СОСТАВЛЯЮЩИЕ ИНТЕЛЛЕКТУАЛЬНОГО ПОИСКА В СЕТИ[1]

Жигалов В.А.
РосНИИ Искусственного интеллекта, Москва
zhigalov@aha.ru
Кешелава В.Б.
Группа компаний СТЭК, Москва
kesha@stack.net

Аннотация

Поисковый сервис в Сети является одним из базовых. Тренд его эффективности в решении актуальных задач отражает степень адекватности современных программных технологий по отношению к обработке информации как таковой. В статье анализируется текущее состояние поисковых технологий в Интернет, тенденции и потенциал развития, а также рассматривается вопрос о применимости прикладных лингвистических систем и ресурсов к повышению качества поиска в Интернет. В качестве одного из способов интеллектуализации поиска предлагается моделирование ресурсов Интернет (сайтов и их разделов, страниц и их составляющих), где под моделированием понимается учет предметной области ресурсов, а также их структурных и жанровых особенностей. Базовая идея качественно нового поиска – использование знаний (как знания предметной области, так и закономерностей строения Сети и ресурсов в ней) и эффективная передача этих знаний от эксперта системе. Таким образом, интеллектуальный поиск должен использовать те же методы, которые сейчас использует обычный пользователь, перебирая вручную массу ссылок и по ряду признаков за доли секунды безошибочно определяя реальную релевантность ресурсов в списке найденных традиционными поисковиками - только это должно осуществляться автоматически. Рассмотрены некоторые интеллектуальные технологии в применении к поиску в Сети.

Интеллектуальный поиск: спрос и предложение

Обсуждение адекватности существующих традидиционных технологий поиска давно стало общим местом. На то, какие здесь возможны перспективы, влияет как спрос на новые технологии поиска, так и предложение. Пожалуй, не ошибемся, если скажем, что спрос на эти технологии до сих пор был очень фрагментарным, несмотря на то, что Сеть уже сейчас представляет собой океан информации, по которому, используя традиционный поиск, можно двигаться только по поверхности [3].

Недостаточный спрос обусловлен значительными усилиями, которые необходимо приложить для построения интеллектуальных технологий поиска (в сравнении с относительной простотой поиска традиционного), т.е. их дороговизной. Скачок сложности технологий объясняется просто: в них необходимо включать механизмы, использующие знания - о ресурсах Сети, о различных предметных областях. Помноженная же на объем накопленной информации в Сети, которую надо обработать новыми технологиями, эта сложность многим представляется непреодолимым барьером.

Между тем, если оценить объем усилий, потраченных информационным сообществом за последние 10 лет на создание текущего наполнения WWW, то задача создания качественного поиска уже не выглядит столь утопичной в смысле необходимых для ее решения ресурсов. Принцип распределения усилий на основе выработанных сообществом стандартов применим в наши дни не меньше, чем 10 лет назад, на заре WWW.

[1] Работа выполнена при поддержке Российского фонда фундаментальных исследований (проект № 02-07-90368).

Возможной точкой кристаллизации может стать проект предметно-ориентированных служб поиска, которые могли бы, используя формализованные знания, постепенно обрабатывать отдельные сегменты Сети, удовлетворяя насущные потребности обычных пользователей (на поверхности лежат идеи качественного поискового сервиса музыкальных MP3-файлов и поиск информации о товарах).

Эти задачи явно стыкуются с инициативой Semantic Web консорциума W3C: она предполагает создание сети ресурсов с их смысловым описанием через онтологии (на языках RDF или OWL). Поисковый сервис в такой Сети Знаний должен опираться на эти описания, но что делать, если их нет? Современное состояние Сети характеризуется довольно слабыми подвижками в направлении описания ресурсов, и необходимы сервисы, которые могли бы или генерировать такие описания (так, как их описал бы эксперт), или заменяли бы собой эти описания, предоставляя другим сервисам необходимую информацию в той или иной нотации, срезе и полноте.

