ОБРАБОТКА ЕСТЕСТВЕННО-ЯЗЫКОВЫХ ЗАПРОСОВ К ПОИСКОВОЙ МАШИНЕ НА ОСНОВЕ ИХ ЛИНГВИСТИЧЕСКОГО АНАЛИЗА

NATURAL LANGUAGE QUERY PROCESSING FOR SEARCH ENGINE BASED ON LINGUISTIC ANALYSIS

Ермаков А.Е., Плешко В.В.

ООО "ЭР СИ О" (www.rco.ru)

Компьютерная лингвистика и интеллектуальные технологии: по материалам ежегодной Международной конференции Диалог 2009 (Бекасово, 27-31 мая 2009 г.). Вып. 8(15). - М.:РГГУ, 2009. - 620с.

Аннотация

Описывается новый способ преобразования запросов на естественном языке в языки запросов поисковых машин, основанный на машинном анализе синтаксических связей между словами и их отображении на соответствующие операторы языка поисковой машины с максимальным сохранением смысла исходного запроса.

Введение

Языки запросов современных поисковых машин, используемых для поиска текстов в базах данных или полнотекстовых хранилищах документов, разрешают задавать различные ограничения на искомые комбинации слов в тексте, определяя обязательность или необязательность присутствия тех или иных слов, допустимое расстояние между словами и порядок их следования в тексте, а также позволяя искать слова во всех грамматических формах, что дает возможность в принципе формулировать очень сложные запросы, точно и полно описывая возможные способы выражения в тексте искомого смысла. Такие возможности поддерживают, к примеру, поисковые машины в СУБД Oracle и СУБД Microsoft SQL Server, поисковые машины компаний Google и Яндекс.

Проблема создания хороших информационно-поисковых систем на базе поисковых машин заключается в том, что пользователь системы часто желает формулировать свой запрос в виде простого набора слов, словосочетаний или фразы на естественном языке, ожидая от системы понимания хотя бы элементарных способов того, в какой форме соответствующий смысл может быть выражен в тексте. Так, большинство поисковых запросов, по которым пользователь может найти требуемые тексты, состоят более чем из одного слова. Здесь начинаются проблемы – как обработать запрос из нескольких слов, в каком виде транслировать его поисковой машине?

Во-первых, не ясно, как искать вхождение слов в документ – как цепочку подряд следующих слов (используя оператор языка запросов для поиска “по фразе”), как набор близко расположенных слов (используя оператор типа "рядом"), просто как набор встречающихся совместно в одном документе слов (используя оператор И), или как набор таких слов, из которых лишь некоторые должны обязательно встретиться в документе (используя оператор ИЛИ).

Во-вторых, не ясно, как расширять слова запроса грамматическими формами. Если искать все формы для каждого слова запроса, то точность поиска оказывается не высока по двум причинам. Во-первых, без учета грамматических связей слов запроса нет возможности разрешить омонимию: например, при обработке запросов *решение суда* и *грузовые суда* все варианты словоформы *суда* следует строить в первом случае от слова *суд*, а во втором - от слова *судно*. Во-вторых, при поиске словосочетаний допустимыми являются не все грамматические формы слов: например, при обработке

запроса *президент России* слово *президент* стоит искать во всех вариантах, а слово *Россия* следует искать только в заданной форме родительного падежа, иначе можно найти фрагменты текста следующего вида: *к встрече американского президента Россия готовилась заблаговременно*. Кроме того, поиск слова во всех грамматических формах обычно увеличивает нагрузку на поисковую машину.

