НОВАЯ КОНЦЕПЦИЯ ПОИСКА ЗНАНИЙ И ИНФОРМАЦИИ В ИНТЕРНЕТЕ.

Бурное развитие информационного сообщества, объединенного Интернет'ом, вызвало к жизни парадокс: по мере увеличения объема информации все труднее найти нужную. Это связано с тем, что :

1. Огромное большинство обычных пользователей Интернета затрудняется составить правильный запрос на поиск информации.
2. Поисковые машины ищут информацию по набору ключевых слов, а не по смыслу содержания web-страниц. Зачастую одно другому не соответствует.
3. Кроме того, небольшой набор ключевых слов вызывает выдачу поисковой машиной ссылок на огромное число web-страниц, просмотреть все их пользователь не в состоянии.
4. Ни одна поисковая машина (AltaVista, Yahoo, Google и т.д.) не покрывает и не может покрыть все информационное пространство Internet'а. По материалам различных исследований посещаемыми являются лишь 5-10% web-страниц.
5. Доступ к информационным ресурсам должен быть мобильным, т.е. пользователь должен получить возможность качественного доступа в Интернет независимо от того, что он использует в качестве терминала -офисный компьютер, PDA, автомобильный компьютер, мобильный или обычный телефон.

Чтобы преодолеть этот парадокс, необходимо решить следующие задачи:

1. Разработать систему распознавания/синтеза естественной речи.
2. Разработать контекстно-зависимую диалоговую систему составления запроса к поисковой машине. Такая система путем ведения диалога с пользователем на естественном языке должна определять область и предмет интересов пользователя и составлять запрос к поисковой машине, содержащий смысловое описание этих области и предмета. При ведении диалога пользователь может применять в качестве терминала офисный компьютер, PDA, мобильный или обычный телефон. При этом есть возможность сделать запрос с одного терминала (напр. С мобильного телефона), а необходимую информацию получить на другой (напр. домашний компьютер).
3. Разработать формализованный язык упрощенного индексирования смыслового описания информации.
4. Разработать поисковую машину, индексирующую смысловое содержание web-страниц, а также поддерживающую соответствующую базу данных.
5. Разработать базу данных, рассчитанную на работу с индексами смыслового описание web-страниц
6. Разработать иерархическую структуру сети поисковых машин Интернета. Сеть строится так, что запрос пользователя попадает к поисковым машинам верхнего уровня, которые «знают» только, к каким поисковым машинам нижних уровней надо обратиться, чтобы получить необходимую информацию. Чем ниже уровень, к которому принадлежит поисковая машина, тем уже область знаний, информация о которых находится в базе данных этой машины, но тем глубже знания о предметах этой области. Т.е. поисковые машины должны стать специализированными.
7. На период, когда предлагаемая иерархическая структура сети поисковых машин Интернета еще не развернута полностью, поисковые машины этой сети должны уметь преобразовывать запросы на языке смыслового описания в запросы, соответствующие форматам запросов современных поисковых машин (AltaVista, Yahoo, Google и т.д.).

Новые возможности Интернета.

1. «Каждому-свое!». Пользователь, независимо от того, хорошо или плохо он владеет компьютером, находится ли он в оффисе или на отдыхе, получает нужную информацию требуемого качества.
2. Появляется возможность адресной подсказки пользователю о тех или иных товарах или услугах, которые могут иметь отношение к его запросу (в т.ч. и адресная реклама).
3. Упрощается возможность контроля за Интернетом с целью противодействия противоправному поведению.

Архитектура системы метапоисковой машины для работы в Интернете и корпоративных сетях.

Основными компонентами системы являются: WWW-интерфейс, база данных, программа подключения поисковых ресурсов, модули лингвистического анализа текста, агентская среда и сами программы-агенты.

Поддержка WWW-интерфейса осуществляется с помощью Microsoft Internet Information Services 5.0 и Microsoft Active Server Pages. Взаимодействие пользователя с системой полностью осуществляется посредством этого интерфейса. При этом необходимо использовать Microsoft Internet Explorer версии 5.5 (в данный момент ведутся работы, чтобы смягчить это требование до MS IE 4.0 и выше).

База данных предназначена для хранения запросов пользователя и результатов их обработки, описания поисковых ресурсов Сети, словарей предикатных слов, синонимов, тезауруса, а также всей служебной информации, которая необходима для работы программ-агентов. Система использует СУБД Microsoft SQL Server 2000.

