ПРОБЛЕМЫ ОБЕСПЕЧЕНИЯ ТОЧНОСТИ И ПОЛНОТЫ ПОИСКА: ПУТИ РЕШЕНИЯ В ИНТЕЛЛЕКТУАЛЬНОЙ МЕТАПОИСКОВОЙ СИСТЕМЕ «СИРИУС» *

Осипов Г. С.

ИСА РАН, Москва

gos@isa.ru

Завьялова О.С.

РУДН, Москва

olga_zavjalova@rambler.ru

Климовский А.А

РУДН, Москва

ydakra@mail.ru

Кузнецов И.А.

ИСА РАН, Москва

i441@mail.ru

Смирнов И.В.

ИСА РАН, Москва

ivanv_smirnov@mail.ru

Тихомиров И.А.

ИСА РАН, Москва

matandra@isa.ru

Доклад посвящен проблемам обеспечения точности и полноты поиска. Рассмотрение ведется на примере интеллектуальной метапоисковой системы «Сириус», в которой применяется метод семантико-синтаксического анализа, основанный на принципах коммуникативно-грамматической школы и использующий неоднородные семантические сети для представления фрагментов ситуаций, описанных в тексте. В докладе демонстрируется процедура семантико-синтаксического анализа текста, которая осуществляется при обработке запроса в системе, и разъясняются основные принципы этого анализа. Кроме того, в докладе рассмотрены другие особенности «Сириуса», такие как вычисление значимости фрагментов текста, расширение поискового запроса синонимами и сходными по смыслу именными группами, ввод запроса на естественном языке, возможности выбора различных стратегий поиска (профилей поиска) и т. д.

* Работы выполнены при поддержке РФФИ проект № 04-07-90097 и программы ОИТВС "Фундаментальные основы информационных технологий и систем" проект № 2.9

Лингвистические основы

Цель разработчиков всех поисковых систем – предоставить пользователю а) документы, в максимальной степени соответствующие смыслу запроса (обеспечить релевантность - точность поиска), и б) как можно большее число документов, содержащих запрашиваемую информацию (обеспечить полноту поиска).

Эта цель в разных системах достигается разными способами. В интеллектуальной метапоисковой системе «Сириус» точность поиска достигается, во-первых, в результате использования метода семантико-синтаксического анализа, основанного на теоретических положениях коммуникативно-грамматической школы [1]. Покажем, в чем состоят отличительные особенности данного метода.

Поиск по ключевым словам часто не удовлетворяет основному требованию пользователя, а именно требованию соответствия найденных документов запросу по смыслу, даже несмотря на то что все слова запроса в этих текстах присутствуют. Так, на запрос **выступление Путина в Думе** приходят документы такого рода: **Выступление** Г. А. Явлинского на заседании Государственной **Думы** по вопросу об утверждении кандидатуры **В.В. Путина** на пост премьер-министра.

Почему же поиск по ключевым словам не может обеспечить отбор только тех документов, которые отвечают запросу по смыслу?

Чтобы ответить на этот вопрос, важно понять, что запрос – это не просто группа слов. Это предложение/словосочетание, в котором слова соединяются друг с другом по определенным законам, которые изучает синтаксис.

Ключевое слово в запросе представляет собой не слово-**лексему**, т. е. единицу “словарного состава языка в совокупности его конкретных грамматических форм и выражающих их флексий, а также возможных конкретных смысловых вариантов” [2]. “Слово-лексема еще не является синтаксической единицей, слово – единица лексики, а в разных его формах могут реализоваться или актуализироваться разные стороны его обще-го значения, разные семы, предопределяющие различия и в синтаксическом употреблении. Формируя и изучая связную речь, синтаксис имеет дело с осмысленными единицами, несущими свой не индивидуально-лексический, а обобщенный, категориальный смысл в конструкциях разной степени сложности. Эти единицы характеризуются всегда взаимодействием морфологических, семантических и функциональных признаков” [1]. Эти единицы получили название **синтаксем**. В конкретном предложении запроса слово выступает как синтаксема.

