УДК 519.21:681.142

АВТОМАТИЗИРОВАННЫЙ ПОИСК КОММЕРЧЕСКОЙ ИНФОРМАЦИИ ДЛЯ РЕШЕНИЯ ЗАДАЧ ОРГАНИЗАЦИИ ВИРТУАЛЬНЫХ ПРЕДПРИЯТИЙ

П.А.Помилуйко

МНУЦ ИТС АН и МОН Украины
pomiluyko@gmail.com

Стаття присвячена опису підходів та архітектурних рішень, які використовуються при створенні засобів для систем пошуку, вилучення комерційної інформації з неструктурованих колекцій документів та організації віртуальних підприємств.
Ключові слова: пошукова система, вилучення інформації, аналіз тексту, організація віртуальних підприємств.

This article describes approaches and architectural solutions are used in the creation of search systems. business information retrieval from unstructured collections and virtual enterprises organization.
Key words: search engine, information retrieval, text analysis, organization of virtual enterprises.

Статья посвящена описанию подходов и архитектурных решений, используемых при создании средств для систем поиска, извлечения коммерческой информации из неструктурированных коллекций документов и организации виртуальных предприятий.

Ключевые слова: поисковая система, извлечение информации, анализ текста, организация виртуальных предприятий.

1. Актуальность задачи

Объемы обрабатываемой информации нарастают по экспоненте – этому способствует активное внедрение мультимедиа, широкое распространение корпоративных и глобальных сетей, уход большинства предприятий от бумажного документооборота и переход на автоматизированные системы управления предприятием, в особенности и в сети Интернет. Проблема поиска и извлечения информации из неструктурированных коллекций документов актуальна и востребована.

Алгоритм поиска информации типичной поисковой системой [11], включает в себя следующие этапы:

- Формализация пользователем поискового запроса (представление пользователем, в том или ином виде, своих информационных потребностей).
- Предварительный отбор документов по формальным признакам наличия интересующей информации (например, наличие в тексте документа одного из слов запроса, если запрос формулируется на естественном языке).
- Анализ отобранных документов (лингвистический, статистический).
- Оценка соответствия смыслового содержания найденной информации требованиям поискового запроса (ранжирование).

Для того чтобы систематизировать и классифицировать коллекцию документов (в нашем случае информацию о предприятиях) необходимо не только найти данные, но и грамотно их извлечь из результатов поиска, выполнить анализ, а также проверить и следить за актуальностью содержания.

Среди типичных подзадач [4, 12] извлечения информации выделим следующие:

- Распознавание именованных элементов: распознавание имён людей, названий организаций, мест, временных обозначений и некоторых типов численных выражений.
- Ссылки: выделение словесных оборотов, ссылающихся на один и тот же объект. Типичный случай таких ссылок — анафора и использование местоимений.
- Выделение терминологии: нахождение для данного текста ключевых слов.

К сожалению, пока ни одна поисковая система не умеет точно и правильно извлекать информацию.

Технологии Интернет делают возможным массовое развертывание приложений для управления знаниями и управления потоками работ. Раньше эти технологии внедрялись в рамках системы клиент-сервер, предназначенной для тщательно отобранных групп пользователей, ограничивавшихся сотрудниками предприятия. Главными причинами таких ограничений являются стоимость установки и сопровождения клиентских приложений на каждой рабочей станции и узкоспециализированные протоколы, используемые этими приложениями. Сегодня же, будучи установлено на сервере, приложение, поддерживающее функции Интернет, становится потенциально доступным с любой рабочей станции. Это открывает путь к использованию интеллектуальных и производственных инструментов управления, почти всем сотрудникам предприятия, причем при минимальных затратах. Одно и то же приложение становится доступно с любой рабочей станции, подключенной к Интернет. Это имеет критическое значение для приложений электронной коммерции. Это необходимо, чтобы открыть доступ к приложениям корпоративным партнерам, что является важнейшей предпосылкой для создания виртуальных предприятий. Решение проблемы организации виртуальных предприятий откроет дорогу для предпринимателей малого и среднего бизнеса в эпоху автоматизации управления предприятием. Это поможет значительно оптимизировать работу предприятия, минимизировать издержки и иметь твердую позицию в конкурентной борьбе.

