смысловой (нечеткий) метод поиска текстов. Это очень существенно, поскольку человек хорошо запоминает смысл фразы, но с течением времени, как правило, не в состоянии воспроизвести ее дословно.

Качество поиска не зависит от лексики предметной области – Следопыт с одинаковой эффективностью производит поиск как по газетным или деловым тек-

стам, так и по узким тематическим направлениям типа глазных болезней или порошковой металлургии.

Основой реализованного в программе Следопыт семантического поиска является метод преобразования исходного естественно-языкового запроса в оптимальную булево-контекстную форму. В данном случае под оптимальностью формы понимается ее максимальная эффективность среди всевозможных булево-контекстных форм в смысле максимизации критерия, выраженного в виде степенной функции от полноты и точности поиска. Дополнительно учитываются все те же самые факторы, что и в случае ранжирования найденных документов.

Экспериментальная оценка проводилась путем анализа результатов поиска программой Следопыт по 8 запросам в массиве компьютерных текстов общим объемом 8 Мбайт.

Сравнивались 2 метода поиска:

- исходный поиск – нечеткий поиск Следопыта с параметром степень расширения запроса, равным 25%;
- поиск с учетом ассоциативных отношений (далее называем его ассоциативным поиском). Множитель, используемый для определения количества выдаваемых документов, взят равным 1,5.

Ассоциативные слова отбирались вручную среди слов, входящих во фрагменты найденных документов. Эти фрагменты (называем их далее релевантными фрагментами) удовлетворяли следующим условиям:

- содержали все слова запроса на небольшом (не более 5–6 слов) расстоянии друг от друга;
- включали по 5 слов слева и справа от появления запроса в документе.

Ассоциативный вес слов вычислялся приближенно с учетом только количества релевантных фрагментов, содержащих слово (чем больше таких фрагментов, тем выше вес). Можно предположить, что при более корректном вычислении ассоциативного веса эффективность поиска не ухудшится.

Отобранные вышеуказанным образом ассоциативные слова добавлялись к исходному запросу, и по этому расширенному запросу выполнялся поиск с помощью Следопыта.

Пример

Для запроса «настольная картографическая система» сформирован расширенный «запрос настольная картографическая система MapInfo».

По всем 8 запросам был проведен поиск и вычислены значения полноты П и точности Т поиска. Эти два параметра являются общепринятыми характеристиками эффективности поиска [4]. Определяются они по следующим формулам:

$$П = Nrf/Nr,$$

$$Т = Nrf/Nf,$$

где Nrf – количество релевантных документов среди документов, найденных по запросу;

Nr – общее количество содержащихся в базе данных документов, которые релевантны запросу. Поскольку определение полного числа релевантных документов требует больших затрат ручного труда, то в качестве оценки Nr принимаем полное число релевантных документов, найденных по запросу двумя сравниваемыми методами поиска. В результате получаем завышенное значение полноты, однако соотношения между значениями полноты при разных методах поиска будут те же самые, что и при корректном определении полноты;

Nf – количество документов, найденных по запросу (из них Nrf документов релевантны запросу).

Полнота и точность, полученные по отдельным запросам, усреднены и сведены в приведенную ниже таблицу:

	Исходный поиск	Ассоциативный поиск
Полнота	0,74	1,0
Точность	0,96	0,89

Отметим, что за счет ассоциативных отношений заметно выросла полнота поиска при относительно небольшом падении точности. Отсюда можно сделать вывод о целесообразности использования автоматически построенных ассоциативных отношений в системах с нечетким поиском.

БИБЛИОГРАФИЧЕСКИЙ СПИСОК

1. *Сэлтон Г.* Автоматическая обработка, хранение и поиск информации. – М.: Советское радио, 1973. – 560 с.
2. *Маркусова В.А.*, *Реброва М.П.*, *Страшко В.П.* Особенности интерактивного поиска проблемно-ориентированной информации в базе данных SCI-SEARCH. НТИ. Сер. 2, № 3, 1988. С. 26–30.
3. *Ашманов И.*, *Григорьев С.*, *Гусев В.*, *Харин Н.*, *Шабанов В.* Применение статистических методов для интеллектуальной компьютерной обработки текстов / Труды Международного семинара Диалог'97 по компьютерной лингвистике и ее приложениям. Ясная Поляна, 10–15 июня 1997 г. С. 33–37.
4. *Солтон Дж.* Динамические библиотечно-информационные системы. – М.: Мир, 1979.
5. *Ашманов И.*, *Харин Н.* Интеллектуальные технологии обработки текстов. Электронный офис, май-июнь 1997, С. 24–25.
6. *Y. Qui*, *H.P. Frei*. Concept based query expansion. ACM SIGIR, 1993.

Е.А Ломако

О НЕКОТОРЫХ АСПЕКТАХ БЕЗОПАСНОСТИ СОВРЕМЕННЫХ ИНФОРМАЦИОННЫХ СИСТЕМ

Современные информационные системы (ИС) *представляют собой комплексные программно-аппаратные решения*, базирующиеся на различных технологиях и способах их реализации. Так, для большинства ИС будет верно распределение составляющих их элементов по следующим ролям:

- хранение и обработка данных;
- представление данных;
- доступ к данным.

Роль хранения и обработки данных обычно выполняют различные базы данных. Применение той или иной системы хранения обусловлено различными факторами, в том числе: объёмом данных, сложностью их обработки, требуемой скоростью обработки и обмена данными, совместимостью с другими подсистемами и т. д. Наиболее значимые на рынке продукты представлены компаниями Oracle, Microsoft, IBM.