ISBN 5-94052-066-0 ПРОГРАММНЫЕ СИСТЕМЫ: ТЕОРИЯ И ПРИЛОЖЕНИЯ. ПЕРЕСЛАВЛЬ-ЗАЛЕССКИЙ, 2004

УДК 519.767.6

Г. С. Осипов, И. А. Тихомиров, И. В. Смирнов

Интеллектуальный поиск в глобальных и локальных вычислительных сетях и базах данных

Аннотация. Статья посвящена методам и инструментальным средствам семантически релевантного метапоиска. Рассматриваются задачи применения описываемых методов для поиска в глобальных и локальных вычислительных сетях и базах данных.

Ключевые слова и фразы: метапоиск, семантический поиск, семантический анализ.

Введение

В настоящее время в связи с тенденцией интеграции локальных сетей с глобальными, а также ростом объемов информации сетевых ресурсов и баз данных, существенно возросла важность проблемы релевантного поиска в источниках различного вида. Однако, хорошо известно, что применяемые в существующих системах методы не позволяют достичь высокой полноты и точности поиска. Одной из причин является узкая специализация систем поиска, которые не позволяют решать широкий спектр задач поиска одновременно в нескольких информационных источниках, таких как ресурсы глобальных и локальных вычислительных сетей, базы данных, локальные документы на жестких дисках и т. д.

В большинстве случаев полнота поиска увеличивается за счет непрерывного мониторирования сети или базы данных с целью выявления новых документов. Иной путь — создание метапоисковых систем [1, 2], позволяющих объединять результаты поиска из различных источников.

Проблема точности традиционно решается на пути использования линейного поиска по ключевым словам с привлечением некоторых лингвистических методов. Ряд систем декларирует возможности

Работа выполнена при поддержке программы фундаментальных исследований ОИТВС РАН «Фундаментальные основы информационных технологий и систем», госконтракт №10002-251/ОИВТС-01/097-111/210503-178 от 6 мая 2003 г.

семантического поиска, ввода запросов на естественном языке, ответов на вопросы пользователя, однако использует для достижения декларируемых целей неадекватные лингвистические и программные средства. Результатом работы таких систем является достаточно большой массив документов, из которых в действительности релевантными являются очень немногие.

В настоящее время в России и за рубежом разработано несколько промышленных систем и прототипов для поиска информации [3, 4]. Среди отечественных работ в этой области наиболее интересной и функциональной является поисковая машина Russian Context Server от компании «Гарант-Парк-Интернет». Поисковая машина Russian Context Server служит для поиска текстовой информации на корпоративном узле Интернет или Интранет. Продукт представляет собой поисковую систему, обладающую возможностями как контекстного, так и реляционного поиска. Russian Context позволяет искать документы с учетом морфологии русского и английского языков, используя SQL-подобный язык запросов и комбинируя поисковые ограничения на контекст с ограничениями на заданные атрибуты документов. Russian Context Server использует компоненты из серии, выпускаемой компанией «Гарант-Парк-Интернет» под маркой RCOTM (Russian Context Optimizer), среди них:

- библиотека морфологического анализа текста RCO Morphology;
- библиотека выделения объектов в тексте RCO Pattern Extractor;
- библиотека работы с поисковым тезаурусом RCO Thesaurus Search;
- библиотека синтаксического анализа текста RCO Syntactic Engine;
- библиотека статистического анализа текста RCO Semantic Network.

Среди зарубежных систем можно отметить промышленную систему Convera RetrievalWare, продукт компании Convera Technologies Corp. Продукт интересен тем, что был локализован отечественной фирмой «Весть-МетаТехнология» и в результате появился Русский Семантический Сервер, представляющий собой совокупность программных средств и информационных ресурсов, позволяющих осуществлять полнотекстовый поиск с учетом специфики русского языка. Среди особенностей RetrievalWare можно отметить:

- технологию адаптивного распознавания образов APRP основанную на нейронных сетях, обеспечивающую нечеткий поиск текстовой информации (исчисление меры близости слов запроса и документа), высокую скорость, точность и полноту поиска, языковую независимость, малые объемы индексных файлов;
- технологию поиска, основанную на семантических сетях и ориентированную на работу с понятиями, содержащимися в текстовых документах;
- поддержку русской морфологии;
- возможность одновременного использования нескольких семантических сетей;
- возможность поиска документов по образцу — при этом система сама выбирает из документа наиболее статистически значимые слова и формирует из них сложный логический запрос, учитывающий и структуру, и содержание документа;

Настоящая работа посвящена изложению подходов и описанию программных средств поиска информации, в которых проблема полноты решается на пути привлечения методов метапоиска [2]. Проблема точности — благодаря использованию солидного арсенала лингвистических средств, в частности, методов морфологического, синтаксического и поверхностного семантического анализа [5]. Вторая особенность описываемых методов — возможность написания запросов на естественном языке.

