# Применение методов искусственного интеллекта в задаче поиска информации в Интернет

Одинцов И.О., Холчева А.В.
Санкт–Петербургский университет, математико–механический факультет,
кафедра информатики
Санкт-Петербург, Россия
E-mail: oio@sparc.spb.su

## 1 Резюме

Выполнен анализ взаимного проникновения результатов исследований в областях искусственного интеллекта и Интернет. Рассмотрены профессиональные методы и средства ведения поиска в Интернет. Предложена система, относящаяся к классу сетевых браузеров нового поколения, включающая возможность создания онтологическо-гипертекстовой модели предметной области.

## 2 Введение

Тематика исследований в двух крупных направлениях информатики – области искусственного интеллекта и Интернет – сближается. Методы искусственного интеллекта все в большей степени ориентируются на задачи практического применения, а Интернет стремится к более сложным приложениям, требующим интеллектуального поведения.

Одним из наиболее мощных информационных ресурсов Интернет является среда World Wide Web (WWW) – глобальная интерактивная распределенная гипертекстовая информационная система. Можно предположить, что в Интернет можно найти информацию практически по любой тематике. Однако, сделать это достаточно сложно, так как наиболее распространенные и известные поисковые инструменты – справочники и поисковые сервера – не позволяют эффективно структурировать результаты поиска. Кроме того, возникает задача отсеивания информации, то есть отсечения ненужной и несвязной информации от той ее части, которая будет полезной. Существует необходимость в совершенствовании программного обеспечения, которое поможет пользователю в интеллектуальном поиске и отборе нужной информации. Именно Интернет становится той средой, в которой с успехом может применяться накопленный опыт решения задач искусственного интеллекта.

Первая Всероссийская научная конференция
ЭЛЕКТРОННЫЕ БИБЛИОТЕКИ:
ПЕРСПЕКТИВНЫЕ МЕТОДЫ И ТЕХНОЛОГИИ,
ЭЛЕКТРОННЫЕ КОЛЛЕКЦИИ
19 - 21 октября 1999 г., Санкт-Петербург

## 3 Профессиональный поиск информации

Результаты профессионального поиска информации играют огромную роль в методах научного познания. Заметим, что в научном исследовании важно не только знание, но и граница между знанием и незнанием.

Этапы, соответствующие началу научного исследования могут быть алгорифмизированы следующим образом:

- Формулировка проблемы, постановка цели (что хотим достичь) и конкретных задач (что необходимо сделать для достижения цели)
- Выбор методов исследования и построение стратегии информационно-аналитического поиска.

Напомним, что общие методы научного познания обычно делят на следующие три большие группы [6]:

1. Методы эмпирического исследования (наблюдение, сравнение и др.)
2. Методы, используемые как на эмпирическом, так и на теоретическом уровне исследования (абстрагирование, анализ и синтез и др.)
3. Методы теоретического исследования (восхождение от абстрактного к конкретному и др.)

Как мы видим, практически все методы требуют поиск информации и использование результатов поиска в качестве входных данных.

Два основных этапа могут быть выделены в случае простейшей (и наиболее распространенной) стратегии поиска информации:

1. запрос информации с целью исследования некоторой предметной области;
2. выборка действительно необходимой информации и генерация отчета по исследуемой проблеме.

Эти этапы и должны быть автоматизированы поисковыми системами. Для повышения результативности поисковых систем предлагается применение методов инновационного поиска [4] с включением необычных, радикальных нововведений, коренным образом влияющих на понимание и решение задачи поиска. Поиск решения может включать эвристические правила (например, четырехступенчатый подход, предложенный Д.Пойа для решения новых задач или алгорифм решения изобретательских задач Г.С. Альтшуллера).

## 4 Краткая история поисковых средств Интернет

Активное внедрение средств компьютерных телекоммуникаций, глобальных компьютерных сетей и сетевых баз данных сделало сеть Интернет доступной и необходимой для многих. Интернет обладает двумя основными свойствами:

- предоставляет информационные ресурсы;
- имеет средства работы с этими ресурсами.

Необходимость создания поисковых средств Интернет была осознана практически с момента создания сети. Исторически известны такие приложения, как Apachie (для ftp), Veronica (для Gopher) и WAIS (поиск в индексированных базах данных). Поисковые сервера – наиболее развитый и удобный в использовании инструмент – можно определить как выделенные компьютеры, просматривающие все ресурсы Интернет и индексирующие их содержимое.

Важную роль с точки зрения определения местонахождения ресурса в Интернет является унифицированный указатель ресурсов, включающий название протокола и унифицированный идентификатор ресурса.

Основной проблемой, возникающей при работе с информацией в Интернет, является недостаточная структурированность информации (наличие так называемых полуструктурированных данных). Для описания такой информации было предложено несколько моделей, в том числе Stanford's Object Exchange Model [9], в рамках которой данные представляются в виде направленного графа с поименованными вершинами и дугами.

Другой проблемой является наличие противоречивых и недостоверных сведений. Типичным примером является включение некорректных ключевых слов в группу инструкций META, специально предназначенных для описания и индексирования документов поисковыми машинами.

