Московский Государственный Университет им. М.В. Ломоносова
Факультет Вычислительной Математики и Кибернетики

Козлов Д.Д.

Информационно-поисковые системы в Internet: текущее состояние и пути развития.

Технологический обзор

Москва 2000

Аннотация

В данной работе рассмотрена проблема поиска информации в Интернет, ее связь с традиционной проблемой поиска информации. Описаны новые задачи, отличающие проблему поиска в Интернет от традиционной проблемы поиска информации, дан обзор существующих методов поиска информации в Интернет. Рассмотрено применение технологии интеллектуальных агентов для поиска информации в Интернет, перечислены основные задачи, связанные с применением интеллектуальных агентов.

Содержание

1. Введение

«We are drowning in information but starved of knowledge»
John Naisbitt [1]

В последнее время сеть Интернет стала основным мировым хранилищем информации. С ростом объемов хранимых данных актуализировалась проблема информационного поиска.
В данном обзоре рассмотрена проблема информационного поиска в Интернет, ее отличия от традиционной проблемы информационного поиска, обусловленные особенностями Интернет как информационной системы. Рассмотрены особенности существующих ИПС в Интернет и задачи связанные с развитием ИПС в Интернет. В работе рассмотрено применение интеллектуальных агентов для решения задачи информационного поиска в Интернет, выделены основные задачи связанные с применением интеллектуальных агентов для ИП в Интернет.

Во второй главе рассмотрена традиционная проблема информационного поиска, вводится терминология, рассматриваются общая архитектура и организация работы автоматизированных информационно-поисковых систем, обсуждаются основные параметры, используемые для оценки эффективности информационно-поисковых систем.

В третьей главе рассмотрены основные особенности задачи информационного поиска в Интернет, отличия от традиционной задачи. Проводится анализ существующих решений, рассматриваются направления развития.

В четвертой главе рассмотрено применение интеллектуальных агентов информационного поиска в Интернет. В данной главе даются основные определения, касающиеся интеллектуальных агентов, рассматриваются примеры применения агентов для информационного поиска в Интернет. В заключение рассматриваются основные тенденции развития агентов для поиска информации в Интернет, и дается обзор основных задач.

[1] Позаимствовано из [1].

2. Традиционная проблема информационного поиска

Начиная с середины пятидесятых годов, поводились активные исследования по проблеме информационного поиска [2]. Хранение и поиск информации производились с помощью ЭВМ. С распространением в девяностые годы Интернет и World Wide Web, методы хранения информации существенно изменились, что позволило говорить о традиционной проблеме информационного поиска и о проблеме информационного поиска в Интернет [6].

2.1 Основные определения

Мидоу [3] определяет информационный поиск (ИП) как «родовое понятие для поиска данных, поиска фактов и поиска документов». В рамках данной работы под *информационным поиском*[2] подразумевается, поиск документов (статей, книг, отчетов) в массиве в соответствии с критериями, предложенными лицом, сформулировавшим запрос. Согласно [4] *информационно-поисковая система* (ИПС) – программная система для хранения, поиска и выдачи интересующей пользователя информации. В соответствии с используемым определением информационного поиска, далее в работе под термином ИПС будет подразумеваться *документальная ИПС* – ИПС, предназначенная для отыскания документов, содержащих необходимую пользователю информацию.

2.2 Функционирование информационно-поисковой системы

Общая схема[3] функционирования традиционной ИПС представлена на Рис. 1. Основными процессами в ИПС являются индексирование документов и поиск документов по запросу пользователя.

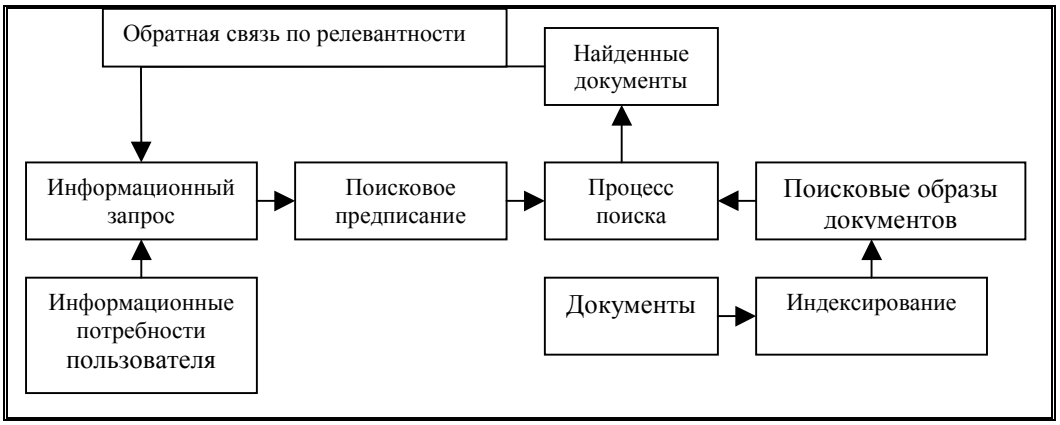

Рис. 1 Схема функционирования ИПС.

Процесс информационного поиска происходит следующим образом. Пользователь выражает свои информационные потребности в виде

[2] Далее по тексту, для краткости наряду с термином «информационный поиск» может использоваться термин «поиск».
[3] Согласно [7].

специального текста - *информационного запроса*[4] к ИПС. Система формирует из информационного запроса *поисковое предписание*, переводя запрос на *информационно-поисковый язык* (ИПЯ). ИПЯ представляет собой формальный язык (по Хомскому [44]), который используется внутри ИПС для представления пользовательского запроса и хранимых документов. Описание документа на ИПЯ называется *поисковым образом* документа. В процессе поиска ИПС должна выбрать из массива документов те, которые *содержательно релевантны* запросу, то есть соответствуют информационным потребностям пользователя, выраженным в запросе. Такое определение релевантности не формально, поэтому определяют *формальную релевантность*, как соответствие, определяемое алгоритмически, путем сравнения поискового предписания и поискового образа документа ([4], стр. 46). *Критерий выдачи* документа – формальное правило, определяющее степень формальной релевантности поискового образа документа и поискового предписания, по которому принимается решение о выдаче некоторого документа в ответ на информационный запрос.