2. Анализ поисковых систем

Даже всемирно известные популярные программы поиска (dtSearch Desktop – официальный сайт: www.dtsearch.com, Ищейка Проф Deluxe – официальный сайт на русском языке: www.isleuthhound.com/ru, Google Desktop Search: www.google.com, SearchInform: www.searchinform.com, Copernic

Desktop Search: www.copernic.com, ISYS Desktop: www.isys-search.com), которые обладают как приличными скоростями, так и неплохим функционалом, не могут похвастаться возможностью извлечения коммерческой информации из результатов поиска. Лидером среди них можно смело назвать Google Desktop Search, который уже научился фильтровать результаты по основным направлениям: новости, книги, обсуждения, товары и цены на них.

К сожалению, в Украине также не существует специализированных поисковых систем коммерческой информации. Среди популярных поисковых каталогов стоит отметить http://all-biz.info, http://prom.ua и http://ukr-firms.com.ua. Общим недостатком этих коммерческих каталогов является ручная модерация и отсутствие автоматического обновления.

3. Общая проблема поиска и извлечения информации

В общем, модели поиска можно разделить на два больших класса [11]:

- Поисковые каталоги
- Поисковые системы

Поисковые каталоги в большей степени ориентированы на структурную организацию тематических коллекций с удобной системой ссылок и иерархией документов по тематическим коллекциям. Это позволяет пользователю самостоятельно находить требуемый документ, просматривая структуру каталога, либо использовать механизмы поиска ориентированные на данный каталог. Основная проблема и недостаток такого варианта поиска - это необходимость выполнения значительного объема работ по предварительной организации, наполнению каталога. Как правило, это ручная классификация на основе привлечения экспертов.

Поисковые системы ориентированы на поиск неструктурированной информации. Как правило, они используются для поиска документов в больших информационных коллекциях. Особенностью таких коллекций является отсутствие четко выраженной структуры, позволяющей упорядочить и классифицировать документы по тематической направленности. Основная проблема такого поиска – это сложность интерпретации содержания текстов документов и формулировки потребности пользователей. Эта проблема актуальна особенно для поиска коммерческой информации.

Ранние информационно-поисковые системы и методы поиска разрабатывались и тестировались на относительно небольших, однородных коллекциях документов. Современные условия поиска и, соответственно, требования к информационно-поисковым системам претерпели значительные изменения [10]. Главным образом, эти условия и требования связаны с развитием Интернета, который имеет свои специфические черты и особенности. Рассмотрим эти особенности:

- **Размер**. Одной из главных особенностей Интернета является огромный объем доступных информационных ресурсов, продолжающий, к тому же,

интенсивно нарастать. По оценкам специалистов, уже сейчас в Интернете содержится более миллиарда страниц, общий размер этих страниц оценивается в терабайтах. В связи с этим возникают высокие требования к масштабируемости используемых алгоритмов поиска.

- **Динамика**. Высокая степень обновления информационных ресурсов Интернета. Очень часто появляются новые и удаляются существующие страницы, меняется их местоположение. Статистика показывает, что среднее время жизни половины страниц в Интернете не превышает десяти дней, ежемесячно примерно 40% страниц подвергается изменениям, а объем всей информации в сети увеличился в два раза за последние два года. Данная особенность значительно затрудняет использование общих статистических характеристик коллекции.
- **Взаимосвязи**. Одной из особенностей информационного пространства Интернета является то, что страницы взаимосвязаны между собой. Эта взаимосвязь реализуется с помощью гиперссылок, что может быть использовано при реализации некоторых методов поиска.
- **Свободная публикация**. В Интернете возможно свободное размещение документов и их удаление из коллекции, т.к. отсутствует централизированное администрирование информационных ресурсов. Вследствие этого могут быть нарушения целостности отдельных документов коллекции и связей между ними.
- **Избыточность**. Для Интернета характерна большая избыточность информационных ресурсов. Очень часто на разных страницах публикуется несколько копий одного и того же документа или его незначительно модифицированных версий. Исследования показывают, что около 30% информации в Интернете - это точные или приблизительные копии других документов.
- **Неконтролируемое качество**. Возможность свободной публикации документов в Интернете, а также отсутствие какой-либо обязательной проверки их содержания зачастую приводит к появлению недостоверной и ошибочной информации, содержащей многочисленные орфографические и грамматические ошибки, опечатки, ошибки, вызванные оцифровкой документов, и просто некорректные и непроверенные данные.
- **Пользователи**. Интернет объединяет многочисленные группы совершенно разных по квалификации и подготовке пользователей. Многие из них не умеют грамотно и эффективно формулировать запросы. Статистика показывает, что более 60% поисковых запросов в Интернете состоят из 1-2 слов, для примера, в классических информационно-поисковых системах эта величина 7-9 слов. Зачастую это приводит к большому количеству обрабатываемых и анализируемых в результате поиска документов. Сами результаты поисков в этом случае могут быть весьма далекими от реальных информационных запросов пользователя,