В процессе поиска, когда мы работаем с текстом, целью поиска должна стать не лексема, а синтаксема, не только лексическое, но и производное от него **синтаксическое значение** компонента запроса (ключевого слова). Это позволит повысить точность поиска, а также сократить количество выдаваемых пользователю ненужных документов.

Важно подчеркнуть, что синтаксическое значение складывается в результате соединения категориального значения и морфологической формы, реализуется в определенной синтаксической позиции. Рассмотрение слова изолированно, в отрыве от текста, не позволит установить синтаксическое значение.

Покажем, что дает использование описанного подхода при информационном поиске, на примере двух запросов: дата изобретения вилки; кто изобрел вилку.

Задачей модуля лингвистической обработки «Сириуса» является построение семантического образа текста запроса и текста документа для последующего сопоставления получившихся образов. Этот образ создается в результате анализа текстов по следующей предложенной нами схеме.

Анализ начинается с нахождения предикатного слова (на данном этапе пока только глагольного или девербатива). На следующем этапе выделяются синтаксемы всех типов, окружающие глагол. Для этих целей в системе имеется словарь глаголов (и девербативов), в который вносится, во-первых, глагол, во-вторых, синтаксемы, которые могут употребляться при этом глаголе, в-третьих, информация, необходимая для идентификации синтаксемы: морфологическая форма, категориальная семантика имен существительных. Обратим внимание на то, что уже сам глагол задает ограничения для возможных прочтений той или иной синтаксемы, например, дать другу денег (синтаксема со значением адресата) – завидовать другу (синтаксема со значением объекта). Кроме того, в алгоритм работы модуля лингвистического анализа заложена система правил, позволяющих установить значение синтаксемы (например, для случаев, когда в предложении у двух или более имен существительных совпадают морфологическая форма и категориальная семантика; для структурно-семантических модификаций предложения).

После того как в запросе выделен предикат и синтаксемы, окружающие предикат, между синтаксемами устанавливаются семантические связи [3]. Семантическая связь является средством описания фрагмента ситуации.

В результате анализа формируется образ запроса, который далее сравнивается с образами документов, проанализированных аналогичным образом.

Вот результаты обработки наших запросов лингвистическим анализатором (приводим только те фрагменты, которые демонстрируют принципы работы анализатора; опущен последовательный перебор всех возможных вариантов):

Пример (1) дата изобретения вилки

Семантические классы существительных:
"дата" ("дата"): темпоративное
"изобретение" ("изобретения"): признаковое
"вилка" ("вилки"): предметное

СОСТАВ РОЛЕЙ ПОСЛЕ ФИЛЬТРАЦИИ
Предикатное слово "изобретения", "изобретение":
объект "вилки"

Пример (2) кто изобрел вилку

Семантические классы существительных:
"кто" ("кто"): анафорический элемент
"вилка" ("вилку"): предметное

СОСТАВ РОЛЕЙ ПОСЛЕ ФИЛЬТРАЦИИ
Предикатное слово "изобрел", "изобрести":
объект "вилку"
субъект "кто"
СВЯЗИ:
CAUS: "вилку"- "кто"

Как видим, в процессе анализа были определены значения и формы синтаксем. Еще раз обратим внимание на то, что идентифицировать синтаксему, определить ее синтаксическое значение помогают как морфологическая форма, так и категориальная семантика имени существительного, образующего синтаксему. Правильно определив синтаксему, ее значение (изобретение вилки - объект), мы можем отсечь большое количество нерелевантных документов. Возьмем в качестве примера запрос строение вилки, в которых вилка выступает субъектом при девербативе. В процессе обработки документов, пришедших по запросу, тексты с ключевым словом вилка в значении субъекта будут рассмотрены как нерелевантные и будут исключены из списка документов, предоставляемых пользователю в качестве результата поиска (это не относится к случаям, когда пользователь выбирает поиск по ключевым словам, как в обычном поисковике).

