УДК 519.688

М о з г о в о й М. В. Семантический анализатор и задача информационного поиска //Вестн. С-Петерб. ун-та. Сер. 10.2005. Вып. 3. С.00-00

Рассматривается один из возможных способов применения семантического анализатора Тузова в задаче информационного поиска. Кратко обсуждается задача информационного поиска, описываются существующие настольные системы, их достоинства и недостатки. Оцениваются новые возможности, связанные с использованием семантического анализа (такие как поиск по сходным понятиям), на примере простого поиска предложений в документе. Результаты обобщаются для случая поиска целых документов в коллекции и для задачи рубрикации. Библиогр. 6 назв. Табл. 2.

Summary

Mozgovoy M. V. Semantical analyzer and the problem of text retrieval.

We consider one of the possibilities of using Tuzov's semantic analyzer in the field of information retrieval. The problem of text retrieval is briefly discussed, some existing desktop systems (with their advantages and disadvantages) are described. We outline new opportunities, related to semantic analysis use (like searching by close conceptions) on the example of simple searching a sentence in a document. The results are generalized for the case of complete document searching and for the categorization problem.

УДК: 519.688 Вестник СПбГУ. Сер. 10, 2005, вып. 3
М. В. Мозговой

СЕМАНТИЧЕСКИЙ АНАЛИЗАТОР И ЗАДАЧА ИНФОРМАЦИОННОГО ПОИСКА

1. Введение. Общеизвестно, что задача анализа текстов на естественном языке давно является объектом пристального внимания специалистов. Кроме удовлетворения амбициозного желания научить компьютер понимать человеческие языки, решив эту задачу, можно достичь заметного прогресса в качестве таких сугубо прикладных систем как автоматические переводчики, вопросно-ответные системы, рубрикаторы и поисковые машины.

Семантический анализатор, разработанный В. А. Тузовым [1], с одной стороны, способен выполнить значительную работу по анализу текстов. Он автоматически выявляет структуру предложений и объясняет смысл каждого отдельного слова (используя для этого формальный семантический язык). С другой стороны, как показывает практика, семантический анализ — лишь первая ступень на пути к решению реально возникающих задач.

Не вызывает сомнений, что семантическая информация текста может и должна быть использована в системах, так или иначе работающих с текстами на естественном языке. Однако найти разумный способ ее применения не так-то просто.

Цель этой работы — показать один из способов использования вывода семантического анализатора в задаче информационного поиска, продемонстрировать его преимущества и указать возможные направления для дальнейшего развития.

2. Задача информационного поиска. Задачу информационного поиска можно сформулировать по-разному. Кроме того, при написании практически пригодной поисковой машины возникает огромное количество, строго говоря, не относящихся к информационному поиску проблем. Например, если речь идет о поиске текстов в Интернете, система должна уметь исследовать и индексировать веб-страницы; к тому же обработка столь больших объёмов данных наверняка потребует грамотного распараллеливания внутри вычислительного кластера.

В рассматриваемом случае можно ограничиться простейшей формулировкой. На локальном диске находится текстовый документ, являющийся исходным источником информации. Требуется определить предложение (или предложения) текста, наилучшим образом соответствующее пользовательскому запросу. В дальнейшем покажем, что эта задача (даже в столь ограниченной постановке) близкородственна задачам поиска целого текстового документа и рубрикации коллекции документов и решается примерно теми же методами.

3. Настольные поисковые системы. В настоящее время информационный поиск связывают прежде всего с поиском документов в Интернете. Однако организацию доступа к веб-страницам вряд ли можно назвать «чистым» информационным поиском. Пожалуй, для интернетовской поисковой машины обеспечение как можно большего объёма проиндексированной информации и быстрое реагирование на изменения в контенте важнее качества поиска. Поэтому разумнее рассмотреть так называемые настольные

поисковые системы, предназначенные для поиска информации на локальных дисках пользователя.

Например, уже несколько лет существуют и развиваются системы The Sleuthhound! фирмы iSleuthHound Technologies [2] и Архивариус 3000 производства Wizetech Soft [3].

Каждая из этих программ предлагает достаточно типичный для современного состояния систем такого класса набор функций:

- индексация файлов, находящихся на локальных дисках;
- быстрый поиск документов, содержащих введенные слова;
- поиск документов с использованием языка запросов (похожего на язык интернетовских поисковых машин);
- поддержка файлов различных форматов (txt, doc, pdf).

В новых версиях авторы обычно расширяют интерфейс пользователя новыми удобными функциями. Кроме того, постоянно пополняется список поддерживаемых форматов документов. Ядро же поискового алгоритма обычно остается неизменным. По сути дела, алгоритм поиска «соответствующего» запросу документа сводится к поиску документа, содержащего слова запроса (возможно, с использованием простых булевых конструкций).

Не отрицая полезности поиска документов с заданными словами, необходимо отметить, что подобный подход значительно сужает представление о «соответствующих» документах по сравнению с повседневным смыслом этого понятия. Например, фраза «президент РФ встретился с лидерами оппозиции», очевидно, соответствует запросу «Путин беседа оппозиция», тем не менее большинство современных поисковых машин с этим не согласятся.