Заметим, что среди информационных ресурсов Интернет особый интерес вызывают базы данных, которые подразделяются [5] на:

- текстовые (полнотекстовые, реферативные, библиографические);
- базы данных, содержащие изображения и использующие средства мультимедиа;
- числовые и табличные;
- базы данных, содержащие программное обеспечение.

Базы данных в Интернет принято разделять на рассчитанные на массового и на профессионального потребителя. Баз данных первого типа – большинство. Ко второй группе относятся профессиональные информационные системы, представляющие собой специализированные базы данных и поисковые программы. Например, крупнейший мировой продавец информации – компания Questel–Orbit, разместила в Интернет базу данных патентов [10].

## 5 Искусственный интеллект и Интернет

Термин "искусственный интеллект" (artificial intelligence) будем использовать как обозначение одной из областей информатики, направленной на моделирование и решение задач, связанных с обработкой символьной информации, логикой и естественным языком. Искусственный интеллект ставит перед собой и более серьезную задачу построения теории интеллекта, базирующуюся на обработке информации. Заметим, что многие методы и приемы искусственного интеллекта нашли применение в конкретных проблемных областях [7]. Перечислим лишь некоторые:

- обработка естественного языка;
- извлечение информации из баз данных;
- экспертные консультирующие системы;
- доказательство теорем;
- роботика;
- автоматическое программирование;
- комбинаторные задачи;
- проблемы зрительного восприятия.

В настоящее время исследование применимости идей искусственного интеллекта к среде Интернет ведется в следующих направлениях [1], [2]:

1. Обработка запросов на естественном языке поисковыми машинами. Система получает запрос на естественном языке, с помощью грамматических и лингвистических правил сопоставляет запрос с информацией в Интернет. Удачным примером такой системы является Яndex [11]. В ней независимо от того, в какой форме употреблено слово в запросе, поиск учитывает все его формы по правилам русского языка.
2. Продукционно-эвристическое распознавание естественного языка. Системы, развивающиеся в этом направлении, как правило, узкопрофильны и предназначены для выполнения очень узко сформированных целей или обработки ограниченных областей данных. Пример системы – FAQFinder [12].
3. Автоматическое наполнение базы знаний. Это и следующее направления используют агентную технологию. Интеллектуальный агент – программа, которая должна уметь реагировать на окружение, генерировать цели без вмешательства пользователя, взаимодействовать с другими агентами и, конечно, действовать рационально для достижения цели. Это направление реализовано в интеллектуальном браузере WebWatcher [13].
4. Использование эвристических правил для установки приоритета. Это направление реализовано в другом интеллектуальном браузере – Letizia [14].
5. Направление нейронных сетей. Основой здесь являются сети искусственных нейронов и других аналогичных конструкций, присущих нервной системе человека, в которых протекают психические процессы, опирающиеся на физиологический и биохимический уровни. Пример реализации этого направления – система работы со знаниями – Autonomy [15].

6. Принцип разделяемых знаний реализуется в виртуальных группах, то есть организациях и рабочих группах людей, которые работают и общаются между собой в интерактивном режиме. Знакомство с такого рода системами можно начать с AIMS [16].

## 6 Система поиска информации в Интернет нового поколения

Практическую реализацию многие идеи искусственного интеллекта нашли в системе поиска информации в Интернет нового поколения.

Подчеркнем три ее основные особенности, отличающие ее от рассмотренных в нашем обзоре:

1. Построение системы на парадигме "усиления информации". Интернет – является информационной подпиткой для поступающего в систему первоначального запроса и дальнейшего развития двухслойной структуры информации.

2. Двухслойная структура представления информации, состоящая из онтологической модели представления знаний и гипертекстовой модели. Это параллельное (и в то же время тесно связанное) представление основано на современных представлениях о метриках левой и правой моделей мира [3], основанных на специфике процессов в левом и правом полушариях головного мозга. Онтологическая модель в системе реализует правополушарное мышление, а гипертекстовая – левополушарное.

3. Методология "анализ–синтез", реализованная в подсистеме вывода.

Отметим четыре следующие особенности системы, обуславливающие ее дружественность пользователю:

- переносимость, определяемая реализацией системы на языке программирования Java;
- быстродействие, определяемое параллельным использованием множества мощных поисковых серверов Интернет;
- расширяемость, определяемая наличием объектно–ориентированной библиотеки быстрых приложений–апплетов ("ежиков") для выполнения небольших конкретных задач;
- наличие мощного графического редактора, позволяющего редактировать двухслойную модель представления информации.

Основная "подсистема вывода" имеет структуру классического компилятора, решающего три основные задачи – анализ, работа с промежуточным представлением и синтез:

1. Рассеянный поиск в информационных ресурсах Интернет на основе анализа ключевой фразы запроса.

2. Переработка собранной информации во внутреннее гипертекстовое представление и ее семантическая обработка.

3. Генерация результирующего документа с традиционными главами и подразделами на основе гипертекстовой модели. Преобразование структуры "предметной области" в структуру "изложения".

