

УДК 004.89:004.912

Е.В. Малащук

Институт проблем искусственного интеллекта, г. Донецк, Украина
evm.it@inbox.ru

Средства семантического поиска близких по общему контексту документов

В статье показано, что при использовании средств семантической обработки текста необходимо учитывать весь предшествующий контекст, использования локальных контекстов недостаточно для выявления правильных значений. Рассматриваются методы сопоставления концептов на основе анализа синтаксиса и семантики. Предлагается способ автоматизированного поиска новых для пользователя документов в сети Интернет.

Целью работы является получение практической методики и необходимых шагов алгоритма семантического поиска, позволяющей отбирать близкие по общему контексту документы независимо от их принадлежности к разным предметным областям для извлечения и обновления знаний.

Введение

Как указывается в [1], за последние несколько лет большинство крупных государственных предприятий и частных фирм перешли на автоматизированные системы управления. Неотъемлемой частью любой такой системы является база данных товаров, клиентов, услуг и т.п. Соответственно необходимо обрабатывать большие массивы жестко структурированных данных. Однако гораздо чаще возникает проблема поиска и выборки необходимой информации из большого неструктурированного массива, содержащего как текстовые, так и мультимедийные данные.

В [2] семантический поиск определяется как поиск по содержательным аспектам всех компонентов документов электронной библиотеки. Но в связи с отсутствием в настоящее время единого понимания «содержательных аспектов» аудио-, видео- и других графических компонентов, в данной работе мы ограничиваемся следующим определением: «семантический поиск – это вид автоматизированного полнотекстового информационного поиска с учетом смыслового содержания слов и словосочетаний запроса пользователя и предложений текстов проиндексированных информационных ресурсов» [3].

Обращение к данной тематике вызвано нуждами проекта по разработке высокоэффективной интеллектуальной системы семантического метапоиска. Как и в любой информационной системе поиска, здесь также возникает задача снятия омонимии для выбора документа, правильно отображающего смысл текстового запроса [4], [5].

Некоторые понятия, однозначно воспринимаемые в естественном языке, в компьютерном словаре приобретают несколько значений, которые должны быть выделены и разделены для генерации правильного результата поиска. Методики,

базирующиеся на использовании тезаурусов [6], в данном случае применимы лишь частично, так как определение смысла, передающего нужное значение, требует более глубокой обработки контекста.

Использования локальных контекстов [7] явно недостаточно для снятия омонимии при осуществлении интеллектуального поиска. В связи с тем, что поисковые фразы передают не значения слов, а концепты, часто описываемые группой слов, предложен метод, расширяющий положения [7]. Описание метода изложено ниже.

Положения, лежащие в основе данного метода, прошли экспериментальную проверку [8] и могут использоваться при поиске в сети Интернет документов, содержащих новые знания.

Постановка задачи. Проблемы понимания текста

Одной из наиболее мощных современных систем, использующих технологии семантического поиска, является RetrievalWare компании Convera (ранее Excalibur), позиционирующей свою систему как «первую платформу по извлечению знаний» [9].

Основой семантического поиска в RetrievalWare является использование семантических сетей, описывающих смысл слов языка и связи между обозначаемыми ими понятиями. В [1] отмечается, что в данном случае под термином «семантическая сеть» понимается тезаурус, позволяющий не только найти слова, связанные по смыслу с данным, но и определить количественно «семантическое расстояние между ними». Но как указывается в [6], «часто различные словарные источники дают различный набор значений многозначных слов, выделяют оттенки значений, причем один и тот же тип многозначности может быть по-разному описан для различных слов даже в одном и том же словаре».

Нужно заметить, что в ряде случаев под многозначностью маскируются разные слова, обозначающие разные понятия и имеющие разные, но частично совпадающие парадигмы. Так, например, в [10] отмечается «относительное равноправие форм “директоры” и “директора” в текстах разных типов», при этом замечается, что «форма множественного числа “директоры” давно признается архаичной».

Справедливо отмечая, что эти слова встречаются в «текстах разных типов», в работе [10] не указывают типы текстов. Очевидно, в первом случае слово «директоры» встречается в технических текстах, где это слово обозначает элемент антенны (например, телевизионной) или монтажной оправки. Во втором случае слово «директора» встречается в текстах, где упоминаются административные должности; именно в этом случае форма множественного числа «директоры» практически не применяется.