В итоге, обычно информационно-поисковые системы инициируют поиск всех слов запроса по ИЛИ либо по И, допуская каждое слово во всех грамматических формах, используя ту особенность поисковых машин, что те обычно ранжируют найденные документы по релевантности таким образом, что первыми в результатах поиска выдаются документы, содержащие наибольшее количество слов из запроса, в которых эти слова расположены наиболее близко в тексте. Поскольку при этом никак не учитывается связанность слов в запросе, результаты поиска могут содержать ошибки, вызванные случайной близостью в тексте не связанных по смыслу слов. Так, например, все слова словосочетаний *президент России* и *российский президент* целесообразно искать в тексте только рядом, поскольку большинство других случаев их близкого положения будут соответствовать совершенно иным смыслам. Напротив, слова словосочетания *зарегистрировать изобретение* могут находиться в тексте рядом в любом порядке, будучи разделенными другими словами, например: *изобретение способа преобразования запросов, которое так и не было зарегистрировано*. Помимо невысокой точности, избыточный поиск по ИЛИ обычно также увеличивает нагрузку на поисковую машину.

Для повышения точности поиска обычно используют информацию о частоте встречаемости слов запроса в найденных документах и во всей коллекции, по которой ведется поиск [1]. Наиболее ярким примером отечественной системы, воплотившей данный подход, является Галактика-Zoom [3].

Для повышения точности поиска в академических коллективах разрабатываются методы, основанные на предварительном лингвистическом разборе текстов [4-6]. Для эффективного практического применения такие методы требуют сохранения полученных описаний грамматической или семантической структуры в специальном индексе, который затем должен использоваться при поиске для сравнения со структурой запроса, получаемой лингвистическим анализатором. При доступных сегодня вычислительных мощностях данный подход не является промышленным, так как требует, во-первых, значительных вычислительных затрат для лингвистического анализа индексируемой коллекции текстов, а во-вторых, разработки специализированной поисковой машины, вследствие чего, в частности, не может быть универсально применен к любой базе данных. Поэтому практические попытки применения лингвистических методов к искомым текстам ограничиваются созданием мета-поисковых систем, которые лишь пытаются переупорядочить документы, найденные другой информационно-поисковой системой, на основании анализа небольших фрагментов текста, выданных поисковиком в качестве рефератов по запросу.

Другие известные анонсированные методы связаны с применением тезаурусов, например [2,7,8], и предназначены для повышения не точности, а полноты поиска (за исключением случаев, когда тезаурус используется для снятия омонимии).

Полезный результат, достигаемый при использовании описываемого способа поиска, заключается в повышении точности поиска при сохранении его высокой полноты, а также в снижении нагрузки на поисковую машину.

Настоящий способ поиска основан на использовании лингвистических знаний о грамматике того естественного языка, на котором формулируется поисковый запрос, и предлагает использовать синтаксические связи между словами поискового запроса для выбора оптимального выражения на языке запросов поисковой машины, а также, при

отсутствии результата поиска документов по этому выражению, для формирования последовательности поисковых выражений с уменьшающейся степенью строгости поисковых ограничений и с максимально возможным сохранением смысла исходного запроса, что обеспечивает последовательное повышение полноты поиска с минимальной потерей точности. Соответствие операторов языка запросов синтаксическим связям между словами устанавливается на основании того принципа, что более сильно связанные в запросе слова должны искаться на более близком расстоянии в тексте и с более жесткими ограничениями на допустимые грамматические формы.

Базовый способ преобразования запроса

Конкретное соответствие типов синтаксических связей между словами поискового запроса и операторов языка запросов поисковой машины может быть различным, поскольку зависит от:

а) обрабатываемого естественного языка (русский, английский и др),

б) используемого синтаксического анализатора и типов выделяемых им синтаксических связей,

в) используемой поисковой машины и поддерживаемых ей операторов языка запросов.

В наиболее сложном случае возможно отображение связей между словами на операторы из следующего множества: И, РЯДОМ, РЯДОМ_УПОРЯДОЧЕННО, ФРАЗА, ВО_ВСЕХ_ФОРМАХ.