1. Особенности предлагаемых подходов

Предлагается использование единой системной архитектуры и методики при поиске информации из различных информационных источников. Центральными задачами настоящей работы являются существенное улучшение характеристик поиска и создание набора инструментальных средств интеллектуального поиска в локальных и глобальных вычислительных сетях, а также базах данных. Принципиальная схема поиска показана на рисунке ниже.

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. Под точностью — степень релевантности найденных по запросу пользователя документов.

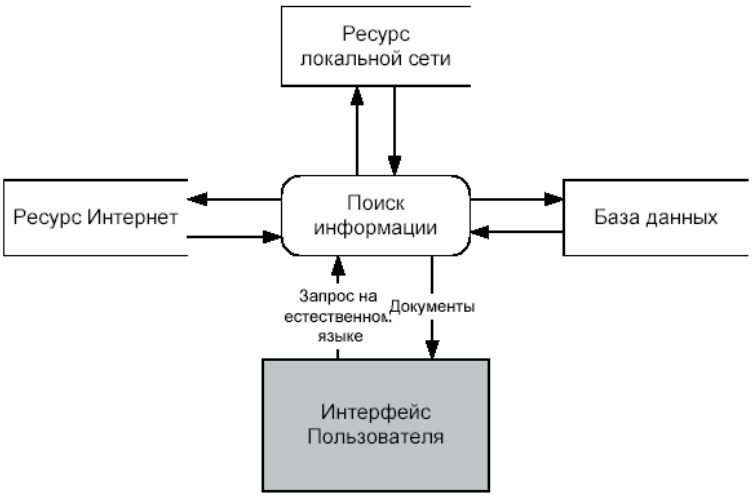

Рис. 1. Схема поиска информации из различных информационных источников

Для повышения полноты поиска разработана метапоисковая система, которая позволяет настраиваться на интерфейсы поисковых ресурсов и отправлять преобразованные запросы пользователя сразу на несколько поисковых машин или любые иные сайты. Разработаны специализированные средства поиска в локальных сетях и базах данных.

Точность поиска повышается за счет последующей обработки информации и семантической фильтрации найденных документов. Сказанное можно продемонстрировать на следующей схеме.

Опишем принципиальный алгоритм работы системы:

(1) пользователь выбирает тип запроса («в сети Интернет», «в локальной БД», «в локальной сети») и вводит поисковый запрос на естественном языке;
(2) запрос обрабатывается, из него извлекаются ключевые слова. При этом используется расширение запроса за счет использования словаря синонимов (используется словарь синонимичных предикатов и словарь синонимичных именных групп), из запроса выбрасываются стоп-слова и т. д.;

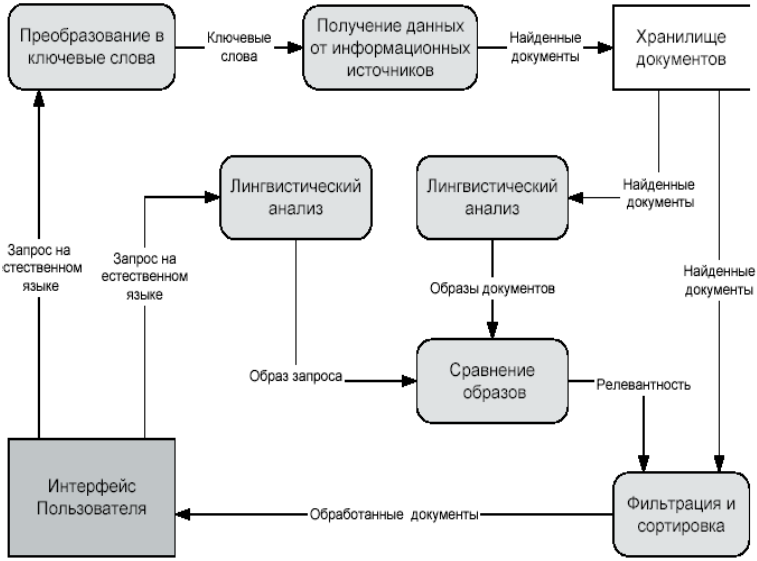

Рис. 2. Схема работы системы поиска

(3) преобразованный таким образом запрос отправляется либо на несколько поисковых машин Интернет (например, на Яндекс или Рамблер) или в локальную БД (поиск осуществляется среди помещенных в ней документов), либо выполняется в папках локальной сети (в данный момент поддерживаются такие форматы, как HTML, TXT, MS Word, MS Excel и MS PowerPoint);