Рассмотрим результаты поиска, выполненного метапоисковой системой «Сириус» с использованием в качестве источника данных поисковую машину Яндекс. Для каждого запроса показаны первые две ссылки, возвращаемые Сириусом и Яндексом (отметим, что в режиме метапоиска «Сириус» анализирует и пересортировывает документы, найденные другими поисковыми машинами):

Пример (3) дата изобретения вилки

Сириус	Яндекс
1.**http://www.koryazhma.ru/usefull/know/doc.asp?doc_id=114** ..... кто изобрел ложку и **вилку** ? / Великие **изобретения**. Логин: Пароль : регистрация . помощь . . > полезное и интересное . > знаете ли , вы . > Великие **изобретения** . кто изобрел ложку и **вилку**?. Что-то похожее на современную **вилку** , только с пятью , а порой и большим количеством зубчиков появилось в Азии в десятом веке. …. Найдено с помощью **www.yandex.ru** 2. **http://www.nalivay.ru/restaurant/79888.html** ..... Кто и когда придумал **вилку** и как ели (+) Кто и когда придумал **вилку** и как ели (+) . [WWW - Конференции .] [Статистика .] [Обменяемся опытом ?! - Ресторанные хроники .] [Вернуться к списку .] Отправлено : Oleenka . , 28 Марта 2003 10 : 15 : 35 . В ответ на : а кто и когда придумал **вилку**? Руками конечно же , как еще ! : -) Вот тебе малюсенький исторический экскурс в историю **изобретения вилок** : Замечательное открытие <…> Найдено с помощью **www.yandex.ru**	1. Константин Мелихан - **Изобретение вилки** Сайт содержит: Тосты; Проведение Свадеб, Банкетов, Юбелеев; Организация застолья; Рецепты приготовления напитков; Алкогольные советы, Юмористические произведения, а также много интересного, **Изобретение вилки** 2. ИНДИВИДУАЛ Всякий прогресс, который нами понимается как прогресс сугубо материальный, научный, начался именно с **изобретения Вилки**. И только с **изобретением Вилки** и укоренением ее в быту, в сознании европейцев произошли существенные сдвиги.

Пример (4) Кто изобрел вилку?

Сириус	Яндекс
1.**http://www.koryazhma.ru/usefull/know/doc.asp?doc_**	1.http://korolenko.kharkov.com/m/viewdoc.pl?num=

id=91 ..... Кто **изобрел вилку** ? / Разное . **Вилка** как столовый прибор создавалась столетиями Кто **изобрел вилку**?... Найдено с помощью **www.yandex.ru** 2. **http://drink.dax.ru/avtor/melihan/melih_067.shtml** ..... Константин Мелихан - Изобретение **вилки** . А **вилку** он **изобрел** так. И он стал изобретать **вилку**.... Найдено с помощью **www.yandex.ru**	128155 Кто **изобрел вилку**?: 2. КТО **ИЗОБРЕЛ ВИЛКУ**? В эпоху Возрождения **вилки** были предметом роскоши, и поэтому их художественно обрабатывали. Черенки **вилок** и ножей изготавливали из серебра, золота, слоновой кости, дерева и украшали различными фигурками, головками, орнаментами.

Нелингвистические методы

Увеличение точности поиска в системе «Сириус» обеспечивается и другими – нелингвистическими методами. Так, с помощью некоторых вычислений осуществляется отбор наиболее релевантной запросу части текста. Для этого вычисляется значимость фрагментов текста. Введены два типа значимости: статическая значимость (значимость фрагмента просто как некоторой структурной единицы текста) и динамическая значимость (значимость, величина которой зависит от запроса, т.е. значимость для запроса). Или, по-другому, статическая значимость элемента текста (вес) - это величина, определяющая значимость фрагмента текста в зависимости от его типа (заголовок, подзаголовок и пр.). Динамическая значимость - это величина, определяющая значимость фрагмента текста в зависимости от его содержания.