Ниже в таблице 1 приведены примеры установления соответствия между основными синтаксическими связями в русском языке и перечисленными операторами. В этих и следующих примерах будет использована нотация, абстрагированная от формальных особенностей языка какой-либо конкретной поисковой машины, но позволяющая записать поисковые выражения с использованием всех общепринятых операторов:

- следующие друг за другом слова должны искаться в тексте "как фраза" (в соседних слово-местах);
- оператор *and* означает, что слова должны встречаться в одном тексте в любых местах;
- операторы *near* и *near_ord* означают, что слова должны находиться в тексте на небольшом расстоянии друг от друга, причем второй маркер дополнительно указывает, что порядок слов в тексте должен соответствовать порядку слов в поисковом выражении. Соответствующие операторы в различных поисковых машинах имеют свои особенности реализации;
- оператор *m:* означает, что соответствующее слово должно искаться во всех грамматических формах (иначе слово ищется только в указанной форме).
- оператор = означает эквивалентность указанных слов при поиске

Тип синтаксической связи	**Оператор запроса**	**Форма подчиненного слова**	**Пример фрагмента запроса**
прилагательное или причастие в составе именной группы (*рыжий <- конь, пишущее -> устройство*)	ФРАЗА	как у главного	`(m:РЫЖИЙ m:КОНЬ)` `(m:ПИШУЩЕЕ m:УСТРОЙСТВО)`
приложение (*царь -> Иван*)	ФРАЗА	как у главного	`(m:ЦАРЬ m:ИВАН)`
генитив (*отношение -> принадлежности*)	ФРАЗА	исходная	`(m:ОТНОШЕНИЕ ПРИНАДЛЕЖНОСТИ)`

Тип синтаксической связи	**Оператор запроса**	**Форма подчиненного слова**	**Пример фрагмента запроса**
деепричастие (*двигаться, -> гремя*)	РЯДОМ	любая	ДВИГАТЬСЯ near ГРЕМЕТЬ
инфинитив (*попытка -> заставить -> работать*)	ФРАЗА	исходная	(m:ПОПЫТКА ЗАСТАВИТЬ РАБОТАТЬ)
предлог (*под <- капотом*)	ФРАЗА	исходная	(ПОД КАПОТОМ)
аргумент предиката при глаголе (*клиент <- арендует -> у предприятия)*	РЯДОМ	исходная, если с предлогом, иначе – любая	m:КЛИЕНТ near m:АРЕНДОВАТЬ near (У m:ПРЕДПРИЯТИЕ)
аргумент предиката при существительном (*аренда -> земли, договор -> (о) разоружении*).	РЯДОМ_УПОРЯДОЧЕННО	исходная, возможно с предлогом	m:АРЕНДА near_ord ЗЕМЛИ m:ДОГОВОР near_ord (О РАЗОРУЖЕНИИ)
обстоятельство (*отчет -> (при) упрощенке*)	И	исходная с предлогом	m:ОТЧЕТ and (ПРИ УПРОЩЕНКЕ)
синтаксически ничему не подчиненные слова (*платежи перечисление*)	И	любая	m:ПЛАТЕЖ and m:ПЕРЕЧИСЛЕНИЕ

Таблица 1. Пример установления соответствия между основными синтаксическими связями в русском языке и основными операторами поисковых машин.

Рассмотрим пример конструирования запроса к поисковой машине для запроса *авансовые платежи налог на прибыль предприятий*.

В терминах приведенной выше нотации по этому запросу может быть сконструировано следующее выражение: *(m:АВАНСОВЫЙ m:ПЛАТЕЖ) and (m:НАЛОГ near_ord (НА near_ord (ПРИБЫЛЬ ПРЕДПРИЯТИЙ)))*.

Соответствующее выражение на языке запросов, поддерживаемом в СУБД Oracle, где оператор *near* требует указания максимального расстояния между словами, (например, 5) и флаг учёта порядка слов (TRUE, что реализует оператор *near_ord*), будет выглядеть так: *($АВАНСОВЫЙ $ПЛАТЕЖ) and (near(($НАЛОГ, НА, ПРИБЫЛЬ ПРЕДПРИЯТИЙ), 5, TRUE))*.