(4) найденные документы обрабатываются и помещаются полнотекстовую базу данных системы;

(5) запрос пользователя и найденные документы подвергаются лингвистическому анализу, включающему морфологический, синтаксический и поверхностный семантический анализ, строятся семантические образы запроса и документов, проводится сравнение образов и вычисление семантической релевантности найденных документов запросу пользователя;

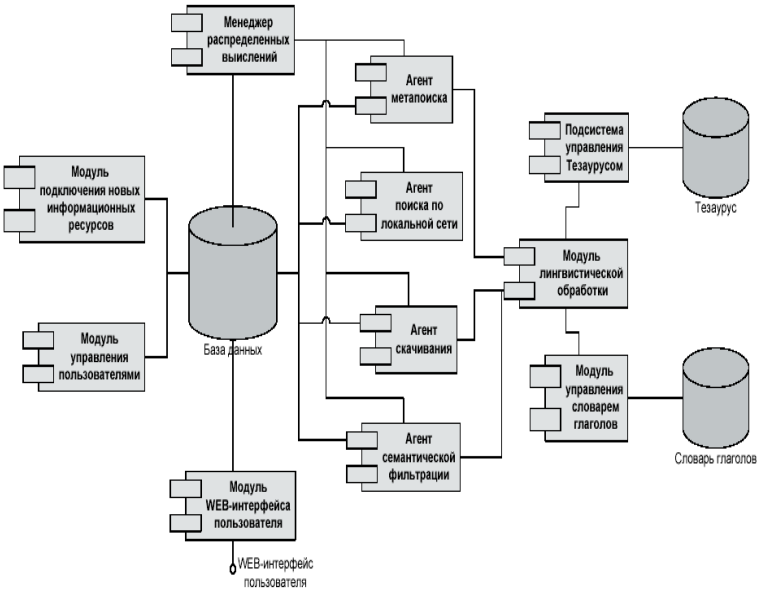

Рис. 3. Компонентная модель системы поиска

(6) найденные документы сортируются в соответствии с вычисленной релевантностью. Низкорелевантные документы отбрасываются.

2. Архитектура и программные средства системы

Концептуально, система состоит из нескольких компонентов, связанных друг с другом [3]. 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 рисунке ниже.

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. Далее остановимся на основных модулях системы и их назначении более подробно.

2.1. Агент метапоиска. Основной задачей агента является выполнение поискового запроса к ресурсам сети Интернет и получение ссылок на найденные документы. В основе алгоритма работы агента лежит предположение о том, что любой ресурс можно описать при помощи некой структуры в терминах тэгов HTML [7]. Указанная структура заполняется в полуавтоматическом режиме при помощи модуля подключения новых информационных источников [6]. Среди вспомогательных задач агента — извлечение дополнительной информации из структуры описания поискового ресурса, эта функциональность позволяет инкапсулировать в агенте всю логику по разбору структуры описания поискового ресурса.

2.2. Агент поиска по локальной сети. Агент выполняеет поиск информации в локальных сетях. Поиск осуществляется по проиндексированным каталогам при помощи службы Microsoft Indexing Service. Агент использует специализированный компонент для извлечения текстов из найденных документов, который позволяет понимать такие форматы, как HTML, TXT, MS Word, MS Excel и MS PowerPoint.

2.3. Агент скачивания. Среди главных задач агента — загрузка документа по ссылке (URL) из сети Интернет, используя протокол HTTP, выделение текста документа и преобразование его в кодировку Windows 1251. Скачанные документы сохраняются в локальной БД системы для последующей обработки.

