# Интеллектуальная поисковая система «Exactus»: Опыт участия в семинаре Ромип

Осипов Г. С., д.ф-м.н., проф. gos@isa.ru
Завьялова О.С., к.ф.н., olga_zavjalova@rambler.ru
Смирнов И.В., ivanv_smirnov@mail.ru
Тихомиров И.А., matandra@mail.ru

Институт системного анализа РАН
г. Москва

**Аннотация**

В статье содержится краткий отчет об участии интеллектуальной поисковой системы «Exactus» в семинаре РОМИП-2005. Главной целью участия в семинаре была апробация поисковых алгоритмов семантического, синтаксического и морфологического анализа с использованием неоднородных семантических сетей на достаточно больших коллекциях документов и тестовых запросах, расположенных на локальных носителях информации. Разработчикам системы было крайне интересно посмотреть на результаты тестирования системы с помощью методик РОМИП-2005.

## Введение

В процессе развития Интернета и роста объемов данных локальных вычислительных сетей возникает следующий парадокс: вероятность существования нужной информации растет, а возможность ее нахождения уменьшается. Надлежащий поиск необходимой информации становится общей проблемой. Требуются новые методы и инструментальные средства, решающие проблемы релевантного поиска в чрезвычайно больших объемах информации [1, 2].

Целью разработчиков всех поисковых систем является предоставить пользователю документы, в максимальной степени соответствующие смыслу запроса (обеспечить релевантность - точность

поиска), и вернуть как можно большее число документов, содержащих запрашиваемую информацию (обеспечить полноту поиска).

Эта цель в разных системах достигается разными способами. В интеллектуальной поисковой системе «Exactus» точность поиска достигается, во-первых, в результате использования метода семантико-синтаксического анализа, основанного на теоретических положениях коммуникативно-грамматической школы [3], а проблема полноты за счет привлечения методов метапоиска [2].

**Особенности интеллектуальной поисковой системы «Exactus»**

Основными рассматриваемыми в работе проблемами являются точность и полнота поиска. Под полнотой поиска будем понимать степень охвата информационных источников, которые могут содержать интересующую пользователя информацию (достаточно нетрадиционное определение). Под точностью – степень релевантности найденных по запросу пользователя документов.

Для повышения полноты разработана метапоисковая подсистема, позволяющая настраиваться на интерфейсы поисковых ресурсов и отправлять преобразованные запросы пользователя сразу на несколько поисковых машин или любые иные сайты.

Точность поиска повышается за счет последующей обработки информации и семантической фильтрации найденных документов.

Рассмотрим по шагам работу системы (см. рис. 1):


Рис. 1. Диаграмма потоков данных при поиске.

1. Пользователь вводит запрос на естественном языке.
2. Запрос подвергается лингвистическому анализу.

3. Производится выделение из запроса ключевых слов. При этом используется расширение полученного множества ключевых слов синонимами, из запроса выбрасываются стоп-слова и т.д.
4. Преобразованный таким образом запрос отправляется сразу на несколько информационных источников.
5. Отклики источников обрабатываются, из них выделяются тексты найденных документов и помещаются в полнотекстовую базу данных системы.
6. Исходный запрос и найденные документы подвергаются лингвистическому анализу, включающему морфологический, синтаксический и семантический анализ.
7. По результатам обработки проводится сравнение семантических образов запроса и найденных документов и вычисление релевантности.
8. Найденные документы фильтруются и сортируются в соответствие с вычисленной на предыдущем этапе релевантностью. Низкорелевантные документы отбрасываются.
9. Результаты поиска выдаются пользователю.

**Лингвистические методы повышения точности поиска**

Хорошо известно, что поиск только лишь по ключевым словам часто не удовлетворяет основному требованию пользователя, а именно требованию семантического соответствия найденных документов запросу.

Понятно, что одним из выходов из этой ситуации может служить более «наукоемкий» анализ, как запроса, так и документов.
Иначе говоря, следует рассматривать запрос не как последовательность слов, а как связный текст, разворачивающийся по законам языка. В интеллектуальной системе «Exactus» высокая точность поиска достигается в результате использования метода семантико-синтаксического анализа, основанного на принципах коммуникативно-грамматической школы [3] и методах описания концептуальной системы предметной области [4]. Покажем, в чем состоят отличительные особенности данного метода.

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”. Слово-лексема еще не является синтаксической единицей, слово – единица лексики, а в разных его формах могут реализоваться или актуализироваться раз-

ные стороны его общего 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 конструк-циях разной степени сложности. Эти единицы характеризуются всегда взаимодействием морфологических, семантических и функциональных признаков" [3]. 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.

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.