Указанные два типа текстов относятся к разным предметным областям, и рассмотренные слова являются не различными формами одного и того же слова, а внешне похожими представителями разных миров, имеющими общего предка.

Несмотря на общепринятое мнение о том, что существует единая научная картина мира, вероятнее всего, что это лишь идеал, к которому стремится наука. У каждого человека имеется собственная внутренняя картина мира, причем она конечна. Это доказывается тем, что всегда найдется вопрос, на который данный человек не сможет ответить. Единство человеческого общества (компания друзей, народности, цивилизации) достигается путем усвоения (в той или иной степени) в результате обучения единых (или близких) понятий явлений окружающей действительности, образующих объекты внутренней картины мира. Но перечень этих одинаково с другими понимаемых явлений разный для каждого человека.

Совокупность совпадающих по значениям объектов внутренних картин мира составляет картину мира (предметную область) сообщества людей, например, профессионального. Но следует еще раз отметить, что внутренняя картина мира любого члена этого сообщества в целом будет отлична (пусть по некоторым элементам) от общей картины мира сообщества, в частности за счет разного понимания даже общепринятых толкований фактов и явлений. Этим объясняются отмеченные в [6] отличия в словарях, составленных разными людьми.

Помимо отличий в картинах мира разных людей, внутренняя картина мира человека, вероятно, представляет совокупность отдельных предметных областей, слабо связанных (или не связанных) друг с другом. По всей видимости, этим объясняется сложность переноса известных в одних условиях взаимодействий объектов в другие условия (в новую предметную область), что является основой методологии ТРИЗ [11], используемой при целенаправленном создании новых изобретений.

Одно и то же слово в разных контекстах приобретает разный смысл. Значения слов омонимов в разных локальных контекстах, очевидно, различны. Но при этом предметные области также различны и не связаны друг с другом. В то же время значения слов могут быть одинаковы, но предметные области могут различаться. И, наконец, одна предметная область примера может частично или полностью поглощать другую предметную область.

Можно констатировать, что понимание одного и того же текста разными людьми в той или иной степени отличается. Соответственно это влияет как на формулировку запроса, так и на оценку результатов его выполнения.

Учитывая наблюдения психологов, что «единицы мысли и единицы речи не совпадают» [12], а также изменчивость значений слов в зависимости от контекста, ясно, что результаты семантического поиска в тексте не могут быть абсолютно точными. Значение результата поиска может лишь с некоторой вероятностью соответствовать смыслу запроса, причем оценка соответствия не является абсолютной, а зависит от позиции оценивающего.

Методика семантического поиска

В приведенных выше примерах мы не выходили за пределы локального контекста, который обычно ограничен одним предложением.

В [7] сформулировано положение, что «Two different words are likely to have similar meanings if they occur in identical local contexts». Это положение используется для выбора нужного значения слова в словаре, используя пояснительные тексты словарных статей.

Но при семантическом поиске использование локального контекста оказывается недостаточным. Полезная информация может быть почерпнута лишь из общего содержания текста [13].

Расширим положение D. Lin [7], сформулировав его следующим образом: два различных слова (или группы слов) могут иметь схожее значение, если они встречаются в идентичных (или схожих) контекстах.

Но в этом случае слово (или группа слов), значение которого определяется путем сопоставления окружающего его контекста и контекста справочного слова (или группы слов), является лишь «заглушкой», резервирующей место в контексте. Эта «заглушка» может быть опущена, если предусмотрены меры, обеспечивающие соответствующий «зазор» в контексте.

Известен способ поиска схожих по семантике документов на основе сопоставления их лексических векторов [14]. Но при задании поискового запроса, включающего все (или наиболее значимые) лексические векторы документа-прототипа (или обобщенные векторы группы документов), документы, относящиеся к другим (хотя и сравнительно близким) предметным областям, будут отсеяны.

В то же время известен так называемый «сленговый метод» [15], используемый, например, при установлении авторства литературных произведений, по которому при сопоставлении документов исключаются наиболее частотные слова.

Предлагаемый метод семантического поиска является модификацией «сленгового метода» и использует замену «заглушкой» наиболее значимого лексического вектора запроса. В этом случае в результате поиска находятся документы, принадлежащие к разным (но сравнительно близким) предметным областям, описывающие разные (в том числе и ранее неизвестные для автора запроса) значения понятия, соответствующего исключенному при запросе лексическому вектору.