2.4. Агент семантической фильтрации. Основными задачами агента семантической фильтрации являются оценка семантической близости запроса и документа (в процентах) и упорядочивание результирующего набора документов в соответствии с этой оценкой (документы с большим процентом семантической релевантности показываются в первую очередь) [5]. Релевантность найденных документов оценивается по трем параметрам:

- **семантическим падежам (ролям).** Тексты документов разрезаются на фрагменты, содержащие ключевые слова запроса. Фрагменты передаются для обработки модулю семантического анализа, который строит их поисковый образ. Поисковый образ — это индекс пар роль, именная синтаксема; именная синтаксема = предлог + падеж следующего существительного [8, 9]. Такая же процедура применяется к запросу. Далее выполняется сравнение семантического образа запроса с семантическими образами фрагментов документа. Релевантность в этом случае есть оценка найденных в образе документа пар из образа запроса;
- **семантическим связям.** Тексты документов разрезаются на фрагменты, содержащие ключевые слова запроса. Фрагменты передаются для обработки модулю семантического анализа, который строит их поисковый образ. Поисковый образ — это индекс троек тип семантической связи, 1-я синтаксема, 2-я синтаксема. Далее сравнивается семантический образ запроса с образами фрагментов документа. Релевантность по связям есть оценка найденных в образе документа троек из образа запроса;
- **ключевым словам.** В документах осуществляется поиск деревьев, растущих от существительных запроса. Синтаксема запроса должна целиком содержаться в синтаксеме документа, релевантность документов по ключевым словам есть процент найденных в документе синтаксем запроса.

2.5. Модуль подключения новых информационных ресурсов. Информационные ресурсы необходимы для первичного поиска информации по ключевым словам, которая в дальнейшем будет отфильтрована и предоставлена пользователю [7]. Информационными источниками являются любые поисковые системы, включая поисковые машины, типа yandex.ru, rambler.ru, altavista.com, Интернет-каталоги, информационные и новостные серверы, Интернет-порталы и любые другие Интернет-ресурсы, имеющие возможности поиска по ключевым словам.

Подключение новых информационных источников позволяет расширять область поиска в Интернете, увеличивая вероятность нахождения необходимого пользователю ресурса. Для облегчения и ускорения подключения новых поисковых ресурсов создан модуль, позволяющий добавлять новые информационные источники в полуавтоматическом режиме [6]. После указания адреса поискового ресурса пользователь в визуальном режиме указывает необходимые для создания запроса элементы управления, чекбоксы, строки редактирования, кнопки и др. На втором этапе, после выполнения фиктивного запроса, пользователь, также в визуальном режиме, указывает необходимые поля в результате запроса в нескольких записях, например, URL, дату, автора, размер и другие, найденного ресурса. По указанным примерам модуль распознает формат представления результата запроса и создает соответствующий скрипт. Созданный таким образом скрипт будет использоваться для извлечения данных о найденных ресурсах.

2.6. Модуль лингвистической обработки. Основной задачей модуля является построение семантического образа текста [5]. Среди прочей информации модуль выявляет, в частности, именные группы и предикатные слова (глаголы и отглагольные формы). Семантические связи между именными группами устанавливаются в результате частичного семантического анализа и основываются на моделях управления найденных предикатных слов. Виды семантических отношений, а также грамматические признаки, позволяющие обнаружить их в тексте документа, отражены в справочнике предикатных слов. Интерфейсы модуля лингвистической обработки, предназначенные для анализа текста, разделяются на две группы: интерфейсы анализаторов и интерфейсы лингвистических данных. Входные и выходные данные не связаны с анализаторами, что позволяет

отделить логику обработки текста от извлекаемой из текста информации. Семантическая обработка текста выполняется в три этапа: морфологический, синтаксический и собственно семантический анализ. Каждый этап выполняет соответствующий анализатор со своими входными и выходными данными и собственными настройками. При этом функции морфологического и синтаксического анализа реализованы в отдельном ActiveX компоненте. Частичный семантический анализ сводится к обработке всех найденных предикатных слов. Для каждого такого слова в пределах сегмента предложения ищутся синтаксически связанные с ним именные группы. Для каждой именной группы делается попытка выяснить, заполняет ли она некоторую валентность (роль) управляющего предикатного слова. При этом формируется запрос к словарю предикатных слов, в котором передается идентификатор предикатного слова, предлог (если он есть) и грамматические характеристики корневого элемента именной группы. В заключение среди заполненных валентностей в словаре предикатных слов ищутся пары, образующие семантические отношения. Результатом работы модуля является семантический образ текста в виде списка семантических ролей и списка семантических отношений. Каждый элемент списка ролей определяет тип роли для одной именной группы текста. Каждый элемент списка отношений включает в себя тип отношения и две связанные данным отношением именные группы. Для быстрого поиска списки проиндексированы по типам соответственно ролей или отношений, а также по частям речи и словарным формам корневых слов именных групп.