Таким образом, система, не ориентированная на распознавание конструкций того или иного естественного языка, не может выполнить полноценный поиск информации. Некоторые поисковые машины (например, Яндекс) умеют работать с морфологией слов. В большинстве случаев эта возможность полезна, но порой нехватка семантического анализа становится очевидной. Так, на запрос «мыть» Яндекс возвращает множество документов, содержащих сочетания вроде «мой город», «моя семья», «мой сайт», никакого отношения не имеющие к мытью. Очевидно, что слово «моя» можно распознать как деепричастие, образованное от «мыть», но семантический анализ мог бы помочь отфильтровать ошибочно выбранные предложения.

4. Поиск с помощью семантического анализатора. Каким же образом семантический анализатор может улучшить качество поиска? Работающая экспериментальная система действует следующим образом.

Сначала генерируются семантические описания всех предложений исходного файла. Как уже было сказано выше, семантический анализатор умеет, во-первых, выявлять структуру предложений. Например, фраза *Президент заинтересован в улучшении отношений с союзниками* преобразуется в

```
@Крат заинтересован<X002.002>
 (@Им Президент<X001.001>,
  @вПред в<X003.101> (@Пред улучшении<X004.001>
     (@Род отношений<X005.003>
       (@сТв с<X006.028> (@Тв союзниками<X007.001>)))))).
```

Во-вторых, любое слово каждого предложения получает описание в виде так называемой семантической формулы, которая способна формально описать смысл слова, определив его как последовательность простых операций над теми или иными базовыми понятиями. Кроме того, слово помещается в соответствующий раздел иерархии понятий. Семантический анализатор в подавляющем большинстве случаев способен правильно выделить значение слова в данном контексте. Так, слово «коса» в контекстах «девушка с русой косой» и «девушка со стальной косой» будет иметь различное описание.

С одной стороны, полная семантическая формула дает полное представление о смысле слова, но ее труднее проанализировать и умело применить. С другой стороны, некоторые наиболее важные части формулы в отдельности также позволяют сделать кое-какие заключения об описываемом слове, не требуя сложной процедуры анализа. В настоящее время в «базе знаний» поисковой машины сохраняется лишь само слово, его класс (позиция в иерархии понятий), а также вид основной операции, при помощи которой это слово было выведено из базового понятия.

Если не использовать дополнительную информацию (класс и вид операции), получим обыкновенную систему поиска по ключевым словам. Нет единого рецепта для «правильного» применения семантических описаний; на данный момент система работает так:

- Если в предложении встретилось искомое слово и оно является именем собственным, рейтинг предложения повышается на 8 баллов.
- Если в предложении встретилось искомое слово, причем его значение в контексте документа совпадает со значением в контексте запроса, рейтинг предложения повышается на 5 баллов. Так, при запросе *русая коса* фраза *Девушка со стальной косой* не получит ни одного балла, поскольку значения слова «коса» в запросе и документе различаются.
- Если в предложении встретился класс, к которому принадлежит какое-либо слово запроса, рейтинг предложения повышается на 3 балла. Такая ситуация возникает, например, во фразе *Председатель Московской Хельсинкской группы Людмила Алексеева рассказала NEWSru.com о ситуации в Благовещенске** при определении соответствия запросу *руководитель*, поскольку слова «председатель» и «руководитель» принадлежат одному и тому же классу **Физический объект** / **Живой** / **Человек** / **Индивидуум** / **Профессия** / **Глава**.
- Если в предложении есть слово, в семантическом описании которого встречается та же операция, что и в описании некоторого слова запроса, рейтинг предложения повышается на 1 балл. К примеру, в формальном описании слова *удалось* встречается запись Hab$12411/033, что означает «иметь успех». Та же самая запись может быть найдена в описании слова *получаться***, поэтому рейтинг предложения, содержащего слово *получаться*, будет повышен на единицу, если в запросе встретилось слово *удалось*.

В табл. 1, 2 приведены примеры ответов системы на запросы пользователя.

Таблица 1. **Результат запроса «глава учреждения сообщил»**

(рейтинг: 16) Такое мнение высказал в воскресенье РИА «Новости» глава комитета Госдумы по международным делам Константин Косачев. *Соответствия:* глава — ГЛАВА; ...

*Все анализируемые тексты взяты с сервера NEWSru.com.