В процессе *индексирования,* для каждого документа, хранящегося в системе, строится поисковый образ. Различают два основных подхода к построению поискового образа - *приписывающее* (*assigned*) и *выводящее* (*derived*) индексирование [7]. В первом случае в процессе индексирования документу присваивается набор ключевых слов из некоторой классификационной системы, и документ помещается в общую классификацию ([5], глава 8). Во втором случае из документа выбирается набор ключевых слов и объявляется поисковым образом, с которым далее работает ИПС [8].

На то, как функционирует ИПС, наибольшее влияние оказывают тип индексирования, тип поискового образа и подход к оценке релевантности. В следующем разделе рассматриваются базовые подходы к внутренней организации ИПС.

2.3 Модели поиска

2.3.1 Модель поиска, базирующаяся на классификации

Модель поиска, базирующаяся на классификации, формально определена в [10]. Частным случаем такой модели является тематический каталог библиотеки. Все поступающие в ИПС документы распределяются по темам, организованным в виде иерархической классификации [11] (приписывающее индексирование). Единственным отношением между классами является отношение включения. Классы между собой не пересекаются. Особенностями работы ИПС с такой организацией является то, что список запросов к системе определен заранее в виде тем классификатора и система не позволяет производить поиск по пересечению классов и упорядочение документов по релевантности.

[4] Далее по тексту, для краткости наряду с термином «информационный запрос» может использоваться термин «запрос».

2.3.2 Булева модель

В булевой модели [12] запрос представляется в виде формулы логики высказываний, где в качестве высказываний используются ключевые слова, высказывание истинно, тогда и только тогда, когда ключевое слово входит в состав документа. Критерием выдачи документа является истинность заданной в запросе формулы. Особенностью метода является невозможность упорядочить результаты запроса по релевантности. Данное ограничение снято в расширенной булевой модели [9].

2.3.3 Векторная модель

В векторной модели [8] документы $d \in D$ идентифицируются с помощью множества независимых атрибутов A, например ключевых слов (выводящие индексирование). Каждый документ представляется в виде вектора $d_i = (a_{i1}, a_{i2}, \ldots, a_{it)}$, где a_{ij} – вес j-ого атрибута в документе. Аналогично, запрос $q = (q_1, q_2, \ldots, q_{t)}$, где q_j – вес j-ого атрибута в запросе. Тогда формальную релевантность можно определить, например[5], как скалярное произведение векторов d_i и q. Отличительной особенностью векторной модели является возможность упорядочивания документов по релевантности запросу. Наиболее популярным на практике является вариант векторной модели, где формальная релевантность вычисляется по формуле TFIDF [8]. Особенностью этого варианта является то, что при вычислении формальной релевантности используется статистическая информация обо всех документах, хранящихся в системе.

2.3.4 Интерактивный поиск

Интерактивный поиск [10] позволяет обеспечить более гибкую, по сравнению с другими методами поиска, реакцию на нужды и желания пользователя. В основе интерактивного поиска лежит обратная связь по релевантности – метод интерактивной работы пользователя с ИПС, при котором ИПС представляет пользователю первоначальные результаты обработки запроса, а пользователь указывает какие из выданных документов релевантны, а какие – нет, после чего ИПС корректирует запрос и продолжает поиск. Интерактивный поиск строится обычно на основе векторной модели. В случае работы высококвалифицированного пользователя он способен давать значительно лучшие результаты, чем автономный поиск[6].

2.4 Параметры оценки информационно-поисковых систем

Для оценки эффективности работы ИПС, помимо стандартных параметров, используемых для оценки эффективности вычислительных систем, используются специализированные параметры для оценки качества работы ИПС. Среди них основными являются следующие параметры [4].

[5] На практике применяются более сложные функции см. [8].
[6] [5], раздел 4.5.

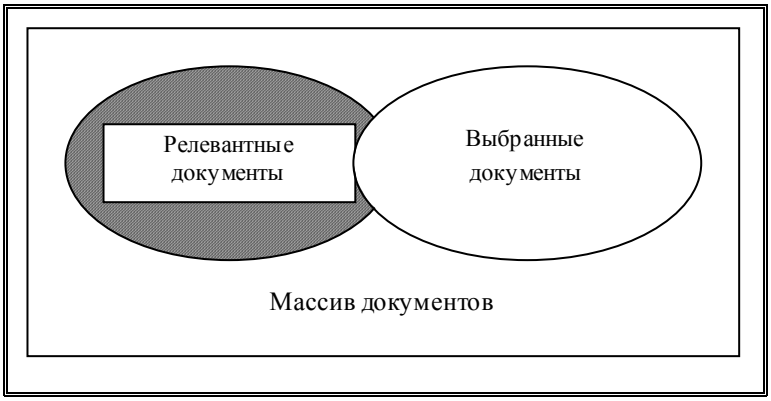

Рис. 2 Подразделение массива документов при обработке запроса.

Точность выдачи – отношение числа выданных релевантных документов к сумме числа выданных релевантных и числа выданных нерелевантных документов.
Полнота выдачи – отношение числа выданных релевантных документов к сумме числа выданных релевантных и числа невыданных релевантных документов.
Потери информации – отношение числа невыданных релевантных документов к сумме числа выданных релевантных и числа невыданных релевантных документов.
Информационный шум – отношение числа выданных нерелевантных документов к сумме числа выданных релевантных и числа выданных нерелевантных документов.
Чувствительность – отношение числа выданных релевантных документов к сумме числа выданных релевантных и числа невыданных релевантных документов.
Специфичность – отношение числа невыданных нерелевантных документов к сумме числа выданных нерелевантных и числа невыданных нерелевантных документов.

На практике для сравнения ИПС используются усредненные графики зависимости полноты от точности [5].
Чтобы избежать сравнения пар полнота, точность используются однозначные оценки [5]. Одной из таких оценок является *E-мера* [7], позволяющая избежать сравнения пар полнота, точность за счет введения отношения их значимости.