2.7. Модуль управления словарем глаголов. Модуль позволяет вносить изменения в словарь глаголов, являющийся одним из главных составляющих системы, а также осуществлять доступ к словарю другим модулям для последующей обработки и использования полученных данных. В словаре глаголов описаны способы синтаксической реализации в тексте различных типов смысловых отношений между понятиями. Словарь представляет собой текстовый файл со списком предикатных слов русского языка (глаголов в прямых и возвратных формах и отглагольных существительных), отражающих некоторую ситуацию. Словарная статья описывает семантические падежи (роли) участников этой ситуации и способы их выражения в тексте, а также различные семантические связи между участниками ситуации [6, 8, 9].

2.8. Модуль управления словарем синонимов. Задачей модуля является управление словарем синонимов и реализация специфических алгоритмов расширения запросов за счет использования словаря синонимов. Алгоритм расширения запроса синонимами можно представить в виде рекурсивной формулы:

$$Query_i = (Word_i \ \& \ Syn(Word_i)) \ \& \ Query_{i+1} \vee Syn(NG),$$

где $Query_i$ — запрос на очередном шаге, $Word_i$ — очередное слово, $Syn()$ — функция добавления синонимов, NG — именная группа для очередного слова.

Действие этих правил лучше всего продемонстрировать на примере. Рассмотрим фразу «Речь президента Российской Федерации»:

(1) корень дерева — слово «речь». По правилу оно должно входить в запрос, значит Query=речь. Дальше ищем синонимы для слова «речь». Предположим, нашли «выступление». Добавляем через связку $\vee$, группируем: Query[1] = (речь $\vee$ выступление). Переходим к потомкам;

(2) слово «президент». Синоним для него — «глава». Добавляем в запрос: Query[2] = (президент $\vee$ глава). Обрабатываем потомков рекурсивно. (пропустим описание этого шага) Получаем Query[2] = ((президент $\vee$ глава) & ((российский) & (федерация)) $\vee$ (Россия)). Нормализуем поддерево «президента», получаем «президент российской Федерации» и находим синоним «Путин». Тогда Query[2] = (((президент $\vee$ глава) & (((российский) & (федерация)) $\vee$ (Россия)) $\vee$ Путин);

(3) далее мы возвратимся на шаг 1, так как слово имеет потомков, нормализуем группу и получаем «речь президента Российской Федерации». Ищем для нее синонимы. Не находим, тогда запрос будет таким: Query[1] = (речь $\vee$ выступление) & Query[2] = (речь $\vee$ выступление) & (((президент $\vee$ глава) & (((российский) & (федерация)) $\vee$ (Россия)) $\vee$ Путин).

Таким образом, по запросу будут найдены документы, содержащие следующие варианты слов:

- речь президента Российской Федерации;
- выступление президента Российской Федерации;
- речь главы Российской Федерации;

- выступление главы Российской Федерации;
- речь президента России;
- речь главы России;
- выступление президента России;
- выступление главы России;
- речь Путина;
- выступление Путина.

Использование алгоритма, идея которого была изложена выше, позволяет значительно повысить полноту поиска.

2.9. Модуль WEB-интерфейса пользователя. Модуль позволяет пользователю управлять системой через WEB-браузер, создавать иерархию тем запросов, ставить задачи поиска информации в Интернет, локальной сети и базе данных, просматривать результаты запросов. Интерфейс системы строится по принципу минимального количества действий пользователя для осуществления типовых операций. Страницы системы однотипные и строятся по одинаковой схеме, типовая страница состоит из заголовка, в который входит тематическое меню, левой части с функциональным меню, рабочей области и подвала страницы. Тематическое меню предназначено для перехода по страницам различного назначения, например, результатов поиска, страницы настроек или помощи. Функциональное меню представляет собой иерархическое дерево тем и запросов. Каждой теме и запросу соответствует набор управляющих иконок — действий, которые можно совершать над данным элементом. В рабочей области страницы располагается основная часть динамического контента, например, результаты поиска, форма редактирования или справка системы. Подвал страницы статический и содержит информацию о разработчиках системы.