т.к. запрос очень короткий. Исследования поведения пользователей показали, что многие из них не готовы к продолжительному ожиданию результатов поиска и анализу результирующего множества для выявления необходимых документов. 58% пользователей ограничиваются изучением первого экрана результатов запроса, 67% не пытаются модифицировать свой первоначальный запрос. При этом критерии качества, используемые в традиционных системах текстового поиска, становятся неадекватными, например, критерий полноты поиска, т.е. процент обнаруженных релевантных документов.

- **Доступ**. Не всегда возможен доступ к информационным ресурсам Интернета, т.к. далеко не все сервера работают круглосуточно в течение всего года. Особенно это важно для предприятий, которые ведут коммерческую деятельность.
- **Многоязычность**. Интернет – это многоязычная информационная среда. Особенно актуальными становятся задачи мультиязыкового и кросс-языкового поиска. Решение этих задач предполагает реализацию алгоритмов поиска, независимых от языка представления анализируемых в процессе поиска документов и языка представления информационных запросов пользователя.

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.

Проблема точности традиционно решается на пути использования линейного поиска по ключевым словам с привлечением некоторых лингвистических методов [12]. Ряд систем декларирует возможности 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.

Вместе с тем существующее положение дел в области информационного поиска не позволяет пока говорить о безусловной эффективности и качественности современных поисковых систем. Существует целый ряд противоречий и проблем, вызванных технической, методологической и организационной сложностью рассматриваемых задач.

4. Задачи и цели исследований

Целью данной работы является рассмотрение аспектов поисковых информационных систем, унификации коллекций документов, организации виртуальных предприятий, а также создание программных средств для

повышения эффективности автоматизированного поиска и извлечения информации.

Коммерческий спрос во многом изменил условия использования систем текстового поиска и выдвинул к ним новые требования. В сжатом виде главные из этих требований можно сформулировать следующим образом:

- эффективная обработка очень больших коллекций документов;
- улучшенное отображение смыслового содержания документов и пользовательских поисковых запросов;
- реализация мультимедийной обработки, т.е. совместной обработки документов разных форматов и представлений - текстовых документов, изображений, аудио, видео и др.;
- реализация эффективных методов поиска в потоках документов (задачи фильтрации (извлечения) коммерческой информации);
- доступ к результатам поиска – виртуальным предприятиям.
- интерактивный поиск (реализация диалога с пользователем во время поиска, уточнение запросов и т.д.).
  Отдельно стоит отметить повышение требований к поисковым системам в отношении так называемого человеческого фактора, определяющего эффективность взаимодействия человека и поисковой системы.

Многочисленные исследования в области поиска коммерческой информации среди неструктурированных коллекций документов не дали хороших результатов. Серьезной проблемой при обработке информации в информационных системах является различное понимание одних и тех же терминов. Смысл терминов меняется как от одной предметной области к другой, так и от одного сообщества людей к другому сообществу.