$E(b) = 1 - \frac{(b^2 + 1.0)PR}{b^2 P + R}$, где P – точность, R – полнота , b – отношение значимости полноты и точности.

3. Информационный поиск в Internet

В рамках данной работы сеть Интернет будет рассматриваться, согласно введенной терминологии, как *документальная информационная система*, то есть информационная система, ориентированная на хранение документов.
Традиционные ИПС осуществляют как поиск, так и хранение документов. Поиск документов производится только по хранимому внутри ИПС массиву документов. В отличие от традиционных ИПС, ИПС для поиска информации в Интернет не могут осуществлять функцию хранения документов, что приводит к необходимости другого подхода к организации работы ИПС.

3.1 Особенности Интернет, как хранилища информации

1. Развитие Интернет как информационного хранилища происходило без учета потребности поиска документов. В результате в Интернет, в отличие от традиционных ИПС, где система хранения документов ориентирована на эффективный поиск, система хранения документов является заданной априори относительно задачи информационного поиска.
2. Интернет представляет собой децентрализованное хранилище документов, не имеющее единого управления организацией и развитием. Сеть Интернет гетерогенна, используются не только различные платформы, но и различные стандарты представления информации. Интернет объединяет как современные, так и унаследованные (legacy) системы. Часть информации хранится в виде, отличном от текста (мультимедиа).
3. Социальная гетерогенность (по данным [13] 83% - коммерческая информация, 6% - научно-образовательная). Большой социальный разброс по авторам, аудитории, читателям.
4. Интернет – распределенное хранилище, время доступа к различным его частям неодинаково и может существенно превосходить время доступа к локальному документу.
5. Объем документов в Интернет (более 800 миллионов документов[7] на февраль 1999 года [13] и более 1 биллиона документов на начало 2000 года [16]) превышает объемы самых больших ИПС и постоянно увеличивается. Большая часть информации, хранимой в Интернет содержится в базах данных (эта часть называется Deep Web [14], см. раздел 3.3) и недоступна для большинства существующих промышленных ИПС. По оценкам [14], количество документов, хранящихся в базах данных, превышает количество документов, известных промышленным ИПС приблизительно в 500 раз.
6. Отсутствие метаинформации[8] об организации отдельных информационных ресурсов Интернет, об их взаимосвязи и о смысловом содержании хранимой информации (по данным [13] только 34% используют HTML тэги “description” и “keywords” и только 0.3% Dublin Core).

[7] Это оценка части Интернет, так называемой publicly indexable Web [13] или Surface Web [14], то есть документов, которые могут быть проиндексированы промышленными ИПС в Интернет (см. раздел 3.2.2).
[8] Под метаинформацией в данной работе, согласно [44], понимается информация об информации, которая описывает хранимую информацию для обеспечения ИП.

7. В отличие от большинства традиционных ИПС, для Интернет характерна динамичность информации (частое появление, исчезновение, изменение) по данным [16] ежедневно в Интернет появляется 1.5 миллиона документов.
8. Реклама на Web-сайтах.
9. Дублирование информации в Интернет (по оценкам [16] около четверти всех сайтов представляют так называемые “зеркала”).

3.2 Существующие ИПС в Интернет.

В данном разделе рассматриваются основные классы промышленных ИПС для поиска информации в Internet.

3.2.1 ИПС, базирующиеся на классификации

Общая схема работы ИПС базирующейся на классификации показана на Рис. 3.

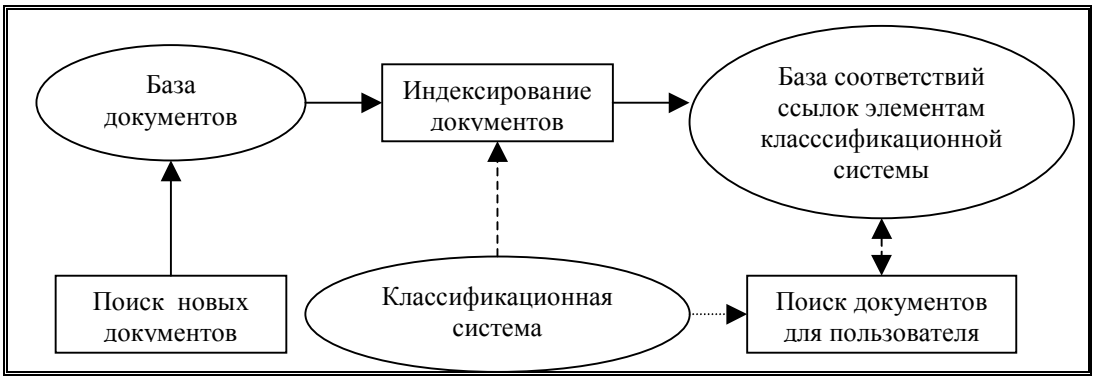

Рис. 3 Общая схема работы ИПС базирующейся на классификации.

В целом, схема работы такой ИПС в Интернет аналогична схеме работы традиционной ИПС, базирующейся на классификации. Основным отличием является появление процесса поиска новых документов. В традиционных ИПС новые документы вводятся в систему хранения оператором и индексируются, в ИПС, ориентированных на работу в Интернет ввод новых документов осуществляется либо вручную оператором, либо автоматически с помощью специальной программы обхода Интернет – индексирующего робота (в английской терминологии также используются термины spider и crawler). Работа индексирующего робота описана в разделе 3.2.2.

Применение для ИП в Интернет ИПС базирующихся на классификации эффективно в случае, когда классификационная система построена по узкой предметной области, как, например, Research Index Computer Science Directory [17]-[19]. Основных недостатков два:

1. Для качественного поиска они вынуждены выкачивать из Интернет все документы для индексирования и хранить их у себя, как это сделано в ResearchIndex [19], что приводит к большому объему хранимой информации, высокой нагрузке на сеть и необходимости постоянно обновлять информацию в базе;

2. поиск документов пользователем может осуществляться только по используемой классификационной системе[9].