Возможно ли создание таких сервисов на текущем уровне развития Сети и информационных технологий? Мы утверждаем, что да. Для этого на вооружение должна быть взята та же идея, которая в свое время породила направление "искусственный интеллект": использовать для решения задач подходы, характерные для человека, но средствами компьютеров. Только ИИ, обязанный своим появлением и своими масштабами военным бюджетам сверхдержав в условиях холодной войны, вместо того, чтобы решать конкретные задачи минимумом средств, стал решать фундаментальные проблемы, растекаясь по всему отпущенному бюджету. История ИИ показала, что заниматься решением проблем столь же более удобно, сколь и более накладно и безрезультативно, чем решать конкретные задачи. Теперь же появилась реальная и очень масштабная, но все-таки конкретная задача, и решать ее следует, более здраво отнесясь как к целям, так и средствам.

Рассмотрим это на трех примерах. В ходе рассмотрения каждого из примеров мы коснемся различных аспектов, которые, как правило, применимы ко всем примерам и к задаче в целом.

Пример №1. Поиск музыки

Рассмотрим задачу поиска информационных объектов в Сети на примере поиска музыкальных файлов. Ситуацию в современном Интернете в этой области можно описать так: имеется громадное количество доступной для скачивания музыки, но традиционные механизмы поиска с трудом помогают найти нужную композицию, при том, что пользователь, как правило, может сформулировать очень точные критерии поиска (название композиции, исполнитель, альбом и т.д.). Распределенные p2p-системы обмена музыкой, как правило, недолговечны и рассыпаются под юридическим натиском звукозаписывающих компаний[2].

Итак, рассмотрим сервис, который:

- Позволяет находить звуковые файлы при указании их характеристик из музыкальной предметной области (например, название композиции)
- Автоматически сканирует Сеть в поисках этих файлов
- Не требует явной регистрации ресурсов (сайтов)
- Публично доступен

Иными словами, рассмотрим поисковую систему, по принципам работы похожую на традиционную (свободный поиск с открытым предоставлением поисковых результатов), но технологически построенный на других принципах. Какие это принципы?

Во-первых, стоит учитывать, что пользователь при поиске использует некоторые знания из предметной области. В нашем случае это современная музыкальная онтология (например, музыкант или группа выпускает альбом, который имеет год выхода в свет, название, список композиций). Система должна знать не только эти базовые понятия, но и обладать знаниями о каждом музыканте, альбоме и композиции в их взаимосвязях, т.е. должна опираться на базу знаний (БЗ). В качестве отправной точки для создания такой БЗ может служить база данных CDDB, содержащая подробные каталоги большинства вышедших в свет музыкальных CD, а также данные об альбомах, композициях, жанрах, артистах и т.д.

Во-вторых, пользователь хорошо знает типичные навигационные "архетипы" сайтов, содержащих музыкальные ссылки: есть страницы поиска, есть страницы, содержащие списки музыкантов (например, сгруппированных по первой букве), списки альбомов конкретных музыкантов, списки композиций конкретных альбомов. Знания

[2] Мы не рассматриваем здесь вопрос о законности или этичности такого поискового сервиса, как и большинство аналогичных вопросов этичности или законности использования технологий вообще: к самим технологиям, на наш взгляд, вопросы этики или закона просто неприменимы, даже если абстрагироваться от вопросов распространения информации в сетевом сообществе.

касаются прежде всего того, что пользователь безошибочно узнает, на какой тип страницы он попал и куда ему двигаться дальше, благо ссылочная модель очень часто практически повторяет строение онтологии. Эти знания порождены опытом "брожения" по сотням и тысячам страниц и составляют навык Web-сёрфера. Здесь учитывается и дизайн страниц, и их структура, и ссылочная модель сайта.

Если наша поисковая система будет учитывать эти знания, то ей надо уметь:

- Категоризировать страницы
- Моделировать структуру сайта (ссылочную и иерархическую).

Категоризация и моделирование, очевидно, невозможно без учета онтологии (надо уметь отделять названия композиций и альбомов от других элементов страниц)[3].

Категоризация страниц отличается от классификации текстов тем, что значительную часть информации несут нетекстовые составляющие: расположение элементов на странице, их стиль - т.е. дизайн. В этом смысле задача категоризации страниц основана на их моделировании, причем на моделировании с позиции пользователя: как бы он воспринял страницу и к какой бы категории ее отнес.