5. Исследования унификации коллекций документов

Другим направлением улучшения качества поиска является унификация описаний документов (метаданных) под определенный стандарт. Дублинское ядро [2] – Dublin Core (DC) – является грамотным и успешным проектом, связанным с разработкой структуры метаописаний ресурсов. Инициативной группой (Dublin Core Metadata Initiative, DCMI) был принят ряд точных концептуальных решений, позволивший найти приемлемый компромисс между выразительностью и простотой, естественностью и полнотой метаописаний. Поскольку проект DublinCore построен весьма элегантно и, в то же время, эффективно, он достоин подробного рассмотрения в качестве эталлоной на сегодняшний день системы метаописаний ресурсов.

Первоначальная версия Дублинского ядра была предложена в 1995 году на состоявшемся в Дублине (США) симпозиуме, организованном Online Computer Library Center (OCLC) и National Center for Supercomuting Applications (NCSA) для описания информационных ресурсов библиотечных систем.

В модели поиска, основанной на Дублинском ядре, представлением k-го документа является множество пар $D_k = \{(N_{ik}, V_{ik})\}$, где:

N_{ik} – имя i-го элемента метаданных Дублинского ядра в описании содержания k-го документа;

V_{ik} – значение этого элемента метаданных.

Представлением запроса также является множество пар некоторых элементов Дублинского ядра и их значений $Q = \{(N_j, V_j)\}$, где:

N_j – имя j-го элемента метаданных Дублинского ядра в описании пользовательского запроса;

V_j – значение этого элемента метаданных.

Критерий релевантности k – го документа выглядит следующим образом:

$$Q \subseteq D_k.$$

На основании стандарта Дублинского ядра была разработана централизированная поисковая система БУД (4bud.biz), главной идеей которого было собрать и структурировать коммерческую информацию, а также автоматизировать создание виртуальных предприятий.

6. Архитектура и программные средства системы

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.

В системе предусмотрено 5 типов пользователей:

- **Неавторизованный пользователь**. Посетитель системы может выполнять поиск и фильтрацию без ограничений и принимать участие в доступных ему сервисах.
- **Авторизованный пользователь**. Посетитель, который зарегистрировался в системе, автоматически получает персональный блокнот, может делать выборку и распечатывать реквизиты предприятий, отправлять почту через интерфейс, просматривать историю своего поиска, изменять регистрационные данные и принимать участие в других доступных ему сервисах.
- **Абонент системы.** Пользователь, который зарегистрировался в системе, как представитель организации, создает и управляет виртуальной организацией.
- **Редактор.** Лицо, обладающее полномочиями для управления содержанием базы данных. Права редактора ограничиваются определёнными разделами системы;

- **Администратор.** Под интерфейсом администратора понимается набор программного обеспечения, позволяющий управлять системой и поддерживать ее основные функции.

Компонентная модель представлена на рисунке 1.