Рассмотрим подходы к решению этих задач (реализуемые "ежиками") более подробно:

На основе анализа ключевой фразы запроса выполняется поиск в информационных ресурсах Интернет. Заметим, что ключевая фраза преобразуется на основе метазнаний в совокупность ключевых слов, соединенных логическими операторами. Для системы разработан язык запросов допускающий булевские связки, скобочные структуры большой вложенности. Пользовательский запрос приводится к стандартному виду, используемому на том поисковом сервере, куда обращается система. На данном этапе с помощью метаправил определяется категория запроса. Система поочередно обходит поисковые сервера (Яndex [11], Alta–Vista, Yahoo и др.) и получает списки URL–адресов, соответствующих пользовательскому запросу. Рассматриваем несколько первых ссылок с каждого сервера и сортируем их по следующему принципу: в первую очередь рассматриваются ссылки, встречающиеся на большем количестве поисковых серверов, среди них отдаем предпочтение тем серверам, которые специализируются по данной тематике.

Далее, при помощи скриптов обрабатываются исходные документы, которые, как правило, уже имеют гипертекстовую структуру. В любом случае, применяя технологию построения гипертекста [8] получаем гипертекстовую модель объектов предметной области и их отношений.

Основной единицей гипертекстовой структуры мы будем считать "логический абзац" - логически связный кусок текста с рядом характеристик (атрибутов). Логический абзац может воплощать метафорную схему – некоторую очевидную абстрактную идею, с которой связываются шаблоны поведения. Атрибуты логического абзаца можно подразделить на 2 части:

1. собственные атрибуты (взвешенное дерево пользовательского запроса, ссылки на релевантные абзацы);

2. атрибуты, наследуемые от документа (ссылка на исходный документ, количество серверов, на которых появился этот документ, категория документа, количество релевантных абзацев в документе)

Приведем пример правила, используемого для при построении гипертекста:

- ЕСЛИ на некоторый логический абзац есть ссылки из совершенно разных контекстов и в его описании важно это указать
- ТО в данном логическом абзаце следует организовать обратные ссылки

Таким образом, на вход генератора итогового документа подается набор логических абзацев со списками атрибутов. Синтезировать отчет можно в нескольких формах, различающихся порядком следования логических абзацев, с предоставлением пользователю переключаться из одного режима в другой. Сортировка гипертекстовых абзацев может производиться по следующим принципам:

- по релевантности логических абзацев;
- по релевантности документов на серверах;
- по рассматриваемым поисковым серверам;
- по частоте появления документа на серверах.

## 7 Результаты

Основным результатом данного исследования является практическое подтверждение тезиса о возможности и необходимости использования подходов и методов искусственного интеллекта в разработке приложений для Интернет и особенно в задачах поиска и обработки данных в Интернет. Благотворно и обратное влияние, заключающееся в возможности использования ресурсов (данных и знаний) Интернет интеллектуальными (в том числе экспертными) системами. Интернет способствует возрождению интереса к тематике искусственного интеллекта.

## Библиография


[1] Daniel E.O'Leary. *The Internet, Intranets, and the AI Renaissance.*//Computer, January 1997.

[2] Kuhanandha Mahalingam, Michael N.Huhns. *A Tool for Organizing Web Information.*//Computer, June 1997.

[3] Грановская Р.М., Березная И.Я. *Интуиция и искусственный интеллект.* –Л: Издательство ЛГУ, 1991.

[4] Косовский Н.К., Хитров Д.В. *Аггрегация инноваций для повышения результативности систем.*//Первый российский философский конгресс. Том V, стр. 122–124. Санкт–Петербург, 1997.

[5] Краснослободцев В.Я., Смирнов А.Б., Лиходедов Н.П. *Инновационный инжиниринг. Практикум. Учебное пособие.* –СПб: Государственный технический университет, 1998.

[6] Кузин Ф. *Кандидатская диссертация: методика написания, правила оформления, порядок защиты.* – М.:"Ось-89", 1998.

[7] Нильсон Н. *Принципы искусственного интеллекта.* –М.:Радио и связь, 1985.

[8] Ованесбеков Л.Г. *Технология построения гипертекстов. Диссертация на соискание ученой степени кандидата физико–математических наук.* – М.,1993. Специальность 05.13.11.

[9] *Working Group on Semistructured Data* [http://www–rocq.inria.fr/ simeon/semistructure/art.html]

[10] *Welcome to QPAT–US* [http://www.qpat.com]

[11] *Яndex* [http://www.yandex.ru]

[12] *InfoLab: FAQ Finder* [http://www.infolab.nwu.edu/faqfinder/]

[13] *WebWatcher Home Page* [http://www.cs.cmu.edu/afs/cs.cmu.edu/project-/theo–6/web–agent/www/project–home.html]

[14] *Letizia Home Page* [http://lcs.www.media.mit.edu/people/lieber-/Lieberary/Letizia/Letizia–Intro.html]

[15] *Autonomy – Knowledge Management and New Media Content Solutions* [http://www.autonomy.com/]

[16] *About AIMS* [http://aims.parl.com/About–AIMS.html]