Далее осуществляется ранжирование документов, являющихся результатом поиска, при этом учитывается значимость фрагментов текста документов, в которых найдены ключевые слова или другие релевантные запросу структуры. Документ в этом случае представляется как набор текстовых фрагментов различного типа, причем множество типов фрагментов упорядочено. Каждый тип фрагмента имеет вес, который учитывается при подсчете конечной релевантности. Например, документы, содержащие ключевые слова в своем названии, будут более значимы для пользователя по сравнению с другими, не имеющими ключевых слов в названии документами, при одинаковой релевантности по всем типам. Иными словами, ранг документов с релевантными запросу структурами, содержащимися в названии, должен повышаться.

Полнота поиска

Полнота поиска в метапоисковой системе «Сириус» достигается за счет:

1. Возможности расширения поискового запроса синонимами. Эта процедура предусматривает поиск в словарях, которые имеются в базе данных, синонимов к предикатным словам и именным группам и добавление найденных синонимов в запрос. Приведем пример.

Запрос: В районе станции Московского метрополитена Выхино обнаружили бомбу

Запрос после предобработки:
В & (районе | области | зоне) & (станции | перегона) & Московского & (метрополитена | метро) & Выхино & обнаружили & (бомбу | взрывное устройство)

2. Возможности расширения поискового запроса конверсивом (в случае если глагол-предикат является членом конверсивной пары). Эта процедура также осуществляется путем обращения к одному из словарей, в котором содержится список пар лексических конверсивов. Пример:

Запрос: где купить дешевые расходные материалы

Запрос после предобработки:
где (купить/продать) дешевые расходные материалы

3. Расширения возможностей пользователя при формулировке запроса, что заключается в а) возможности строить запрос в форме вопроса. В этом случае в ходе анализа устанавливается синтаксическое значение и синтаксическая функция вопросительного местоимения в предложении-запросе.

4. возможности выбора различных стратегий поиска (профилей поиска): от ключевого слова до целой ситуации. Поисковый профиль задается пользователем перед отправкой запроса на поисковую машину. В системе имеется 4 профиля:

1. "обычный" - как в обычном поисковике;
2. "поиск объекта " - применяется, когда пользователь хочет найти какой-то объект;
3. "факт" - применяется, когда пользователь хочет найти какой-то факт;
4. "ситуация" - применяется, когда пользователь хочет найти какую-то ситуацию.

В настоящее время реализована вторая версия исследовательского прототипа мета-поисковой системы «Сириус» [4]. Проводимые с прототипом эксперименты позволяют с высокой степенью достоверности сделать вывод о работоспособности описанных здесь методов. Работа в каждом из описанных направлений продолжается. Планируется расширить «область деятельности» лингвистического анализатора, за счет включения безглагольных предложений. Указанные работы позволят расширить область применимости системы и существенно повысить точность, качество и скорость поиска.

Список литературы:

1) Золотова Г.А., Онипенко Н.К., Сидорова М.Ю. Коммуникативная грамматика русского языка. М., 1998; 2 изд. – М., 2004.
2) Лингвистический энциклопедический словарь. Под. ред. Ярцевой В.Н. 2-е изд., доп., М.: Большая Российская Энциклопедия, 2002.
3) Осипов Г.С. Приобретение знаний интеллектуальными системами: Основы теории и технологии. – М.: Наука, Физматлит, 1997.
4) Осипов Г.С., Тихомиров И.А., Смирнов И.В. Интеллектуальный поиск в глобальных и локальный вычислительных сетях и базах данных.// Труды международной конференции "Программные системы: теория и приложения". - ИПС РАН, Переславль-Залесский 2004. т 2.