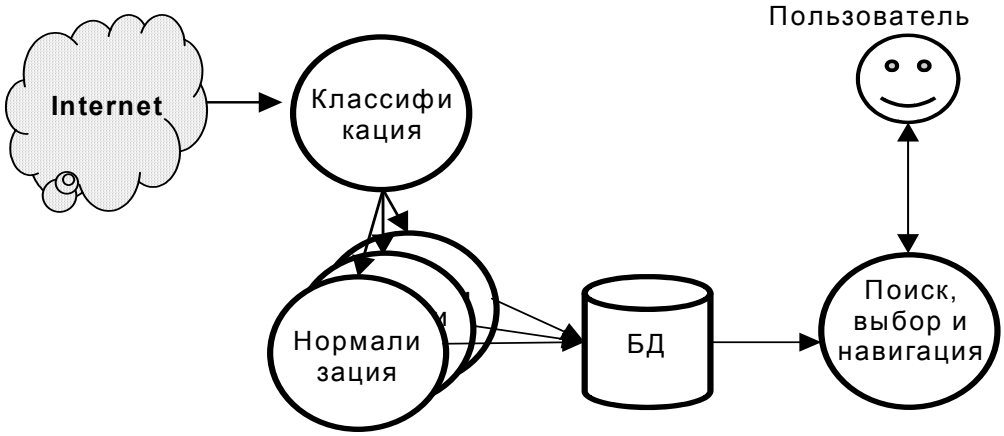
Задача моделирования страниц может решаться описанием структурных шаблонов. В базовые элементы (хотя каждый из этих элементов - в общем случае тоже шаблон) этих шаблонов входят: списки, таблицы, ссылки, баннеры, текстовые фрагменты, причем необязательно выраженные явно как соответсвующие HTML-структуры.

Пример №2. Поиск описаний товаров

Те, кто выбирал (не обязательно покупал) себе товар в Интернете, знают, что приходится просеивать много страниц, чтобы выделить нужную информацию. Критерии поиска можно сформулировать примерно так: «описание товара А», «статья о товаре Б», или «цены на В». Как и в случае с поиском музыки, пользователь хочет найти довольно ограниченную в жанре и предметной области информацию (в варианте «статья» - в основном текстовую, в вариантах «описание» и «цена» - структурно обогащенную).

Эти жанры поддаются моделированию, но усилий здесь потребуется гораздо больше. Дело в том, что предметная область товаров (особенно так называемых гаджетов – высокотехнологичных устройств) гораздо богаче предметной области каталога музыкальных произведений. Для гаджетов количество свойств товара (если подходить с позиций схемы классов), влияющих на выбор товара, несколько десятков, причем для каждого класса устройств параметры могут быть свои. Усугубляет сложность покрытия этой предметной области и тот факт, что описания довольно разнообразны в форме подачи. Человек без труда поймет эти описания – они составлялись человеком для человека, но требуется довольно много усилий, чтобы научить компьютерную систему так же гибко воспринимать разнообразие в подаче информации даже в узкой предметной области.

По нашим оценкам, при использовании технологий анализа, основанных на настройке сложных шаблонов, для нескольких первых источников пересечение шаблонов составляет не более 30% - и повторно используемая часть настройки для различных источников растет сначала довольно быстро, и в конце концов приближается к известному принципу 80/20, который в данном случае можно сформулировать так: 20% усилий тратится, чтобы покрыть 80% разнообразия.



[3] Кроме того, часто ссылки ведут на закрытые паролем или уже на несуществующие страницы. Система должна уметь такие ссылки отслеживать и помечать как недоступные - как и любая другая поисковая система.

Рис. 1. Схема организации сервиса поиска информационных объектов в Сети.

На основе технологий, которые мы упомянем ниже, вполне реальна система, включающая следующие процессы (Рис. 1): процесс классификации выделяет представляющие интерес страницы, которые затем проходят процесс нормализации, т.е. выделения нужной информации. Затем эта информация складывается в базу данных, где доступна для части, отвечающей за поиск и выбор информационных объектов.