Соответствующее выражение на языке запросов, поддерживаемом в СУБД MS SQL Server, где не существует оператора, эквивалентного *near_ord* (используется только *near*), а оператор, эквивалентный *m:*, может применяться только ко всем словам фразы, будет выглядеть так: *FORMSOF(INFLECTIONAL, "АВАНСОВЫЙ ПЛАТЕЖ") AND (FORMSOF(INFLECTIONAL, НАЛОГ) NEAR НА NEAR "ПРИБЫЛЬ ПРЕДПРИЯТИЙ")*.

Построение последовательности поисковых выражений

В информационно-поисковой системе, если полнота поиска по исходному запросу оказалась неудовлетворительной (найдено недостаточно документов), запрос может быть преобразован с некоторым ослаблением поисковых ограничений и вновь передан поисковой машине, что может в итоге привести к построению целой последовательности поисковых выражений до тех пор, пока не будет достигнута требуемая полнота поиска. Ниже описываются те преобразования, которые ослабляют поисковые ограничения, одновременно пытаясь максимально сохранить смысл исходного запроса.

Во-первых, при конструировании поискового выражения возможно из запроса на естественном языке исключать слова, синтаксически или семантически подчиненные другим словам.

Так, при исключении зависимых слов *авансовый* и *предприятие* может быть сконструировано выражение *(m:ПЛАТЕЖ) and (m:НАЛОГ near_ord (НА near_ord (ПРИБЫЛЬ)))*. А при исключении из исходного запроса зависимого слова *на* может быть получено *(m:АВАНСОВЫЙ m:ПЛАТЕЖ) and (m:НАЛОГ near (m:ПРИБЫЛЬ ПРЕДПРИЯТИЙ))*.

Однако, часто именно синтаксически зависимые слова наиболее точно определяют предмет поиска и их удаление делает запрос бессмысленным, например, *общий принцип изменения и расторжения договора*. Поэтому, критерий синтаксической зависимости при исключении слов из запроса является менее приоритетным, чем следующий критерий, который учитывает лексическую значимость слов.

Во-вторых, при конструировании поискового выражения возможно из запроса исключать слова, входящие в заданный стоп-словарь, с сохранением грамматики фразы. Например, существительное исключается вместе с согласованными определениями.

Так, при конструировании выражения для запроса *высокий рост детской смертности в Никарагуа* можно исключить общеупотребимое слово *рост* вместе со своим определением *высокий*. В результате получим следующее выражение: *(m:ДЕТСКИЙ m:СМЕРТНОСТЬ) and m:НИКАРАГУА*.

И напротив, возможно из поискового запроса исключать слова, не входящие в заданный словарь терминов предметной области. Так, при исключении из запроса *система современного налогообложения в малом бизнесе* слов *система* и *современный*, не входящих в словарь юридических терминов, получим следующее выражение: *m:НАЛОГООБЛОЖЕНИЕ and (m:МАЛЫЙ m:БИЗНЕС)*.

В-третьих, при конструировании поискового выражения возможно любое из слов или словосочетаний запроса заменить на слово или словосочетание из словаря синонимов, гипонимов или других связанных по смыслу слов, объединяемых в поисковом запросе оператором =, *or* или другим эквивалентным по смыслу оператором.

Так, для запроса *рост доходов населения в России* соответствующее выражение при расширении синонимами будет выглядеть следующим образом: *(m:(РОСТ=ПОВЫШЕНИЕ=ПОДЪЕМ) ДОХОДОВ=ПРИБЫЛИ НАСЕЛЕНИЯ=ЖИТЕЛЕЙ) and m:(РОССИЯ=РФ)*.

В-четвертых, если в результате поиска по любому из приведенных способов найдено недостаточно документов, возможно для каждого поискового выражения строить несколько похожих выражений с меньшей степенью строгости путем замены одних операторов поиска на другие *(ФРАЗА -> РЯДОМ -> И -> ИЛИ)*.