Модули лингвистического анализа текста – агент «Лингвист» и модуль семантического анализа – выполняют всю необходимую лингвистическую обработку текста – морфологический, синтаксический и семантический анализ.

Агентская среда предназначена для организации распределенной обработки запросов пользователя, при которой отдельные, функционально независимые этапы работы, реализованные модулями-агентами, могут выполняться на разных компьютерах локальной сети. Управление задачами в распределенной среде выполняет менеджер распределенных вычислений.

Типичный сценарий работы пользователя с системой следующий:

1) Используя WWW-интерфейс, пользователь задает новый поисковый запрос – в Интернет, либо в локальной базе данных. Здесь же он указывает расписание выполнения запроса.
2) В ответ на запрос система формирует список задач для его выполнения и график их запуска.
3) Менеджер распределенных вычислений опрашивает базу на предмет текущих задач, определяет программу-агент, отвечающий за выполнение этой задачи и назначает выполнение задачи в соответствии с загруженностью компьютеров локальной сети.
4) В случае поиска в Интернет, система сформирует следующий список задач:
   a. Перенаправление запроса подключенным поисковым ресурсам и выгрузка результирующего списка URL (агент метапоиска).
   b. Загрузка в базу текста документов из списка отклика каждой поисковой машины (агент загрузки документов из Интернет).
   c. Анализ текста запроса и документов и построение семантического представления этих текстов (агент «Лингвист» и модуль семантического анализа).
   d. Сопоставление полученных семантических образов с целью выяснения семантической релевантности документа (агент семантической фильтрации).
5) После выполнения запроса формируется список отклика, в котором документы ранжированы по семантической релевантности и релевантности по ключевым словам. Этот список пользователь также просматривает через WWW-интерфейс.

АГЕНТ ЗАГРУЗКИ ДОКУМЕНТОВ ИЗ ИНТЕРНЕТ

В задачи агента входят:

1) Загрузка документа по ссылке (URL), используя протокол HTTP. В настоящий момент поддерживаются документы в формате plain text и HTML.
2) Выделение текста документа и преобразование его в кодировку Windows 1251.

АГЕНТ «ЛИНГВИСТ»

Для реализации информационного поиска необходимо анализировать лингвистические характеристики текста документа. Это необходимо для описания поискового образа и при его сопоставлении информации в тексте. Задача агента «Лингвист» предоставить другим агентам всю необходимую лингвистическую информацию о тексте. В настоящее время агент способен обрабатывать только русскоязычные тексты.

Агент выполняет следующие виды лингвистического анализа:

1. Морфологический анализ. Анализ основан на словаре основ и флексий. Производится предсказание парадигмы в случае, когда слово не обнаружено в словаре. Результатом является множество омонимов графемы с полными морфологическими характеристиками.
2. Синтаксический анализ. Этот вид анализа разбит на две концептуальные фазы:
   2.1. Микросинтаксический анализ. Включает в себя выделение именных групп различной структуры.
   2.2. Макросинтаксический анализ. Осуществляет фрагментацию предложений и выделение клауз. Завершает построение полного синтаксического дерева предложения.
3. Разрешение местоименной анафоры. Под этим понимается сопоставление местоимению того объекта, который он обозначает. В настоящее время обрабатывается девять следующих местоимений: он, она, оно, они, его, ее, их, который, чей.

Результатом работы агента является специальное описание лингвистической информации текста.

МОДУЛЬ СЕМАНТИЧЕСКОГО АНАЛИЗА

Основной задачей модуля является построение семантического образа текста. Модуль использует результаты обработки текста агентом «Лингвист», который осуществляет морфологический и синтаксический анализ. Среди прочей информации, модуль «Лингвист» выявляет, в частности, именные группы и предикатные слова (глаголы и отглагольные формы). Семантические связи между этими именными группами устанавливаются в результате частичного семантического анализа, основываясь на моделях управления найденных предикатных слов. Виды семантических отношений, а также грамматические признаки, позволяющие обнаружить их в тексте документа, определяются лингвистами в справочнике предикатных слов.

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 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 отношений. Каждый элемент списка включает в себя тип отношения, и связанные именные группы. Для быстрого поиска список проиндексирован по типам отношений, частям речи и словарным формам корневых слов из каждой именной группы.