Пример №3. Интеллектуализация поиска документов

Среди пользователей уже сложилось представление о «видимой» (информации, которую можно найти) и «подводной» (информации, которую практически найти невозможно) части Интернет. Эксперты расходятся в точной оценке соотношения этих частей, но сходятся в одном: «подводная» часть во много раз больше «надводной». Существующая в Интернет и будучи доступной теоретически информация становится практически недоступной из-за неэффективности работы традиционных поисковых механизмов и неадекватности архитектуры поискового сервиса.

Проблема поиска в Сети тесно связана с проблемой поиска знаний вообще. С этой точки зрения любой пользователь ищет именно знания, а не просто документы, данные или объекты. Представляется очевидным, что такой поиск информации должен быть ориентирован на соответствие «смысла» запроса и получаемых документов. Обеспечить такой поиск можно, если строить средства поиска на основе некоторого «понимания» машиной поиска как самого запроса, так и документа. Однако понимание документов должно выполняться лишь в контексте необходимости решения основной задачи - поиска. Иными словами, полное понимание документа не обязательно, если для нахождения этого документа можно обойтись лишь частичным пониманием.

Необходимое «понимание» документов возможно за счет использования систем, основанных на заранее сформированных базах знаний. Этот подход использует смысловые сети слов (онтологии), организованные в виде концептуальных отношений, связывающих узловые понятия, и чаще всего реализуется в виде семантической сети. Метод предопределяет ключевые слова и понятия, объединяющиеся в базу знаний, которая отражает содержание конкретного информационного массива. Затем база знаний может использоваться для поиска и ранжирования групп родственных документов. Подход на основе семантических сетей реально обладает достаточной гибкостью, доступен для расширения и не слишком громоздок при эксплуатации.

Отметим также, что использование онтологий может не только помочь решить задачу повышения эффективности поиска, но и сделать язык формулирования запроса на поиск простым и удобным для пользователя, максимально приблизив его к естественному языку.

Основные проблемы использования данного подхода обычно возникают из-за широты постановки задач: как правило, в такую задачу включают проблему понимания текстов и запросов на естественном языке, их обработку и вывод результатов в естественноязыковой форме. Согласно нашим оценкам, общелингвистическая и смысловая проблема в ближайшее время не будет решена (если такое решение вообще возможно), но задачу ограниченного понимания с целью интеллектуализации поиска можно успешно решить отдельно в пределах каждой узкой предметной области.

Описанные онтологии предлагается применять в качестве посредника между пользователем и поисковой системой на всех этапах обработки запроса:

1) обработка запроса с целью формирования поискового образа релевантного документа;

2) построение запроса к поисковой системе;

3) формирование списка релевантных документов.

При этом онтологии предметных областей могут выступать и в качестве единого «канонического» описания, помогающего решить проблему несовместимости и противоречивости понятий, используемых при составлении запросов для извлечения информации из сети Интернет.

Ядром такой поисковой системы является интеллектуальное хранилище документов, которое хранит не сами документы (или ссылки на документы), а и их метаописания. К функциям такого хранилища относятся создание и поддержание шаблонов моделей предметных областей (МПО), создание метаописаний документов на основе их соотнесения с предметными областями, хранение созданных метаописаний и обеспечение доступа к ним.

Реализация предлагаемого решения для интеллектуализации поискового сервиса представляется в следующем виде:

1. для каждой предметной области создается своя Модель (МПО), которую можно рассматривать как шаблон метаописания;

2. каждый документ, попадающий в поле зрения хранилища, соотносится со своей предметной областью (рубрикация, каталогизация);
3. на основе существующего набора МПО для каждой единицы хранения (документу, сообщению) в хранилище создается ее метаописание;
4. поиск должен производиться не по документу, а по его метаописанию в интеллектуальном хранилище;
5. сортировка полученных документов производится по соотнесению документа с различными МПО.

Рассмотрим основные этапы обработки поискового запроса системой, основанной на описанных выше принципах.

1. На этапе формирования образа искомого документа из пользовательского запроса выделяются смысловые структуры: идентифицируется предметная область и выделяются значимые слова и термины. При этом диалог запроса кроме поисковых слов и выражений может предусматривать введение дополнительной информации. Такая информация может уточнять жанр искомого документа (обзор, техническое описание, научная статья, художественное произведение и т.д.), идентифицировать предметную область (технологии, здравоохранение, производство, торговля и т.д.) и пр.