3.2.2 ИПС, базирующиеся на поиске по ключевым словам

ИПС базирующиеся на поиске по ключевым словам (далее системы поиска по ключевым словам, в английской терминологии search engines) позволяют искать Web-страницы по их содержанию, формулируя запрос в виде набора ключевых слов, которые должны присутствовать в документе. В настоящее время, системы поиска по ключевым словам представляют собой наиболее распространенные ИПС в Интернет.

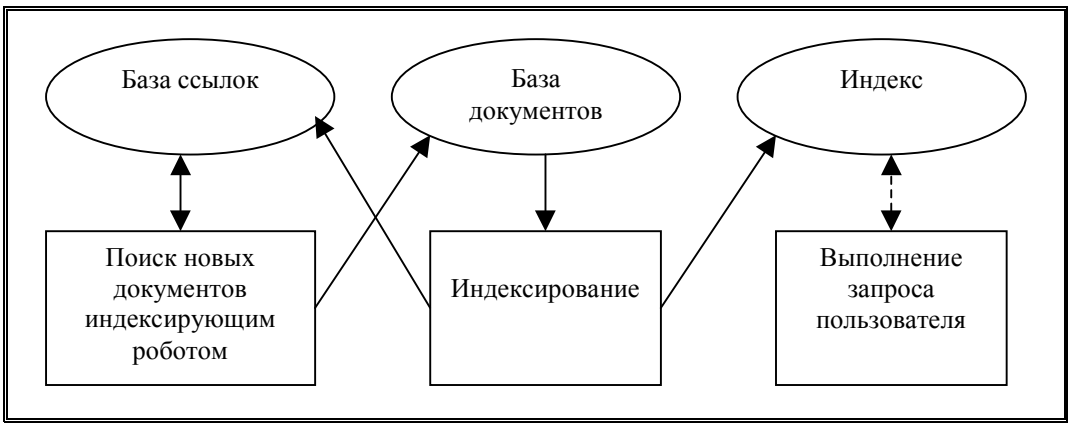

Рис. 4 Общая схема работы системы поиска по ключевым словам.

На Рис. 4 показана общая схема работы системы поиска по ключевым словам.[10] Основными процессами в работе системы являются поиск новых документов индексирующим роботом, индексирование найденных документов и выполнение запроса пользователя.

Индексирующий робот представляет собой автономный процесс, постоянно или периодически обновляющий и пополняющий базу документов. Изначально роботу дается список Web-серверов, которые необходимо проиндексировать. В процессе работы индексирующий робот осуществляет обход Web-серверов по гиперссылкам между страницами и собирает все найденные документы в базу документов, а ссылки – в базу ссылок. Таким образом, на основе начального списка серверов строится документов для всех достижимых Web-страниц. Периодически, индексирующий робот проверяет хранящуюся информацию на корректность и целостность путем повторного обхода проиндексированных страниц.
По известным ИПС документам строится индекс, позволяющий эффективно осуществлять поиск по ключевым словам (более подробно об этом см., например, [21]). Дальнейшее хранение всего документа после индексирования не нужно, для экономии дискового пространства хранится короткий поисковый образ.

---

[9] Здесь есть тенденция дополнения возможности булева поиска по ключевым словам (ResearchIndex).
[10] Данное описание основано на [14], [20], [21].

Запрос пользователя представляет собой набор ключевых слов с булевыми связками (см. 2.3.2). Выбор документов по запросу осуществляется с помощью индекса. ИПС возвращает в ответ на запрос список ссылок на документы.

Достоинством систем поиска по ключевым словам является простота использования. К недостаткам можно отнести следующие особенности.

1) В ответ на запрос выдается много нерелевантной информации. Это происходит в частности из-за того, что с помощью списка ключевых слов практически сложно точно сформулировать информационные потребности пользователя, а пользователи чаще всего дают короткие запросы из 1-2 ключевых слов, используя при этом только базовые возможности для формулирования запроса, так как для использования развитых возможностей формулирования запроса требует знания специального синтаксиса, а у каждой из систем синтаксис свой.
2) Индексирующие роботы сильно загружают сеть. Так как робот не имеет возможности перемещаться по сети он вынужден скачивать большие объемы информации для локальной обработки. Хранимый ИПС объем информации исчисляется терабайтами и эту информацию надо постоянно обновлять.
3) Невозможность работы с часто изменяемой информацией. Обход всех документов для проверки их неизменности продолжается несколько недель и даже месяцев [14].
4) По данным [13] охват Интернет любой из имеющихся систем поиска по ключевым словам не превышает 16%. Не охвачены многие информационные источники в Интернет (FTP, GOPHER,ARCHIE). Алгоритм поиска новых документов на основе обхода ссылок не позволяет достичь некоторых Web-страниц.
5) Возможность работы пользователя только в интерактивном режиме.
6) Промышленные ИПС индексируют Интернет неравномерно, предпочтение отдается более популярным Web-серверам, коммерческой тематике [13].

В настоящее время проводятся исследования по повышению эффективности систем поиска по ключевым словам. Основное внимание уделено следующим направлениям развития: повышение качества оценки релевантности документов, например, PageRank [21] строит оценку релевантности с использованием рейтинга цитирования документов; построение модели пользователя для выдачи более релевантной для конкретного пользователя информации.

3.2.3 Метасистемы

Метасистемы [22]-[25] для ИП в Интернет, являются надстройками над существующими системами поиска по ключевым словам. Они позволяют преодолеть следующие недостатки промышленных систем поиска по ключевым словам.

- Малый охват Интернет. По данным [13] ни одна из существующих систем поиска по ключевым словам не охватывает более 16% Интернет. Использование 11 поисковых систем позволяет охватить 42% Интернет [11].
- Наличие в результатах обработки запроса ссылок на отсутствующие документы.

[11] Имеется ввиду только Surface Web (см. раздел 3.3).

- Низкое качество оценки релевантности выдаваемых системами поиска по ключевым словам документов.
- Наличие мусора (spam), злоумешленно внесенного в результаты обработки запроса.
- Большое количество документов выдаваемых в ответ на запрос.
- Различие пользовательских интерфейсов систем.