Рассмотрим пример построения последовательности поисковых выражений для приведенного выше запроса *авансовые платежи налог на прибыль предприятий*. Исходное выражение *(m:АВАНСОВЫЙ m:ПЛАТЕЖ) and (m:НАЛОГ near_ord (НА near_ord (ПРИБЫЛЬ ПРЕДПРИЯТИЙ)))* может быть далее преобразовано в следующие последовательности выражений:

1.1. удаление зависимых слов: *(m:ПЛАТЕЖ) and (m:НАЛОГ near_ord (НА near_ord (ПРИБЫЛЬ)))*;

1.2. расширение синонимами: *(m:ПЛАТЕЖ=ПЛАТА=ПЛАТИТЬ=ЗАПЛАТИТЬ) and (m:НАЛОГ near_ord (НА near_ord (ПРИБЫЛЬ=ДОХОД)))*;
или
2.1. замена части операторов на "более слабые": *(m:АВАНСОВЫЙ m:ПЛАТЕЖ) and (m:НАЛОГ near (m:ПРИБЫЛЬ ПРЕДПРИЯТИЙ))*;
2.2. замена части операторов на "более слабые": *(m:АВАНСОВЫЙ m:ПЛАТЕЖ) or (m:НАЛОГ and (m:ПРИБЫЛЬ ПРЕДПРИЯТИЙ))*;
2.3. удаление зависимых слов: *(m:ПЛАТЕЖ) or (m:НАЛОГ and m:ПРИБЫЛЬ)*.

Заключение

Описанный способ обработки поисковых запросов на естественном языке был успешно апробирован в одном из проектов компании "ЭР СИ О" (http://www.rco.ru) на базе специализированной поисковой машины заказчика и показал заметное повышение точности поиска на многословных запросах. Дальнейшее практическое исследование, в том числе сравнение качества поиска с другими информационно-поисковыми системами, планируется провести в рамках очередного семинара РОМИП (http://romip.narod.ru/). На изобретение подана и зарегистрирована патентная заявка "Способ выполнения поиска в компьютерной системе" №2008138379 от 26.09.2008.

Литература

1. C. J. Van Rijsbergen. Information Retrieval, 2nd edition. – Butterworths, London, 1979.

2. Солтон Дж. Динамические библиотечно-информационные системы. – Пер. с англ. – М.: Мир, 1979. – 558 с.

3. Антонов А.В., Курзинер Е.С., Новые возможности поисково-аналитической системы "Галактика-Zoom" (ранжирование документов по значимости). // Материалы конференции "Диалог-2003". (http://www.dialog-21.ru/archive/2003/Antonov.htm)

4. Осипов Г.С. и др. Проблемы обеспечения точности и полноты поиска: Пути решения в интеллектуальной мета-поисковой системе "Сириус". // Компьютерная лингвистика и интеллектуальные технологии: труды Международной конференции Диалог'2005. – Москва, Наука, 2005. - С. 390-394.

5. Тихомиров И.А. Вопросно-ответный поиск в интеллектуальной поисковой системе Exactus // Российский семинар по Оценке Методов Информационного Поиска. Труды четвертого российского семинара РОМИП'2006. – Санкт-Петербург: НУ ЦСИ, 2006, - С. 80-85.

6. Окатьев В.В., Баркалов К.А., патент № 2320005 на изобретение «Способ поиска информации», опубликовано 20.03.2008, ООО "Диктум".

7. Лукашевич Н.В., Добров Б.В., Тезаурус русского языка для автоматической обработки больших текстовых коллекций // Компьютерная лингвистика и интеллектуальные технологии: Труды Международного семинара Диалог'2002 / Под ред. А.С. Нариньяни – М.: Наука, 2002. Т.2. С.338-346.

8. Брин С., Гомес Б., Тонг С. патент № 2324220 на изобретение «Оснащение пользовательского интерфейса расширением поисковых запросов», опубликовано 10.05.2008, Гугл Инк. (US).