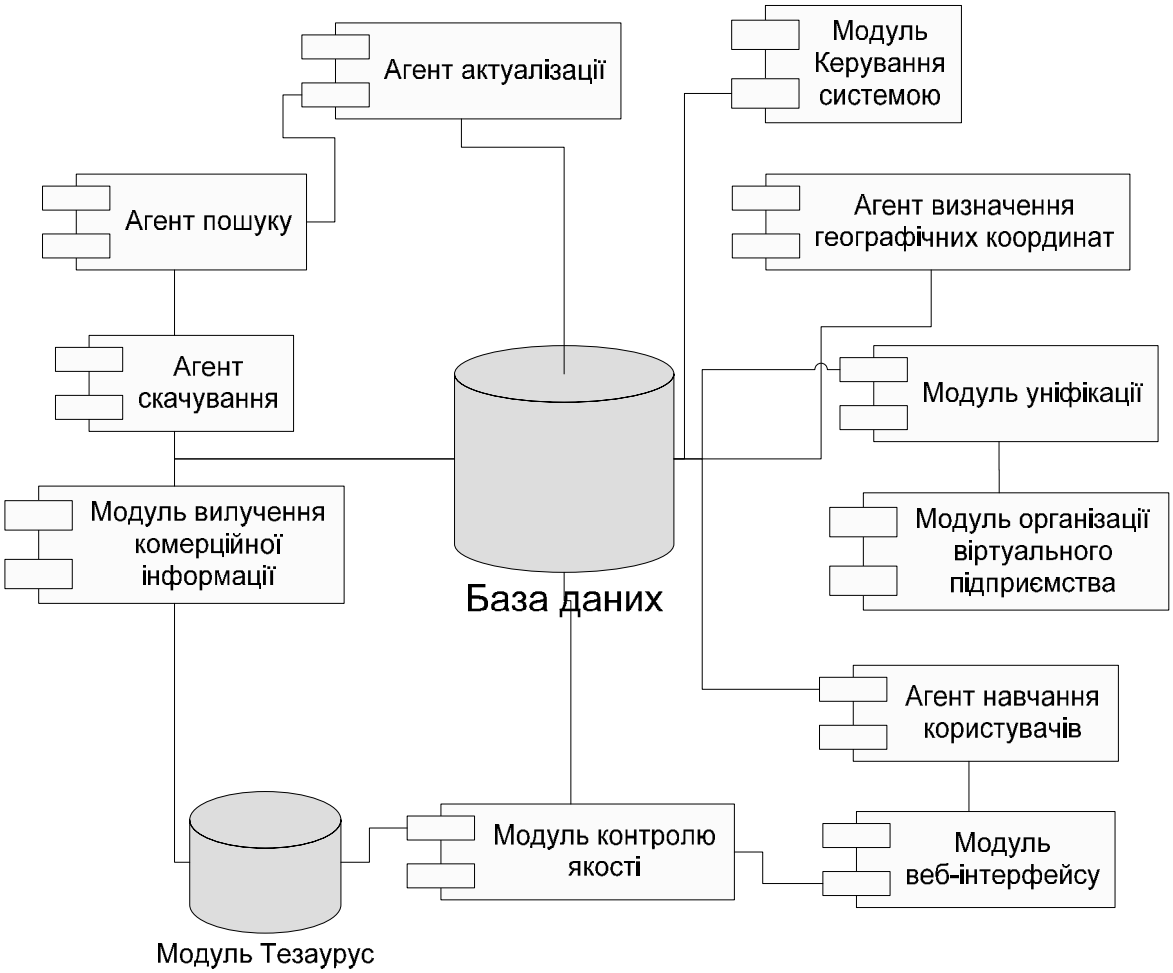

Рис.1.

Прежде чем какая-либо задача будет исполнена, она попадает в очередь задач. Задачи исполняются параллельно несколькими агентами. 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.

Рассмотрим подробнее основные компоненты системы.

6.1. Агент поиска

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. Указанная структура заполняется в полуавтоматическом режиме при помощи

модуля подключения новых информационных источников. 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.

6.2. 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 UTF-8. Скачанные документы сохраняются в централизованной базе данных системы для последующей обработки.

6.3. Модуль извлечения коммерческой информации

Модуль извлекает из записанного в базу данных текста, распознает основные поля и ключевые слова. На основании этих данных модуль проводит анализ и определяет какие виды услуг оказывает предприятие, разделы классификатора базы, к которым можно отнести предприятие и географическое расположение. Для точного определения структура документа не должна быть нарушена.

6.4. Агент определения географических координат

После точного определения адреса предприятия агент системы связывается с агентом Яндекса и получает координаты расположения предприятия на географической карте.

6.5. Модуль Тезаурус

Модуль Тезаурус – специализированный словарь предметной области, в котором указаны семантические отношения (ассоциативность, эквивалентность и иерархическая расширяемость) между лексическими единицами. Таким образом, на основании запроса пользователя строится поисковый образ.

6.6. Агент обучения пользователей

Далеко не всегда уровня грамотности пользователя достаточно, чтобы продуктивно работать с системой. Агент помогает правильно сформулировать запрос и предложить на основании Тезауруса другие наиболее релевантные. Интерфейс построен на интерактивном диалоге между пользователем и системой.