Метасистемы являются посредниками между пользователем и системами поиска по ключевым словам. Они получают от пользователя запрос в виде булевой комбинации ключевых слов и некоторой дополнительной информации касательно типа искомого документа и решают, к каким системам поиска по ключевым словам его перенаправить, далее запрос к каждой из переформулируется в с учетом синтаксиса конкретной системы, уточняется, (например, дополнительными ключевыми словами [22]) и посылается системе. Параллельно опрашивается несколько систем поиска по ключевым словам. По полученным ссылкам считываются документы, анализируются на предмет соответствия запросу и сортируются по своему ([22]), более точному, чем в системах поиска по ключевым словам алгоритму оценки релевантности. Метасистемы предоставляют также более развитые, по сравнению с системами поиска по ключевым словам, возможности по сортировке результатов запроса.

3.2.4 Общие особенности существующих ИПС в Интернет.

Для всех рассмотренных классов ИПС в Интернет характерны следующие особенности.

- Малый объем охвата Интернет. Для систем поиска по ключевым словам объем охвата Интернет за последние три года постоянно снижается ([13],[14]). Эта проблема менее актуальна для метасистем, но тенденция использования технологий баз данных в WWW усложняет ситуацию, расширяя часть Интернет, не видимую для рассмотренных систем (более подробно см. в разделе 3.3).
- Большинство рассмотренных ИПС содержат большое количество ссылок на несуществующие документы [26]. Период обновления информации об известном документе может составлять месяц. Данная проблема решается только в метасистемах путем непосредственной проверки наличия документа перед выдачей ссылки.
- Все рассмотренные системы ориентированы на поиск популярных ресурсов и на обработку наиболее популярных запросов. Для редких запросов и малопопулярных ресурсов результат ИП имеет более низкое качество, чем для популярных.
- Все системы ориентированы на разовое удовлетворение запроса пользователя и не имеют развитых средств настройки на конкретного пользователя.
- Для всех систем проблемой является оценка релевантности документа, в этой области сосредоточены основные исследования.
- Объем данных (в терабайтах), хранящихся в нескольких наиболее крупных ИПС превышает объем информации, хранящейся в Интернет, а около 50% трафика в Интернет приходится на роботов ИПС ([14]).

Важной тенденцией существующих развития ИПС в Интернет является тенденция интеграции ИПС основанных на классификации и систем поиска по ключевым словам. В результате получаются тематические ИПС типа ResearchIndex, по сути, – электронные библиотеки, построенные на основе документов, собранных в Интернет, обеспечивающие высокое качество поиска.

3.3 Направления развития информационного поиска в Интернет

Исходя из особенностей Интернет, как системы хранения информации возникает вопрос о возможности эффективного ИП в Интернет при существующей организации Интернет. Наибольшими препятствиями на пути разработки эффективных алгоритмов ИП в Интернет являются малая структурированность информации в Интернет, высокая скорость изменения информации и отсутствие метаинформации.

Информационные источники в сети Интернет можно разделить на две категории Всемирная Паутина (World Wide Web, WWW, Web) и унаследованные системы. Всемирная Паутина объединяет в себе большую часть Интернет, внешнее представление информации на Web-серверах стандартизовано, доступ к информации осуществляется единообразно. Унаследованные системы (например, Gopher, Archie, Usenet News и т.п.) не соответствуют стандартам представления информации, используемым в WWW. Часть этих систем постепенно отмирает. Многие из систем имеют шлюзы для представления информации в виде принятом в WWW. Поэтому в данной работе основное внимание уделено поиску в WWW.

В 1996 году в работе [15] Etzioni высказал следующий тезис: «Структурированность WWW достаточна для построения эффективных систем поиска информации». С 1996 года Всемирная Паутина существенно изменилась. Важно отметить следующие тенденции ее развития: применение технологий баз данных для создания Web-сайтов, появление средств динамической генерации Web-страниц, коммерциализация WWW, появление электронной коммерции, и, как следствие, постоянное снижение охвата WWW системами поиска по ключевым словам [13].

На текущий момент в WWW можно выделить две основные части.

- Surface Web [14] – часть WWW, доступная большинству промышленных ИПС, индексирующие роботы которых обходят Интернет по ссылкам. Относительный размер этой части постоянно уменьшается. Сюда относятся статические и динамические Web-сайты в которых возможна навигация по ссылкам.
- Deep Web [14] – часть WWW, недоступная для большинства промышленных ИПС из-за необходимости явно формулировать запрос на получение требуемого документа (нет ссылок). Относительный размер ее постоянно увеличивается. Сюда относятся электронные библиотеки, ИПС для поиска по ключевым словам, Web-интерфейсы к базам данных.

При ИП по Surface Web основными проблемами являются социальная гетерогенность документов, высокая скорость изменения информации при почти полном [13] отсутствии метаинформации.

При ИП по Deep Web основными проблемами являются отсутствие метаинформации по информационным источникам и отсутствие единого интерфейса для доступа к информации. Данные задачи могут решаться только со стороны поставщика информации, путем стандартизации интерфейсов доступа и описаний метаинформации. Такие работы проводятся в рамках Open Archives Initiative [42] и W3C Metadata Activity [43].

Таким образом, решение проблемы ИП в Интернет на сегодняшний день состоит не только в построении эффективных ИПС, но и в изменении структурной организации информации Интернет.

3.3.1 Задачи, связанные с развитием информационно-поисковых систем в Интернет.

1. Развитие средств семантического анализа текстов на естественном языке. Сюда относятся задачи реферирования текстов, рубрикации, кластеризации, смыслового поиска по текстам. Интернет вносит большее разнообразие в качество и социальную ориентированность текстов по сравнению с традиционными системами, что существенно усложняет задачу семантического анализа.