2. Полученные смысловые структуры затем используются для формирования поискового образа с применением эвристических правил вывода на онтологии. Образ релевантного документа представляет собой описание желаемого результата работы поисковой системы, которое включает в себя:

- расширенный набор терминов, которые должны включаться в документ;
- набор характеристик документа;
- набор требований к результату поисковой системы, таких как количество документов и т.п.

3. На этапе построения запроса к поисковой системе на основе поискового образа документа формируется соответствующее выражение на языке запросов конкретной поисковой системы. Обычно такой запрос представляет собой соединенный логическими связками набор терминов и понятий. На этом этапе возможен дополнительный диалог с пользователем для уточнения поискового предписания.

4. Расширенный (специализированный) и уточненный запрос автоматически модифицируется в запрос к поисковой системе. При этом задаются параметры поиска, специфичные для каждой поисковой системы.

5. Как известно, результатом работы поисковой системы является множество ссылок на целевые документы (в форматах HTML, TXT, PDF, DOC и т.д.). Так как среди полученного множества ссылок все еще могут содержаться ресурсы, недостаточно релевантные запросу пользователя, то необходимо выполнить проверку соответствия полученных документов поисковому образу документа. Этот анализ также выполняется с использованием онтологии. Заметим, что из-за возможно большого количества ссылок, анализ документов должен проводиться на основе жестких критериев отбора. После проведенного «отсечения» лишних документов результаты поиска отображаются в удобном для пользователя виде.

Реализация описанной схемы организации поиска с использованием онтологий предполагает наличие следующих этапов:

1. Построение онтологии;

2. Выбор средств анализа результатов поиска;

3. Создание интерфейса между онтологией и пользователем; этот интерфейс должен обеспечивать настройку онтологии на конкретного пользователя, построение и ввод запроса в комфортной для неспециалиста форме, а также просмотр результатов запроса;

4. Создание модуля взаимодействия онтологии с поисковой системой; данный модуль должен обеспечивать перевод запроса, преобразованного онтологией, в запрос в формате соответствующей поисковой системы и оценивание результатов его выполнения.

На Рис. 2 приведена схема организации интеллектуального хранилища.