**Нелингвистические методы**

Увеличение точности поиска в системе «Exactus» 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 системе «Exactus» 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 за-проса, что заключается в:
   a. возможности строить запрос в форме вопроса. В этом случае в хо-де анализа устанавливается синтаксическое значение и синтаксическая функция вопросительного местоимения в предложении-запросе.
   b. 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.

**Архитектурные решения**

Exactus предлагает использование единой архитектуры и методики при поиске информации в гетерогенных источниках информации. Центральной задачей при разработке системы было создание набора инструментальных средств интеллектуального поиска в локальных и глобальных вычислительных сетях, а также базах данных.

Концептуально система состоит из нескольких компонентов, связанных друг с другом. Под компонентом понимается набор логически связанных модулей, имеющих общее назначение и представляющих собой законченную подсистему. Основное связующее звено компонентов системы – база данных, в которой централизованно хранится основная информация. Для данных, которые нецелесообразно хранить в реляционной БД, используются файловые хранилища. Система поддерживает параллельную обработку данных, при этом используется мультиагентная среда распределенных вычислений. Компонентная модель представлена на рис. 2.

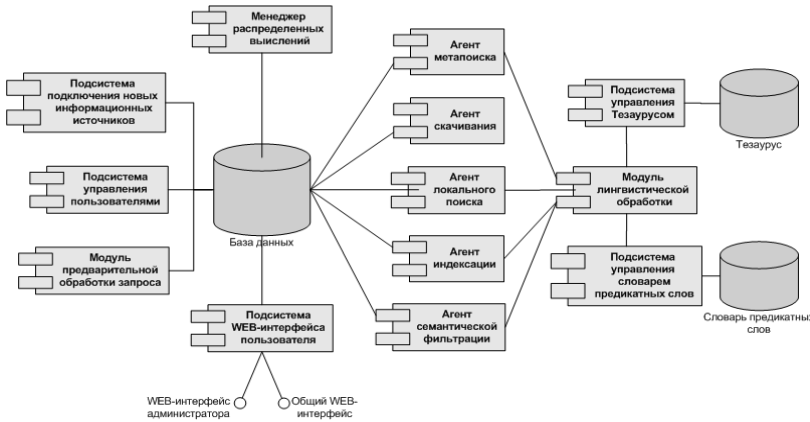


Рис. 2. Компонентная модель системы.

Любое действие в системе инициируется пользователем, для этого предусмотрено два интерфейса – Интерфейс администратора и Интерфейс пользователя. Под интерфейсом администратора понимается набор программного обеспечения, позволяющий управлять

системой и поддерживать ее основные функции, такие как пополнение и редактирование словарей, настройка на новые поисковые ресурсы, управление пользователями, управление настройками системы и т.д. Под интерфейсом пользователя понимается WEB-интерфейс, с помощью которого выполняется постановка задач на поиск информации в Интернет, а также просмотр и обработка полученных в результате поиска данных.

Прежде чем какая-либо задача будет исполнена, она попадает в очередь задач. Задачи исполняются параллельно несколькими агентами, причем обработка может проходить на нескольких компьютерах локальной сети. Каждый агент записывает результаты своей работы в базу данных, которые потом будут выданы пользователю, разумеется, в уже обработанном виде. Задачи могут выполняться несколькими агентами, причем различного класса, отработка агента может повлечь за собой постановку новых задач.

Ниже перечислены основные модули системы и их назначение:

- менеджер распределенных вычислений - обеспечивает функционирование мультиагентной среды параллельных вычислений и параллельную обработку данных;
- база данных - обеспечивает централизованное хранение данных;
- агенты - обеспечивают реализацию конкретных задач поиска;
- модуль управления пользователями - обеспечивает управление пользователями системы;
- модуль подключения новых информационных источников - обеспечивает подключение новых и управление уже существующими поисковыми ресурсами;
- модуль предварительной обработки запроса - осуществляет обработку запроса для повышения полноты и точности поиска;
- модуль управления тезаурусом - обеспечивает добавление новых словарей и управление уже существующими;
- модуль управления словарем предикатных слов - обеспечивает добавление новых предикатов и управление уже существующими;
- модуль лингвистической обработки - содержит набор модулей по лингвистической обработке тестов;
- подсистема WEB-интерфейса пользователя - позволяет пользователям работать с системой через WEB-браузер.

## Опыт участия в РОМИП

Этот раздел представляет собой «хронику» событий участия в семинаре.

### Модификация системы и выбор тестовой коллекции

«Exactus» является поисковой системой специально разработанной для поиска в гетерогенных источниках информации. В результате потребовались минимальные изменения в системе: в качестве внешнего источника данных был просто подключен дисковый массив – тестовая коллекция документов.

К сожалению, ввиду ограниченности материально-технической базы для тестирования пришлось выбрать самую минимальную по размеру тестовую коллекцию – по коллекции нормативно-правовых документов.