2. Огромные скорости роста Интернет привели к тому, что на практике не существует стандарта для организации информации в Интернет, что усложняет доступ к имеющейся информации. Задача, заключается в необходимости обеспечения единообразного доступа ко всем информационным ресурсам Интернет. С одной стороны, задачу можно рассматривать как необходимость выработки единого стандарта для организации данных и приведения всей хранящейся в Интернет информации к этому стандарту, с другой стороны, задачу можно рассматривать как необходимость обеспечения возможности получения метаинформации о любой информации, хранящейся в Интернет.

3. Обеспечение контролируемого и безопасного доступа к хранящейся в Deep Web информации.

4. Покрытие ИПС максимального количества информации в Интернет, устранение дублирования ИПС с целью экономии ресурсов. Организация единой распределенной технологии поиска.

5. Поддержание информации в ИПС в соответствии с реальностью. Максимально быстрый учет изменений документов Интернет.

4. Использование интеллектуальных агентов для информационного поиска в Интернет

В последнее время в контексте проблемы ИП наблюдается повышенный интерес к применению для ИП в Интернет так называемых программных агентов и их разновидности – интеллектуальных агентов.

4.1 Общие понятия о программных агентах

В самом общем виде *программный агент*[12] (*software agent*) можно определить, как проблемно-ориентированную программную сущность, которая может выполнять за человека рутинную работу, которую человек мог бы выполнять сам, при наличии времени.[13]

Более конкретно, термин программный агент можно определить[14] как проблемно-ориентированную программу, обладающую следующими свойствами.

- *Автономность* (autonomy). Агент работает в окружающей среде, состояние которой он воспринимает с помощью сенсоров и может изменять средствами воздействия в процессе работы. Агент выполняет воздействия на окружающую среду согласно собственной цели, и контролирует собственные действия путем наблюдения изменений окружающей среды.
- *Восприимчивость* (reactivity, responsiveness). Агент наблюдает изменения окружающей среды и своевременно отвечает на них.
- *Целеориентированность* (goal orientedness). Агент должен не только отвечать на внешние воздействия, но и выполнять поставленную задачу.
- *Рациональность поведения* (rationality). Агент не должен делать действий, препятствующих выполнению поставленной задачи. Для каждой возможной воспринимаемой последовательности *идеально рациональный агент* [1] делает то действие, которое принесет наибольший успех (результат).

Кроме обязательных основных свойств агент может обладать одним или несколькими из следующих дополнительных свойств.

- *Коммуникабельность* (*social ability*). Возможность общаться с человеком и с другими агентами для достижения своей цели и для помощи другим агентам.
- *Предусмотрительность (proactiveness)*. Возможность не только выполнять поставленную задачу, но и должен собирать при этом полезную для пользователя информацию, относящуюся к запросу пользователя.
- *Адаптируемость поведения*. Агент должен уметь настраиваться под привычки и методы работы конкретного пользователя.
- *Мобильность* - возможность перемещения агента в сети.
- *Обучаемость* - возможность приобретения агентом знаний в процессе работы с пользователем и с другими агентами.

[12] Далее по тексту термин «агент» используется как синоним термина «программный агент».
[13] Ted Selker, IBM Almaden Research Center, позаимствовано из [1]
[14] Данное определение основано на [27],[28].

Частным случаем программных агентов являются *интеллектуальные агенты* (intelligent agent, ИА), то есть агенты, обладающие свойством интеллектуальности. Согласно [29] «интеллектуальность агента является степенью способности к рассуждению и обучаемости. Интеллектуальность подразумевает, как минимум, возможность задавать пользовательские предпочтения агенту и наличие у агента механизма рассуждения, чтобы действовать в соответствии с этими предпочтениями. Более высокий уровень интеллектуальности подразумевает наличие у агента модели пользовательских потребностей и механизма поиска способа их удовлетворения. Дальнейшее развитие интеллектуальности связано с обучаемостью и адаптацией к окружающей среде, как с точки зрения пользовательских потребностей, так и с точки зрения доступных ресурсов».

В настоящее время разрабатывается теория построения ИА и соответствующий инструментарий (обзор в [30]).

4.2 Предпосылки к использованию интеллектуальных агентов для информационного поиска в Интернет

Как отмечалось выше (см. раздел 3.3.1) важными задачами в развитии ИП в Интернет являются

- обеспечение эффективного взаимодействия пользователя с ИПС, в частности упрощение процесса формирования запроса, настройка ИПС на индивидуальные особенности пользователя, предотвращение перегрузки пользователя большими объемами найденной информации;
- повышение качества поиска и индексирования документов ИПС;
- накопление, обобщение и повторное использование опыта ИП различных пользователей.

Применение интеллектуальных агентов для решения вышеперечисленных задач является перспективным по следующим причинам[15].

- Свойства, которыми обладают ИА, необходимы для построения персональных помощников и интеллектуальных пользовательских интерфейсов.
- Свойства мобильности и автономности работы агентов позволяют предложить новый подход к поиску новых документов и индексированию документов в ИПС.
- Технология на основе взаимодействующих программных агентов предоставляет основу для моделирования систем высокой сложности, позволяя разбивать задачу на подзадачи меньшей сложности. Модель системы при этом близко соответствует модели реального мира за счет использования понятия агента.
- Агенты позволяют разрабатывать систему на высоком уровне абстракции, скрывая детали реализации и проблемы низкоуровнего взаимодействия.
- Каждая подзадача, которой соответствует агент, может быть решена наиболее оптимальным образом, решение может быть легко модифицировано

[15] Составлено на основе [1], [31]

и повторно использовано, что особенно актуально ввиду динамичности Интернет.

4.3 Примеры использования интеллектуальных агентов для информационного поиска в Интернет

Можно выделить следующие основные области применения ИА для ИП в Интернет.

- Фильтрация информации. Агент WebMate [32] позволяет сократить количество Web-серверов, просматриваемое пользователем в поисках новостей. Он формирует персональный дайджест новостей по интересующим пользователя темам.

- Предоставление пользователю помощи при работе с некоторым информационным источником. Агент-гид является Web Watcher [33] который позволяет задать интересующую пользователя тему и в процессе навигации по Web-сайту рекомендует ссылки, которые относятся к этой теме.