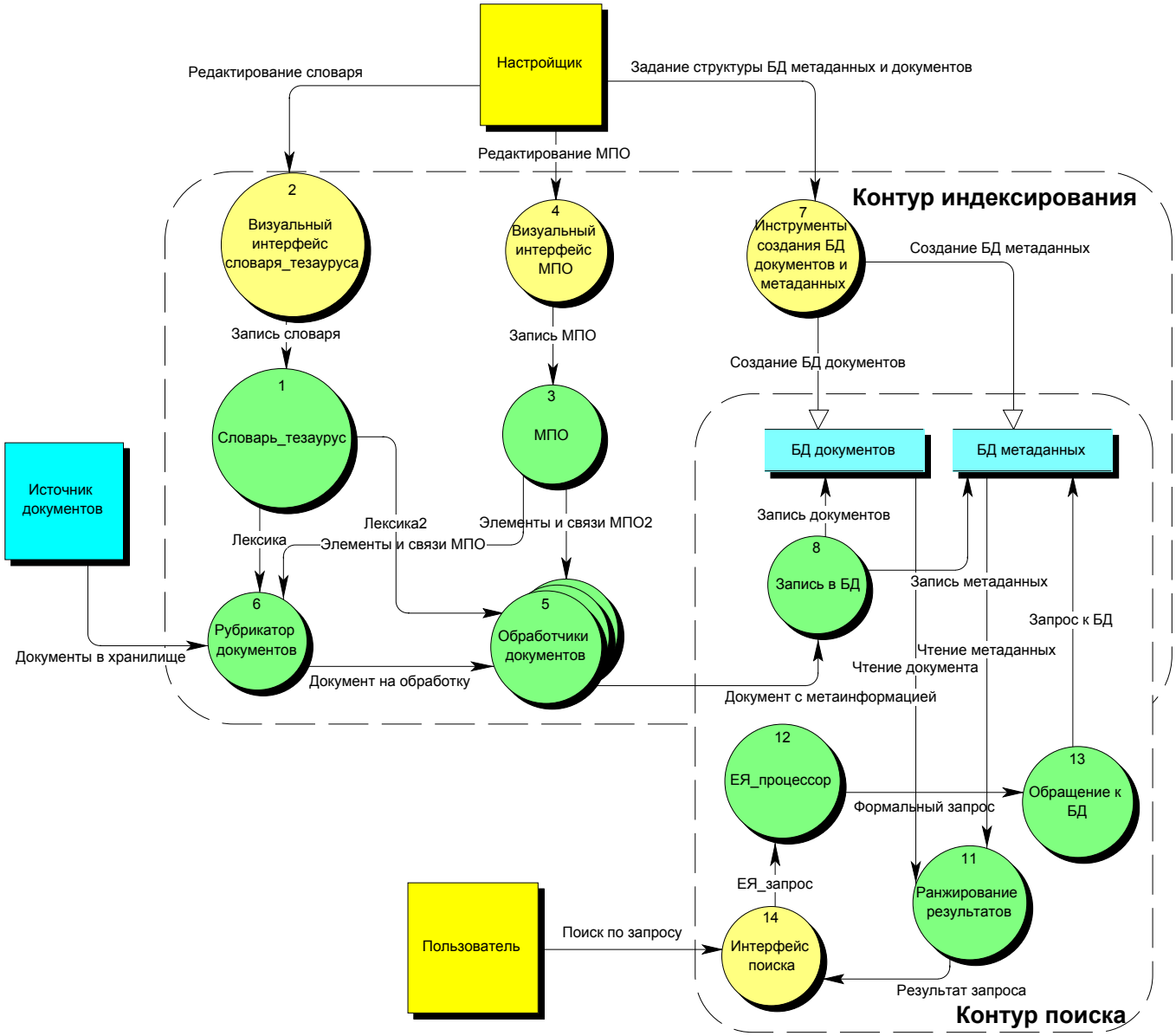


Рис. 2. Схема потоков данных интеллектуализированного хранилища.

Преимуществом данного подхода является то, что оценка релевантности результатов выполнения запроса его смыслу может проводиться на всех этапах движения (обработки) запроса от пользователя к поисковой системе и, наоборот, от поисковой системы к пользователю. Это делает возможным использовать данный подход для разработки эффективных систем поиска в сети Интернет, легко настраиваемых на выбранную предметную/проблемную область [1].

Такая поисковая система может рассматриваться как стартовая точка для разработки глобальной распределенной по множеству участников сети поисковой системы. Разработанная архитектура и использованные технологии предусматривают возможность клонирования машины поиска и согласованной работы сети независимых клонов, что позволяет решить проблему поиска информации в больших информационных пространствах. При таком подходе весь Интернет может быть разделен (по регионам, государствам, направлениям бизнеса или др.) между несколькими специализированными машинами, поддерживающими и ищущими информацию независимо, но отвечающими на запрос согласованно

Технологии

Какие же технологии должны быть использованы для решения этих задач?

Необходим полный спектр - от обработки ЕЯ-текстов и моделирования предметных областей и ресурсов, до организации распределенных вычислительных систем:

- Поиск в тексте на основе сложных шаблонов
- Классификация документов
- Нормализация слабоструктурированной информации
- ЕЯ-интерфейсы
- Построение моделей предметных областей и онтологий
- Построение моделей страниц и сайтов
- Ведение предметных словарей и тезаурусов
- Индексирование
- Распределенные хранилища данных
- Балансировка нагрузки

и т.д.