### «Жертва» полнотой поиска и скорость работы

«Exactus» не имеет собственного индексатора документов, поэтому перед поиском пришлось осуществить индексирование тестового массива документов с помощью Microsoft Indexing Service с подключенной к нему морфологией. Однако после индексирования и предварительных тестовых прогонов запросов оказалось, что алгоритм поиска Indexing Service, несмотря на все надстройки, слабо корелирует с поисковым алгоритмом «Exactus» и зачастую выдает релевантные документы в хвосте списка. Поэтому пришлось пожертвовать еще больше полнотой за счет динамического обрезания списка возвращаемых Indexing Service документов. Кроме того, пришлось отключить в предварительной обработке запроса его расширение синонимами. Явно в методики они не участвуют, однако увеличивают число возвращаемых документов в среднем в два-три раза, что сказывается на скорости работы системы. Таким образом, в жертву точности поиска была отдана его полнота.

В результате по факту время обработку 12 тыс. запросов на коллекции из 65 тыс. документов составило 120 часов на машине AMD Athlon 2800+. То есть среднее время составило около 1800 в байт текста в секунду, что при наличии в системе морфологического, синтаксического и семантического анализаторов, не так уж плохо.

**Результаты – немного критики в адрес методики тестирования РОМИП**

Сразу оговоримся, что о полноте поиска в результатах мы говорить не будем – ясно почему она не будет приемлемой, по крайней мере применительно к нашей системе.

Итак, точность поиска. Во-первых, следует отметить сам субъективизм понятия точности и численной оценки точности поиска. Во-вторых, если обратиться к методике тестирования точности поиска РОМИП и тому, сколько человек проводило тестирование, станет ясно, почему результаты могут оказаться столь неожиданными для участников. В методике не учтены такие вопросы, как:

1. Расширение запроса близкими по смыслу словами
   Как показывает практика, иногда бывает очень полезным расширять запрос близкими по смыслу словами. Ровно так же поступает и пользователь, когда не находит по запросу нужных документов.

2. Ситуативный поиск
   Поиск давно уже вышел за рамки того, чтобы просто искать документы по набору ключевых слов. Нужны новые методы поиска, адаптированные под конкретные задачи. Например, пользователя может интересовать ситуация, когда «кто-то где-то что-то взорвал». То есть ситуацией близкой к «шахидка в Московском Метро взорвала взрывное устройство» будет «на Тайване опять взорвали бомбу». Искать и оценивать, опираясь только на ключевые слова тут просто бессмысленно.

3. Вопрос-ответ
   Рассмотрим также проблему вопросно-ответного режима поиска, когда пользователь задает системе вопрос, а она возвращает ему список найденных документов с ключевыми словами, зачастую имеющими мало общего с ключевыми словами запроса, но высокорелевантными. К сожалению, подобный режим никак не представлен на РОМИП.

4. Осмысленность запроса
   И главная проблема, на наш взгляд, всех запросов к коллекции нормативно-правовых документов – осмысленность поставленных перед системой запросов. При неосмысленном запросе оценивать точность поиска эксперту очень сложно, даже если он является экспертом предметной области. Если

требуется действительно точный поиск нужно более точно формулировать запрос, не указывая только лишь фрагмент или топик документа.

5. Экспертная оценка – есть ли альтернатива?
   К сожалению, на наш взгляд, альтернативы экспертной оценке поисковых алгоритмов не существует. И тем более точная будет экспертная оценка, чем больше экспертов из разных групп пользователей будет привлечено к этому процессу. На наш взгляд, РОМИП страдает из-за малого количества подобных экспертов.

Мы искренне надеемся, что наши пожелания будут учтены в последующих методиках тестирования РОМИП, что только положительным образом скажется на результатах и общем уровне семинара РОМИП.

## Заключение

В заключении хотелось бы выразить огромную признательность организационному комитету РОМИП. Этот семинар единственный в России дает возможность (мы не говорим слов объективно и пр.) оценить результаты работы поисковых машин и на наш взгляд у этого семинара большое будущее.

## Литература


[1] Осипов Г.С., Куршев Е.П., Кормалев Д.А., Трофимов И.В., Рябков О.В., Тихомиров И.А. Семантический поиск в среде Интернет. // Препринт. – Переславль-Залесский: ИПС РАН, 2003.

[2] 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. т2.

[3] Золотова Г.А., Онипенко Н.К., Сидорова М.Ю. Коммуникативная грамматика русского языка. – М., 2004.

[4] G. S. Osipov. Semantic Types of Natural Language Statements. A Method of Representation. //10th IEEE International Symposium on Intelligent Control Monterey, California, USA, Aug. 1995.