- Формирование модели пользователя и повышение качества ИП на основе этой модели. Персональный информационный помощник Personal Web Watcher [34] составляет модель информационных потребностей пользователя на основе наблюдения за навигацией пользователя по Интернет и потом использует эту модель для рекомендации потенциально интересных пользователю ссылок.

- Поиск информации по запросу пользователя. Агент CiteSeer [35] позволяет частично автоматизировать процесс поиска научных статей в Интернет. Он ищет статьи, выкачивает их и сортирует по предполагаемой релевантности теме, заданной пользователем. Информационный помощник Softbot [28] позволяет автоматизировать рутинные задачи пользователя по управлению файлами. При этом пользователь формулирует задание в высокоуровневой форме, а агент формирует соответствующую цепочку действий и выполняет ее.

- Повторное использование опыта пользователей по поиску информации [40]. Основной идеей подхода является наличие у каждого пользователя персонального агента, накапливающего опыт поиска информации пользователем. При необходимости искать что-то новое агент использует опыт, накопленный другими агентами в ходе наблюдения за их пользователями.

4.4 Тенденции развития интеллектуальных агентов для информационного поиска в Интернет

Исходя из приведенного выше рассмотрения примеров применения ИА для ИП в Интернет, можно выделить два направления использования агентов:

- агенты как интеллектуальные проблемно-ориентированные системы и
- агенты как модель для построения сложных динамических систем.

Подробное рассмотрение второй области применения агентов выходит за рамки данной работы. В первом же направлении дальнейшее развитие связано с повышением интеллектуальности агентов и с расширением круга решаемых задач. Актуальность этого развития связана с тем, что в последние годы объем информации в Интернет достиг такого размера, что пользователь просто теряется в информационном пространстве, агрессивное распространение рекламы (spam) в Интернет очень мешает поиску информации вручную. Поэтому возникла потребность снижения нагрузки на человека, путем перенесения части проблем связанных с поиском и обработкой информации имеющейся в Интернет на компьютер.

Для решения этой задачи можно предложить построение интеллектуального агента - *информационного представителя*, в идеале полностью замещающего человека в информационном пространстве Интернет. Информационный представитель представляет информационные интересы своего пользователя в Интернет, является посредником между пользователем и Интернет и сочетает в себе следующие функции.

- Поиск информации по заданию пользователя (одновре́менный и постоянный).
- Возможность доставки информации на компьютер пользователя.
- Фильтрация потоков информации.
- Предоставление вновь появляющейся в Internet информации, которая может заинтересовать пользователя.

Таким образом, информационный представитель становится «глазами и ушами» пользователя в Internet, избавляя его от рутинных операций поиска нужной информации и защищая от огромного потока ненужной информации.

4.5 Задачи, возникающие при построении персональных информационных агентов

1. Осуществление единообразного доступа ко всем информационным ресурсам Интернет.
2. Осуществление контролируемого, безопасного для информационного источника доступа. Эта задача особенно актуальна в следующем контексте. ИА осуществляющий поиск информации на некотором корпоративном Web-сервере создает очень высокую загрузку сети при просмотре информации. Наиболее выгодным, с точки зрения уменьшения сетевого трафика, является выполнение агента на Web-сервере, что позволит передавать по сети только нужную информацию. Такой способ организации вычислений сопряжен с угрозой безопасности сервера и запрещен практически во всех существующих системах. Задача состоит в том, чтобы разработать такой метод выполнения, который бы гарантировал безопасность выполнения агентов на сервере и давал владельцам сервера ограничивать доступ к различным данным для разных агентов. Данная задача решается, например, в рамках проекта Aglets Workbench [36].
3. Обеспечение мобильности агента, например для сокращения сетевого трафика [36].
4. Обеспечение взаимодействия агентов друг с другом с целью обмена информацией. Данная задача решена в рамках проекта ARPA Knowledge Sharing Effort [37] апробирована в рамках проекта JKQML [38]. Результатом проектов являются предложения по организации взаимодействия агентов

(проект стандарта) и обмену знаниями между агентами и набор библиотек для обеспечения этого взаимодействия.

5. Разработка моделей построения распределенных приложений на основе многоагентных систем (обзор см. в [39]).
6. Создание новых средств обеспечения интеллектуальности агента в условиях такой сложной и динамичной среды как Интернет.
   Создание модели окружающей среды внутри агента и средств для поддержания актуальности модели. Модель должна адекватно представлять окружение агента и должна быть пополняема, чтобы обеспечить возможность рациональной работы агента (см. п. 4.1).
   Создание модели пользователя и его информационных потребностей.
   Создание алгоритма планирования выполнения запроса. Задачей минимум алгоритма является, построение плана выполнения запроса, а задачей максимум является построение оптимального плана на основе имеющейся у агента информации.
   Разработка аппарата принятия решений на основе недостоверной и неполной информации.
   Данные задачи исследовались в рамках проектов Personal WebWatcher [34] Softbot [28].
7. Создание инструментальных средств разработки агентов [38], [36].

5. Заключение

Данный обзор показывает, что проблема информационного поиска в Интернет, имеет ряд существенных отличий от традиционной проблемы информационного поиска, обусловленных особенностями Интернет как информационной системы. В обзоре рассмотрены основные особенности Интернет как документальной информационной системы и их влияние на ИП. Рассмотрены особенности существующих ИПС в Интернет и задачи связанные с развитием ИПС в Интернет. В работе рассмотрено применение интеллектуальных агентов для решения задачи информационного поиска в Интернет, выделены основные задачи связанные с применением интеллектуальных агентов для ИП в Интернет.

В заключение автору хотелось бы поблагодарить проф. Смелянского Р.Л. за помощь при создании этой работы и Козлова Д.Г. за ценные замечания по стилистике.