АГЕНТ ИНДЕКСИРОВАНИЯ

Агент индексирования служит для сохранения временного индекса ключевых слов в постоянном индексе базы данных. В процессе обработки документов Интернет, агент

семантической фильтрации строит для каждого документа временный индекс по ключевым словам. Просматривая результаты поиска, пользователь может отметить те документы, которые он хотел бы сохранить в базе данных. Поскольку таких документов со временем может накопиться довольно много, пользователю понадобится поиск в локальной базе. В связи с этим временный индекс каждого документа интегрируется в общий индекс ключевых слов. Семантический индекс в силу своего большого объема, не хранится в базе данных постоянно.

АГЕНТ СЕМАНТИЧЕСКОЙ ФИЛЬТРАЦИИ

Основными задачами агента семантической фильтрации являются

1) оценка семантической близости запроса и документа (в процентах);
2) упорядочивание результирующего набора документов в соответствии с этой оценкой (документы с большим процентом семантической релевантности показываются в первую очередь).

Агент использует модуль семантического анализа для построения поискового образа запроса. Затем выясняется тип запроса. На данном этапе поддерживаются два типа запросов: поиск в локальной базе данных и в Интернет. В первом случае агент обрабатывает все документы, сохраненные пользователями в локальной базе, во втором – все документы, ссылки на которые вернули поисковые машины Интернет, и текст которых был успешно скачан в базу. Для документов Интернет строится временный индекс ключевых слов.

Далее происходит поиск ключевых слов из запроса в индексе ключевых слов документов. По результатам этого поиска из текста документов вырезаются фрагменты, содержащие ключевые слова запроса. Эти фрагменты передаются для обработки модулю семантического анализа, который строит их поисковый образ. Наконец, агент сравнивает поисковый образ запроса с поисковыми образами фрагментов документа и вычисляет процент совпадающих семантических связей. Если в документе найдены все связи из запроса, такой документ считается 100% релевантным. Попутно агент выделяет фрагменты с наибольшим семантическим весом, включая их в краткую сводку по документу.

Интерфейс пользователя

Постановку задач и просмотр результатов работы системы пользователь осуществляет с помощью Web-интерфейса.

В начале работы с системой пользователь задаёт пароль доступа. По этому паролю определяются права пользователя и его персональные настройки. После входа в систему, пользователь может создавать поисковые задачи и просматривать результаты поиска предыдущих задач.

Для удобства организации информации пользователь может создать несколько тематических папок. В них он может создавать другие папки и задачи. При формировании поисковой задачи у пользователя запрашивается вся необходимая для ее решения информация. Во-первых, задача может выполняться однократно или периодически (*режим мониторинга*). Во-вторых, можно указать дату и время запуска задачи. В-третьих, можно указать приоритет задачи. Чем выше приоритет, тем быстрее будет получена информация по этому запросу. Обязательным параметром поисковой задачи является поисковый запрос.

После постановки задачи и наступления времени ее запуска система начинает выполнять поиск. Результаты работы поиска представляются в виде таблицы. В ней приводится ссылка на сайт, где была найдена информация, краткая выдержка из содержимого найденного документа, а также оценка релевантности найденного документа поисковому запросу. Для детального ознакомления с содержимым документа, пользователь может по гиперссылке перейти на сайт-источник.

Web-интерфейс реализован по технологии Active Server Pages и должен базироваться на Microsoft Internet Information Server.

АГЕНТ ЛИНГВИСТИЧЕСКОЙ БАЗЫ ДАННЫХ

Лингвистическая база данных содержит сведения о предикатах, а также о ролях и связях свободных синтаксем для каждого предиката. База данных заполняется лингвистом, для чего разработан специальный интерфейс лингвиста. Для доступа к базе данных используется динамическая библиотека, которую могут использовать другие приложения.

Библиотека выполняет несколько функций и имеет несколько вариантов входов. При определении ролей входом является предикат и пары "предлог+падеж". Выход – возможные варианты роли. При установлении связей между синтаксемами вход - глагол и пара ролей, полученных на предыдущем шаге, а выход - варианты связей.

АГЕНТ ПОДКЛЮЧЕНИЯ ПОИСКОВЫХ РЕСУРСОВ

Основной задачей агента является обеспечение полуавтоматического подключения к системе новых поисковых ресурсов. Агент составляет формализованное описание поискового ресурса и помещает его в XML-структуру. В последствие данная структура используется агентом мета-поиска для получения ссылок на найденные по поисковому запросу документы.