Описанный метод использован при поиске в сети Интернет и получил экспериментальное подтверждение [8]. В результате проведенных экспериментов получена индикативная математическая модель релевантности запроса:

$$Y_{21} = 0,517 + 0,157A + 0,207AB - 0,264AC + 0,122BC - 0,246ABC.$$

Здесь символами В и С соответственно обозначены учет синтаксиса (порядка слов в запросе) и морфологии. Характерно, что значимым оказалось лишь взаимодействие этих факторов. Это свидетельствует, что при семантическом поиске необходимо одновременно учитывать как морфологию слов запроса, так и его синтаксис.

Очевидно, что для этого необходимо проводить анализ текстов полученных в результате поиска документов, так как индексы поисковых машин, сохраняя информацию о положении слов в документе (при координатном индексе), не хранят знаки препинания, играющие существенную роль при синтаксическом анализе текстов [16].

Роль знаков препинания обычно в системах синтаксического анализа принижается, отводя им лишь роль разделителей текста. Однако учет знаков препинания в ряде случаев облегчает семантический анализ текста, особенно в таких случаях, как в задаче о постановке запятой во фразе «КАЗНИТЬ НЕЛЬЗЯ ПОМИЛОВАТЬ». Разделение этих значений может быть осуществлено на основе анализа знаков препинания без сложного анализа контекста. Но справедливости ради заметим, что это не всегда возможно.

Область применения

Предлагаемый способ семантического поиска рассматривался выше применительно к задаче поиска новой полезной информации. Но он может использоваться не только для этого, а, например, и в системах автоматического понятийного перевода.

Для преобразования WWW в Семантическую сеть необходимо в миллионах существующих документов расставить тэги, задающие значения URI [17] хотя бы основных терминов документа. Очевидно, что этот процесс может быть осуществлен только соответствующими программными агентами.

Метод его работы в общем случае тривиален: поиск онтологии, в которой есть описание данного термина; при наличии нескольких таких онтологий – уточнение значения термина по контексту (снятие омонимии) и выбор соответствующей онтологии; определение и запись в документ значения URI термина, что в дальнейшем обеспечивает однозначное определение этого термина.

Но что делать, если нужно задать URI термина, для которого не удастся установить связь с какой-либо из онтологий, значение которого отсутствует в доступных словарях синонимов?

Представляется, что формирование запроса на основе предложения, содержащего такой термин, но в котором этот термин заменен пропуском, позволит найти в Сети синонимы этого термина, имеющие такой же окружающий контекст. Тогда значение этого термина (URI) может быть определено на основе значения синонима.

Результаты экспериментов показывают, что предлагаемый способ поиска оказывается достаточно эффективным и при поиске новой информации по заданной тематике. При этом возможен следующий алгоритм работы агента, ищущего для пользователя новую информацию в Сети WWW:

1. Имеющиеся документы по определенной тематике объединяются в одну группу.

2. В этой группе документов определяются наиболее часто встречающиеся слова и словосочетания, а также определяются мера близости документов и критерий отнесения документов к указанной группе.

3. Выделяются наиболее часто встречающиеся предложения, содержащие указанные наиболее часто встречающиеся слова и словосочетания.

4. Эти предложения используются в качестве поисковых запросов, причем указанные наиболее часто встречающиеся слова и словосочетания исключаются и заменяются операторами, указывающими поисковой машине, что в запросе имеется пропуск величиной *n* слов.

5. Результаты запроса сравниваются с документами группы (п. 1). Из результатов исключаются документы, совпадающие с уже имеющимися в группе, а также те, которые по заданному критерию не относятся к данной группе.

6. Релевантные результаты запроса добавляются в соответствующую группу документов.

7. Пункты 1 – 6 периодически повторяются.

Выводы

Предложенная методика семантического поиска позволяет отбирать близкие по общему контексту документы, даже если они принадлежат к разным предметным областям. Тем самым можно собирать или извлекать, обновлять и обобщать знания, рассредоточенные в различных областях.

Предполагается, что предложенная методика может быть полезной при преобразовании сети Интернет в Семантическую сеть.