6. Литература

[1] Hermans B., Intelligent Software Agents on the Internet: an inventory of currently offered functionality in the information society and a prediction of future developments, http://www.hermans.org/agents, 1996

[2] Salton G., Historical Note: The Past Thirty Years in Information Retrieval, Cornell University Technical Report 87-827

[3] Мидоу Ч., Анализ информационно-поисковых систем, М., Мир, 1970

[4] Мальковский М. Г., Грацианова Т. Ю., Полякова И. Н., Прикладное программное обеспечение: системы автоматической обработки текстов, Учебное пособие для студентов факультета ВМиК МГУ, Москва, МГУ, 2000

[5] Солтон Дж., Динамические библиотечно-информационные системы, М., Мир, 1979.

[6] Etzioni O., The World Wide Web: quagmire or gold mine?, Communications of the ACM, Nov.'96.

[7] C.J. van Rijsbergen, Information Retrieval, London, Butterworths, 1979

[8] Salton G., Buckley C., Term Weighting Approaches in Automatic Text Retrieval, Cornell University Technical Report 87-881

[9] Salton G., Fox E., Wu H., Extended Boolean Information Retrieval, Cornell University Technical Report 82-511

[10] Сэлтон Г., Автоматическая обработка, хранение и поиск информации, М., «Советское радио», 1973

[11] Черный А. И., Введение в теорию информационного поиска, М., Наука, 1975

[12] Salton G., Mathematics and information retrieval Cornell University Technical Report 78-332

[13] Lawrence S., Giles C., Accessibility of Information on the Web, Nature, vol.400, pp. 107-109, 1999

[14] Bergman K., The Deep Web: Surfacing Hidden Value, BrightPlanet.com LLC, http://www.completeplanet.com/Tutorials/DeepWeb/index.asp

[15] Etzioni O., The World Wide Web: quagmire or gold mine?, Communications of the ACM, November1996.

[16] Inktomi Corp., Web Surpasses One Billion Documents, press release issued January 18, 2000, http://www.inktomi.com/new/press/billion.html

[17] Lawrence S., Bollacker K., Giles C., Digital Libraries and Autonomous Citation Indexing, IEEE Computer, pp. 67-71, June 1999

[18] Lawrence S., Bollacker K., Giles C., Indexing and Retrieval of Scientific Literature, Proceedings of CIKM 1999 Conference, pp. 139-146

[19] Research Index Computer Science Directory http://citeseer.nj.nec.com/directory.html

[20] Stolt H., Agents, filter and search engines an evaluating survey on technologies for effective search for information from internet resources, Graduation thesis, Umea University, 1997

[21] Brin S., Page L., The Anatomy of a Large-Scale Hypertextual Web Search Engine, Stanford University, http://www-db.stanford.edu/pub/papers/google.pdf

[22] Glover E., Lawrence S., Birmingham W., Giles C., Architecture of a Metasearch Engine that Supports User Information needs, Proceedings of CIKM-99 Conference, pp. 210-216, ACM, 1999

[23] Lawrence S., Giles C., Context and page analysis for improved web search, IEEE Internet Computing, July 1998, pp.38-46
[24] Lawrence S., Giles C., Inquirus: The NECI Search Software, http://www.neci.nj.nec.com/homepages/lawrence/inquirus.html
[25] Seberg E., Etzioni O., The MetaCrawler Architecture for Resource Aggregation on the Web, http://www.cs.washington.edu/research/metacrawler, 1998
[26] Lawrence S., Giles C., Searching the Web: General and Scientific Information Access, IEEE Communications Magazine, January 1999, pp 116-122
[27] Franklin S., Graesser A., Is It an Agent, or just a Program?: A Taxonomy for Autonomous Agents, Proceedings of the Third International Workshop on Agent Theories, Architectures, and Languages, Springer-Verlag, 1996
[28] Etzioni O., Weld D., Intelligent Agents on the Internet: Fact, Fiction, and Forecast, 1995
[29] Gilbert, Aparicio, The Role of Intelligent Agents in the Information Infrastructure, IBM 1995
[30] Wooldrige M., Jennings N., Agent Theories, Architectures, and Languages: A Survey, Intelligent Agents pp 1-22, Springer-Verlag, 1995
[31] Stolt H., Agents, filter and search engines an evaluating survey on technologies for effective search for information from internet resources, Umea University, 1997
[32] Chen L., Sycara K., WebMate: A Personal Agent for Browsing and Searching, in Proceedings of Conference on Autonomous Agents, ACM, 1998, Pages 132-139
[33] T. Joachims, D. Freitag, T. Mitchell, WebWatcher: A Tour Guide for the World Wide Web, Proceedings of IJCAI97 Conference, August 1997, http://www.cs.cmu.edu/~webwatcher/ijkai97.ps
[34] Mladenic D., Personal WebWatcher: Implementation and Design Technical Report IJS-DP-7472, Department of Intelligent Systems, J.Stefan Institute, Slovenia, 1996, http://www.cs.cmu.edu/afs/cs/project/theo-4/text-learning/www/pww/papers/PWW/pwwTR.ps.Z.
[35] Bollacker K., Lawrence S., Giles C., CiteSeer: An Autonomous {Web} Agent for Automatic Retrieval and Identification of Interesting Publications, Proceedings of the Second International Conference on Autonomous Agents, ACM Press, 1998, pp. 116--123
[36] Aglets Workbench, http://www.trl.ibm.co.jp/aglets/information.html
[37] Finin T., Fritzson R., KQML – A Language and Protocol for Knowledge and Information Exchange, UMBC Univercity
[38] JKQML, http://alphaworks.ibm.com/tech
[39] Jennings N., Sycara K., Woodridge M., A Roadmap of Agent Reasearch and Development, Autonomous Agents and Multy-Agent Systems, 1, pp275-306, 1998
[40] Voorhees E., Agent Collaboration as a Resource Discovery Technique
[41] Oates T., Prasad N., Lesser V., Networked Information Retrieval as Distributed Problem Solving. University of Massachusetts
[42] Open Archives Initiative http://www.openarchives.org
[43] W3C Metadata Activity http://www.w3c.org/RDF/
[44] Ахо А., Ульман Д., Теория синтаксического анализа, перевода и компиляции, М., Мир, 1978.
[45] Metadata Activity Statement, http://www.w3c.org/Metadata/Activity.html