Совершенно не претендуя на полноту и объективность рассмотрения, просто перечислим некоторые примеры реализаций таких технологий (в основном те, которые близки и хорошо известны авторам данной статьи):

- Alex - технология лексического анализа текстов на основе шаблонов произвольной сложности ([2], http://www.artint.ru/projects/alex.asp).
- InBASE - технология построения естественно-языковых интерфейсов к базам данных ([4], http://inbase.artint.ru).
- InDOC - технология интеллектуализации документооборота InDOC: автоматическая классификация и аннотирование документов ([5], http://www.artint.ru/projects/indoc.asp).
- Semp - технология разработки сложных интеллектуальных систем на основе интегрированной модели представления знаний Semp ([6], http://www.artint.ru/projects/semp-t.asp);
- Turtle - поисковая система с распределенной масштабируемой архитектурой ([7], www.turtle.ru)

Заключение

Закономерен вопрос: как можно решить такую огромную задачу? Очевиден традиционный путь: найти финансирование и взяться за работу. Таким путем, по-видимому, пошла команда Дугласа Лената (http://www.cyc.com/staff.html). Процесс создания системы Cyc в компании Cycorp Inc. идет уже более 15 лет. За это время была создана собственно система, оперирующая почти миллионом вручную введенных правил. Разработка продолжается, с промежуточными результатами можно ознакомиться на сайте проекта (http://www.cyc.com/products.html).

Финансирование работы такого масштаба может позволить себе либо очень богатое общество, либо (как уже отмечалось выше) сверхдержава, находящаяся в состоянии напряженности (например, холодной войны) с другой сверхдержавой.

Другой возможный путь состоит в том, чтобы использовать существующий потенциал и стремления сетевого сообщества.

Принципиальное отличие ситуации сорокалетней давности от текущей: гораздо большее количество людей, вовлеченных в информационное сообщество (если считать информационным сообществом в то время людей, занятых в отрасли информатики и кибернетики). Почему это так важно? В инициативе моделирования Сетевых ресурсов две краеугольные составляющие:

- знания экспертов (а в качестве эксперта вполне может выступать здесь и 15-летний подросток, в совершенстве знающий некоторый узкий срез Сети - времени на сетевую общественную жизнь у него гораздо больше, чем у, скажем, 35-летнего кандидата наук);
- технологии их формализации и использования в сервисах.

Если описание ресурсов будет распределено между членами информационного сообщества так же, как оно было распределено при создании этих ресурсов, и будет сопровождаться созданием и использованием эффективных технологий автоматизации описания, то за ограниченное время (5-10 лет) возможно заметное продвижение в решении этой насущной задачи.

Основной вектор движения – переход от поиска слов и страниц к поиску *информационных объектов*, представленных на страницах Сети. На наш взгляд, осуществление пилотных проектов в этой области возможно уже сегодня, на основе уже разработанных интеллектуальных технологий.

Литература

1. О.И.Боровикова, Ю.А.Загорулько. Организация порталов знаний на основе онтологий // Труды международного семинара Диалог'2002 по компьютерной лингвистике и ее приложениям. Протвино, 2002. -Т.2. -с.76-82.
2. В.А.Жигалов, Д.В.Жигалов, А.А.Жуков, И.С.Кононенко, Е.Г.Соколова, С.Ю.Толдова. Система Alex как средство для автоматизированной обработки текстов экспертом и перспективы ее развития // Труды Международного семинара Диалог’2002
3. В.А.Жигалов, Ю.А.Загорулько, А.С.Нариньяни, О.И.Россеева. Предел однородности поиска в Интернет. // Системная информатика. Выпуск 8: Теория и методология программирования – Новосибирск: Наука. Сибирская издательская фирма РАН, 2002. – с.29-71.
4. Жигалов В.А., Соколова Е.Г. InBASE: технология построения ЕЯ интерфейсов к базам данных // Труды Международного семинара Диалог’2001 по компьютерной лингвистике, Том 2, Аксаково, Июнь 2001, с. 123-135.
5. Ю.А. Загорулько, И.С. Кононенко, Ю.В. Костов, Е.В. Сидорова. Представление знаний в интеллектуальной системе документооборота // Труды 8-й национальной конференции по искусственному интеллекту - КИИ'2002. –Москва: Физматлит, 2002. -Т.2. -С.867-875.
6. Ю.А.Загорулько, И.Г.Попов. Описание сложных предметных областей на основе интеграции средств представления знаний // Труды международного семинара Диалог’97 по компьютерной лингвистике и ее приложениям. – Москва, 1997. – с.110-115.
7. Д.В.Крюков. Поисковая система Turtle. Физиология и анатомия. http://www.turtle.ru/db/architecture/