Литература

1. Бойцов И. Системы поиска по массивам неструктурированной информации // РЕЛИБ, 2003 // <http://www.relib.com/articles/article.asp?id=216>
2. Зацман И.М. Семантический поиск научной информации: неоднородные коммуникативные компоненты и цветовая палитра объектов поиска // Компьютерная лингвистика и интеллектуальные технологии: Тр. Междунар. семинара «Диалог'2002». – М.: Наука, 2002. – Т. 2 – С. 214-227.
3. Попов А. Поиск в Интернете – внутри и снаружи // Компьютеры + программы. – 1996. – № 4. – С. 62-63.
4. Ландэ Д.В. Поиск знаний в Internet. Профессиональная работа. – СПб.: Диалектика, 2005. – 272 с.
5. Voskressenski A. Signs and speech: two forms of human communication // Proceedings of the Ninth International Conference «Speech and Computer» SPECOM'2004. – Saint-Petersburg (Russia), 2004. – P. 666-669.

6. Лукашевич Н.В., Добров Б.В. Тезаурус русского языка для автоматической обработки больших текстовых коллекций // Компьютерная лингвистика и интеллектуальные технологии: Тр. Междунар. семинара «Диалог'2002». – М.: Наука, 2002. – Т. 2. – С. 338-346.
7. Lin D. Using syntactic dependency as local context to resolve word sense ambiguity // Proceedings of the 35th annual meeting on Association for Computational Linguistics. – Madrid (Spain), 1997. – P. 64-71.
8. Воскресенский А.Л., Хахалин Г.К. Формирование запросов к поисковой машине для извлечения знаний из Интернета // Компьютерная лингвистика и интеллектуальные технологии: Тр. Междунар. конф. «Диалог'2005». – М.: Наука, 2005. – С. 86-91.
9. Papadopoulos A., Van Winkle J. RetrievalWare 8 – the Knowledge Discovery Platform // A Convera technical overview // <http://www.convera.com/whitepapers.TO-RW8-031111.pdf>
10. Беликов В.И. Yandex как лексикографический инструмент // Компьютерная лингвистика и интеллектуальные технологии: Тр. Междунар. конференции «Диалог'2004». – М.: Наука, 2004. – С. 39-46.
11. Альтшуллер Г.С. Творчество как точная наука: Теория решения изобретательских задач. – М.: Советское радио, 1979.
12. Выготский Л.С. Психология. – М.: ЭКСМО-Пресс, 2000.
13. Браславский П.И. Использование стилистических параметров документа при поиске информации в Internet // Доклады VI рабочего совещания по электронным публикациям. – Новосибирск: ИВТ СО РАН, 2001.
14. Андреев А.М., Березкин Д.В., Морозов В.В., Симаков К.В. Автоматическая классификация текстовых документов с использованием нейросетевых алгоритмов и семантического анализа // Тр. Пятой Всероссийской научной конференции «Электронные библиотеки: перспективные методы и технологии, электронные коллекции». – Санкт-Петербург, 2003 // <http://rcdl2003.spbu.ru/proceedings/B1.pdf>
15. Хайтун С.Д. Наукометрия. Состояние и перспективы. – М.: Наука, 1983.
16. Хахалин Г.К. Лингвистическая трансляция сложных и эллиптических ЕЯ-предложений // Тр. VIII Национальной конференции по искусственному интеллекту «КИИ – 2002». – Коломна, 2002. – С. 251-256.
17. Berners-Lee T., Hendler J., Lassila O. The Semantic Web // Scientific American. – May 17, 2001.

Є.В. Малащук

Засоби семантичного пошуку близьких за загальним контекстом документів

У статті показано, що при використанні засобів семантичної обробки тексту необхідно враховувати весь попередній загальний контекст, використання локальних контекстів недостатньо для визначення правильних значень. Розглядаються методи зіставлення концептів на основі аналізу синтаксису та семантики. Пропонується спосіб автоматизованого пошуку нових для користувача документів у мережі Інтернет.

E.V. Malaschuk

The Semantic Search Methods of Documents with the Relative General Context

Achieved that in using semantic search methods of documents it is necessary keep in mind all general context, because using only syntactic dependency as local context to resolve word sense ambiguity is not enough for extracting rule means. Described concept comparison methods based on syntax and semantic analysis. Propose method of the automated search for the new documents for user in to the Internet network.

Статья поступила в редакцию 27.06.2006.