Описание поискового ресурса содержит следующую информацию:

1. Формат запроса к поисковому ресурсу (HTTP-запрос, в котором задается текст запроса).
2. Структура отклика поискового ресурса:
    1. ссылки на найденные документы;
    2. ссылки на следующие страницы с документами.
3. Возможности расширенного поиска:
    1. графическое представление операторов AND, OR, NOT, NEAR;
    2. возможности поиска по словосочетаниям.

Подключение поисковых ресурсов представляет собой постраничного помощника, в котором XML-структура уточняется шаг за шагом. Если агент не может автоматически распознать формат поискового ресурса, то используется обучение с учителем.

АГЕНТ МЕТАПОИСКА

Агент метапоиска решает две основные задачи, первая - выполнение поискового запроса и получение ссылок на найденные документы, и вторая - разбор структуры описания поискового ресурса и передача этой информации другим агентам.

1. Получение ссылок на документы.

Это основной метод агента, позволяющий получить URL найденных по поисковому запросу документов.

Входные параметры:

1. поисковый запрос (оформляется агентом, отвечающим за формирование запроса);
2. структура описания интерфейса поискового ресурса (XML-структура, составленная агентом подключения поисковых ресурсов);
3. количество ссылок из результирующего списка, которые нужно вернуть.

Выходные параметры:

1. агент возвращает ссылки на найденные документы.
2. статус выполнения операции (0 – успех, код ошибки – в случае неудачи)

2. Извлечение информации из структуры описания поискового ресурса.

Эта функциональность позволяет инкапсулировать в агенте мета-поиска всю логику по разбору структуры описания поискового ресурса.

Входные параметры:

1. структура описания интерфейса поискового ресурса.

Выходные параметры:

1. графическое представление операторов AND, OR, NOT, NEAR соответственно;
2. есть ли поиск по словосочетаниям и его параметры;
3. статус выполнения операции (0 – успех, код ошибки – в случае неудачи).

Агент представляет собой Win32 DLL

МЕНЕДЖЕР РАСПРЕДЕЛЕННЫХ ВЫЧИСЛЕНИЙ

Решаемые системой задачи требуют для своей работы больших вычислительных мощностей. Чтобы обеспечить высокий уровень производительности системы имеется возможность использовать для работы несколько компьютеров. Эти компьютеры должны быть объединены сетью, чтобы обеспечить возможность взаимодействия частей системы, работающих на них. Для такой организации работы разработана специальная агентная архитектура (*агентная среда*). А чтобы координировать работу частей системы, создан менеджер распределённых вычислений.

Общая задача обработки документа может быть разбита на ряд подзадач. Это делает возможным выделение этих подзадач в отдельные приложения (*агенты*) и относительно изолированное их выполнение. Такой подход позволяет организовать распределённую обработку информации, а в некоторых случаях и параллельную. Последнее, например, имеет место при обработке сразу нескольких документов.

Менеджер распределенных вычислений осуществляет управление агентами. Он является ключевым элементом агентной среды. Его функцией является обнаружение задач (*работ*), планирование их исполнения и запуск агентов, предназначенных для решения этих задач. Поставщиками работ являются пользователи и другие агенты. После обнаружения работ менеджер принимает решения о том, какие агенты, в каком порядке и на каких компьютерах запускать. Планирование может осуществляться автоматически или в соответствии с правилами, заданными при настройке системы. В текущей версии планирование осуществляется на основе приоритетов задач, приоритетов агентов, распределения исполняющихся агентов и черных списков (эти списки запрещают запуск конкретных агентов на конкретных компьютерах). Существует механизм для оценки мощности аппаратных средств компьютеров. Эту информацию тоже можно использовать при планировании, но в настоящее время механизм не используется. После выполнения планирования менеджер запускает на компьютерах агентов и передает им параметры задачи. По выполнении работы агенты сообщают об это менеджеру. По умолчанию агенты запускаются динамически. Но возможно и стационарное распределение агентов по компьютерам. Это может обеспечить небольшое увеличение производительности.

Кроме работы в штатном режиме, менеджер способен функционировать в нестабильной среде. Он содержит мощные средства восстановления рабочего режима в случаях сбоев сети и даже сбоев компьютеров. При этом гарантируется, что ни одна работа не будет потеряна.

Менеджер имеет пользовательский интерфейс и позволяет визуально контролировать ход выполнения работ и распределение вычислительных нагрузок по компьютерам. Это удобное средство для анализа работы системы и точной настройки агентной среды.