3. Результаты и дальнейшие работы

В настоящее время реализована вторая версия исследовательского прототипа системы, позволяющая осуществлять интеллектуальный поиск в сети Интернет, локальных сетях и базах данных [3]. Прототип включает в себя мощные средства лингвистического анализа текстов, включая средства морфологического, синтаксического и семантического анализа, которые могут быть использованы отдельно

друг от друга в различных приложениях, в том числе и коммерческих. В рамках прототипа разработана общая расширяемая архитектура системы, которая позволяет наращивать функциональные характеристики системы, расширяя тем самым области ее возможного применения. Уже в данный момент систему можно использовать в таких областях, как:

- поиск в Интернет/Интранет/Экстранет средах;
- поиск в локальных массивах документов;
- поиск в базах данных;
- использование в качестве комбинированной корпоративной поисковой системы, объединяющей все выше перечисленное.

Заключение

Проведенные эксперименты позволяют с высокой степенью достоверности сделать вывод о работоспособности описанных здесь методов. Они же позволили сформулировать дальнейшие направления работ. Среди них — совершенствование алгоритмов настройки на поисковые ресурсы и повышение скорости работы, а именно — временных параметров и качества семантической фильтрации. Весьма перспективно совершенствование методов расширения естественно-языковых запросов синонимической и иной семантически эквивалентной информацией, что должно повысить полноту поиска. Предполагается, кроме того, разработка новых алгоритмов распознавания и анализа различных моделей предложения, параметрических алгоритмов вычисления релевантности, алгоритмов машинного обучения, адаптированных к работе с текстовой информацией и пополнение словарей системы. Указанные работы позволят расширить область применимости системы и существенно повысить точность, качество и скорость поиска, что позволит ставить вопрос о реализации промышленного прототипа системы.

Список литературы

[1] Козлов Е. Б., Метелкин А. В., Хорошевский В. Ф. *Мультиагентная система поиска информации в Интернет* // Труды седьмой национальной конференции по искусственному интеллекту с международным участием КИИ'2000. — М.: Физматлит, 2000, с. 840–850. ↑(document)

[2] Куршев Е. П., Осипов Г. С., Рябков О. В., Самбу Е. И., Соловьева Н. В., Трофимов И. В. *Интеллектуальная метапоисковая система* // Труды международного семинара Диалог'2002 «Компьютерная лингвистика и интеллектуальные технологии». — М.: Наука, 2002, с. 320–330. ↑(document)

[3] Осипов Г. С., Куршев Е. П., Кормалев Д. А., Трофимов И. В., Рябков О. В., Тихомиров И. А.: Препринт // Семантический поиск в среде интернет. — Переславль-Залесский, ИПС РАН, 2003. ↑(document), 2, 3

[4] Ермаков А. Е. *Эксплицирование элементов смысла текста средствами синтаксического анализа-синтеза* // Компьютерная лингвистика и интеллектуальные технологии: труды Международной конференции Диалог'2003. — М.: Наука, 2003. ↑(document)

[5] Кормалев Д. А., Куршев Е. П., Осипов Г. С., Сулейманова Е. А., Трофимов И. В.: Препринт // Методы поиска и анализа информации. Автоматическое извлечение данных. — Переславль-Залесский, ИПС РАН, 2003. ↑(document), 2.4, 2.6

[6] Осипов Г. С. Приобретение знаний интеллектуальными системами: Основы теории и технологии. — М.: Наука, Физматлит, 1997. ↑2.1, 2.5, 2.7

[7] Куршев Е. П. *Метод извлечения полуструктурированных данных из Интернет* // Труды седьмой национальной конференции по искусственному интеллекту с международным участием КИИ'2000. — М.: Физматлит, 2000, с. 260–263. ↑2.1, 2.5

[8] Золотова Г. А. Синтаксический словарь. Репертуар элементарных единиц русского синтаксиса. — М.: Эдиториал, 2001. ↑2.4, 2.7

[9] Золотова Г. А., Онипенко Н. К., Сидорова М. Ю. Коммуникативная грамматика русского языка. — М.: МГУ, 1998. ↑2.4, 2.7

Исследовательский центр искусственного интеллекта ИПС РАН

Российский Университет Дружбы Народов РУДН

G. S. Osipov, I. A. Tikhomirov, I. V. Smirnov. *Intelligent Search in Global and Local Area Networks and Databases.* (in russian.)

Abstract. The article outlines methods and tools for semantic metasearch. Efforts are focused on applying these methods to perform search in Global and Local Area Networks and Databases.