**В случае многозначных слов или омонимов наличие тех или иных операций в описании зависит от контекста слова.

учреждения — РИА_НОВОСТИ, ГОСДУМА, КОМИТЕТ; сообщил — ВЫСКАЗАТЬ
(рейтинг: 11) «Положение в стране крайне нестабильное» — отметил в интервью РИА «Новости» вице-спикер Госдумы Вячеслав Володин. *Соответствия:* глава — ВИЦЕ-СПИКЕР; учреждения — РИА_НОВОСТИ, ГОСДУМА; сообщил — ОТМЕТИТЬ, ИНТЕРВЬЮ
(рейтинг: 10) Такой неутешительный прогноз дал в интервью ИТАР-ТАСС заместитель директора Института океанологии РАН Леопольд Лобковский. *Соответствия:* глава — ЗАМЕСТИТЕЛЬ, ДИРЕКТОР; учреждения — ИТАР-ТАСС, ИНСТИТУТ; сообщил — ИНТЕРВЬЮ
(рейтинг: 3) Кроме того, в регионе ЮВА работают два российских госпиталя. *Соответствия:* учреждения — ГОСПИТАЛЬ

Таблица 2. **Результат запроса «сложные погодные условия»**

(рейтинг: 20) По словам собеседника агентства ситуацию осложняют погодные условия — сильный боковой ветер, минусовая температура и непрекращающийся снегопад. *Соответствия:* сложные — ОСЛОЖНЯТЬ; погодные — ПОГОДНЫЙ, ВЕТЕР, СНЕГОПАД; условия — УСЛОВИЯ
(рейтинг: 3) Аэропорт Ростова-на-Дону закрыт из-за сильного обледенения. *Соответствия:* погодные — ОБЛЕДЕНЕНИЕ
(рейтинг: 3) В выходные буран в Москве стихнет. *Соответствия:* погодные — БУРАН
(рейтинг: 3) В дневные часы термометр покажет минус 8–10 градусов в Москве и 7–12 градусов мороза в ее окрестностях. *Соответствия:* погодные — МИНУС, МОРОЗ

Как видно из них, большинство выданных системой предложений не содержат в точности указанных в запросе слов, тем не менее все они справедливо сочтены релевантными.

5. О задаче поиска документов и задаче рубрикации. Задача информационного поиска выше определялась как задача нахождения отдельных предложений входного документа, некоторым образом соответствующих пользовательскому запросу. Однако эта постановка ука-

зывает скорее на наши текущие интересы, а не на ограничения используемого алгоритма. Поскольку предложения исходного текста рассматриваются просто как наборы слов (пусть даже расширенных семантической информацией), нетрудно представить целый документ как совокупность слов, его составляющих, а затем применить описанный выше алгоритм расчета рейтинга. При желании несложно также воспользоваться известной моделью векторного пространства [4], изменив в ней лишь схему вычисления весов слов.

Имея пространство документов с введенной в нем метрикой, несложно решить задачу рубрикации, воспользовавшись любыми популярными алгоритмами классификации и кластеризации в метрическом пространстве [5, 6].

6. Выводы и перспективы развития. Приведенные примеры доказывают полезность использования даже базовой семантической информации в задаче информационного поиска. Поскольку в настоящее время исследования сконцентрированы на принципиальных возможностях применения семантического анализатора в самых разнообразных практических системах, разработанная программа как пригодный для повседневного использования продукт еще очень далека от совершенства. В частности, пользователю необходимо предоставить расширенный язык запросов, с помощью которого можно было бы сформулировать свои предпочтения более точно (этот язык мог бы содержать простые инструменты для семантической разметки).

Серьезные изменения могут быть связаны с более глубоким использованием предоставляемой анализатором семантической информации. Так, анализатор способен определить, что слово «черный» в контексте «черный кофе» относится к кофе, а в контексте «кофе в чёрном пакете» — к пакету. Однако поисковая система на данный момент игнорирует подобные различия.

Очень интересным направлением представляется создание пользовательских иерархий понятий. Стандартная иерархия, заложенная в семантический анализатор, пытается воспроизвести окружающий мир с достаточно субъективной (хотя и разумной) точки зрения. Например, с одной стороны, в этой иерархии слова «монета» и «бусинка» не имеют ничего общего. Слово «монета» принадлежит классу *Деньги*, а «бусинка» — классу *Украшения*. С другой стороны, предположим, что пользователю требуется сортировать предметы по размеру, и тогда как монета, так и бусинка попадают в один и тот же класс мелких предметов (допустим, имеющих размер меньше 5 см в поперечнике). Точно так же профессии можно сортировать не только по типам деятельности, но и по степени квалифицированности того или иного труда. К счастью, имея созданную специализированную иерархию (с распределенными по ней словами), существующую поисковую систему можно перестроить в автоматическом режиме без особых затруднений, раскрывая для пользователя целый набор новых возможностей.

Литература

1. *Тузов В. А.* Компьютерная семантика русского языка. СПб.: Изд-во С.-Петерб. ун-та, 2004. 400с.

2. *Вебсайт компании:* www.isleuthhound.com

3. *Вебсайт проекта:* www.wizetech.com/ru/document-search

4. *Baeza-Yates R., Ribeiro-Neto B.* Modern Information Retrieval. Boston: Addison Wesley Longman, 1999. 544p.

5. *Witten I. H., Frank E.* Data mining: practical machine learning tools and techniques with Java implementations. San Francisco: Morgan Kaufmann, 1999. 525p.

6. *Hand D. J., Mannila H., Smyth P.* Principles of data mining. Cambridge, MA: The MIT Press, 2001. 425p.