6.7. Модуль организации виртуального предприятия

Для реализации задач организации виртуального предприятия предусмотрено два режима: автоматический и ручной. В ручном режиме пользователь (представитель предприятия) регистрируется в системе, заполняет анкету, загружает прайс-листы, фотографии и другие медиа документы. Модуль обрабатывает данные и создает виртуальное предприятие. В автоматическом режиме модуль связывается с модулем извлечения коммерческой информации, получает информацию и также создает виртуальное предприятие.

6.8. Агент актуализации информации

Для решения задачи содержания базы данных о коммерческих предприятий в актуальном состоянии агент связывается с различными источниками для подтверждения информации или, в случае ручной регистрации, в режиме диалога обновляет базу данных.

6.9. Модуль контроля качества

Модуль контролирует ввод текста пользователем, проверяет входящие данные на орфографические и грамматические ошибки, выполняет типографическую обработку текста и записывает в базу данных.

6.10. Модуль унификации

Модуль унификации предоставляет данные о виртуальных предприятиях в виде RSS и по стандартам Дублинского ядра. В системе документ коллекции представлен (предприятие) следующим образом:

- Title – название предприятия;
- Creator – представитель предприятия (необязательное поле);
- Subject – классификация по разделам;
- Description – описание видов деятельности;
- Date – дата создания документа;
- Format – формат документа;
- Identifier – идентификатор документа;
- Language – язык документа.

6.11. Модуль веб-интерфейса

Модуль позволяет пользователю управлять системой через веб-браузер, ставить задачи поиска информации в Интернет, 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 нижней части страницы.

6.12. Модуль управления системой

В модуле реализован набор необходимых административных функций и компонентов для управления системой, пользователями и редакторами.

6.13. Другие стандартные библиотеки

Система построена на MVC-платформе Zend Framework 1.10.6 использует javascript-библиотеки jQuery 1.4.3 и базу данных MySQL 5.5.

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 поиска, извлечения информации и повышение скорости работы, а именно – временных параметров и качества фильтрации коммерческой информации.

Так как Интернет являет собою свободную платформу для разработок программного обеспечения, то «навязать» разработчикам определенный формат структурирования информации пока не представляется возможным. Формат унификации Дублинского ядра, предложенный в статье, наиболее качественно и в полном объеме позволяет описать любую коллекцию документов, в том числе и коммерческую информацию об организациях. Возможность свободного экспорта коллекции в данном формате позволяет работать с системой практически любым внешним агентам, что сопутствует распространению актуальной и структурированной информации в целом.

Указанные работы позволят расширить область применимости системы и предложить простой и удобный инструмент для решения организации виртуальных предприятий. Это направление является весьма перспективным и будет приоритетным в дальнейшей разработки системы.

Список использованной литературы

1. OpenCyc, http://www.opencyc.org.
2. The Dublin Core Metadata Initiative (DCMI), http://dublincore.org
3. Scientific Discovery, http://www.aaai.org/AITopics/html/discovery.htmI.
4. Semantic Web Community Portal, http://www.semanticweb.org/.
5. Web-сайт автономного программирования IBM, http://www-3.ibm.com/autonomic/index.html.
6. Web-сайт группы программирования эмоций в MIT (Affective Computing Group at MIT), http://affect.mediamit.edu/.
7. Web-сайт компании Cycorp, http://www.cyc.com.
8. Вагман М. Процесс научного открытия для людей и компьютеров: теория и разработка в психологии и искусственном разуме (Wagman M. Scientific Discovery Processes in Humans and Computers: Theory and Research in Psychology and Artificial Intelligence. Praeger Publishers, 2000).
9. Козлов Е. Б., Метелкин А. В., Хорошевский В. Ф. Мультиагентная ситема поиска информации в Интернет.
10. Куршев Е. П., Осипов Г. С., Рябков О. В., Самбу Е. И., Соловьева Н. В., Трофимов И. В. Интеллектуальная метапоисковая система.
11. Кормалев Д. А., Куршев Е. П., Осипов Г. С., Сулейманова Е. А., Трофимов И. В.: Препринт. Методы поиска и анализа информации. Автоматическое извлечение данных.
12. Осипов Г. С. Приобретение знаний интеллектуальными системами: Основы теории и технологии.
13. Куршев Е. П. Метод извлечения полуструктурированных данных из Интернет.