УДК 681.518:004.93.1

ПРОЕКТИРОВАНИЕ ПОРТАЛА ЗНАНИЙ ПО ИНФОРМАЦИОННЫМ ТЕХНОЛОГИЯМ

М.С. Бабий, А.П. Чекалов, С.С. Шевченко
Сумский государственный университет, г. Сумы

Предложен проект портала знаний по информационным технологиям и искусственному интеллекту. Рассмотрена схема индексирования и поиска данных с возможностями уточнения запроса и ранжирования результатов поиска.

ВВЕДЕНИЕ И ПОСТАНОВКА ЗАДАЧИ

В настоящее время в сети Интернет представлен большой объем знаний по информационным технологиям и искусственному интеллекту. Однако ресурсы Интернет недостаточно систематизированы, т.е. практически случайным образом распределены на сайтах различной тематической направленности, что значительно затрудняет их поиск и использование.

Проблема осложняется еще и тем, что различные группы людей используют при работе с поисковыми системами как свои специальные термины, так и термины, используемые другими группами в ином контексте. Вследствие этого возникает проблема несовместимости используемых терминов.

Исходя из необходимости организации эффективного поиска, хранения, классификации, анализа и обработки все возрастающего объема знаний в области информатики, поставлена задача спроектировать *портал знаний* по искусственному интеллекту и информационным технологиям и разработать схему индексирования и поиска документов с возможностями уточнения запроса и ранжирования результатов. На основании анализа тенденций развития корпораций и научных сообществ [1] можно предполагать значительное увеличение в будущем роли и количества подобных порталов.

СТРУКТУРА ПОРТАЛА

Традиционно портал представляет собой сайт или совокупность сайтов, предоставляющих пользователю широкий набор услуг. В их число входят информационный сервис, сервис реализации бизнес-функций, сервис для общения, инструментарий для продвижения собственного контента, прежде всего бесплатный хостинг и e-mail.

В отличие от классического корпоративный портал знаний ориентирован на более узкий тип информационного наполнения. Современные порталы знаний позволяют проводить многофакторные исследования, выполнять глубинный анализ текста, выявлять направления развития.

Логическую основу проектируемого портала знаний составляет онтология. Каждая база знаний или система, основанная на знаниях, разрабатывается в рамках некоторой концептуализации, а онтология чаще всего определяется как *точная спецификация концептуализации* [2]. Независимо от вида основу онтологии составляет словарь представленных в ней терминов, дополненный описанием отдельных терминов для их правильного толкования. Особенностью онтологии порталов знаний, ориентированных на поиск информации в Интернете, является наличие в них описания сетевых ресурсов наряду с традиционным описанием предметной области.

Общая схема онтологии разрабатываемого портала представлена на рис.1.

Наиболее важной частью портала является поисковая система. Поисковая система состоит из трех основных частей.

Робот – подсистема, которая обеспечивает сканирование Интернет и является основным средством сбора информации о состоянии информационных ресурсов сети по искусственному интеллекту и информационным технологиям. Подсистема создает и периодически обновляет индексную базу данных по данной тематике.

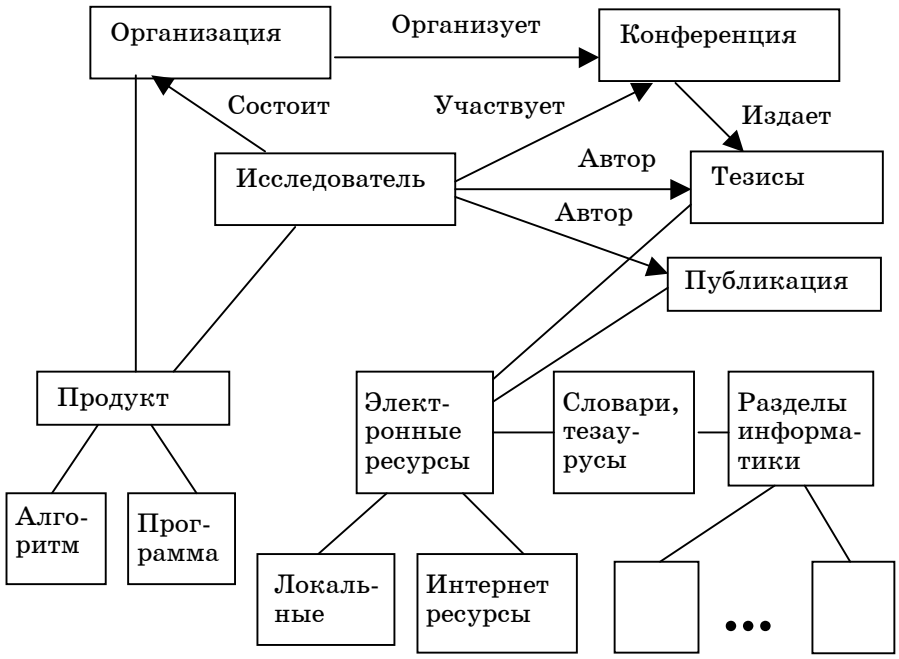

Рисунок 1 – Онтология портала

Индексная база данных – структура данных, включающая, прежде всего, инвертированный файл, состоящий из лексических единиц из проиндексированных Web-документов. База содержит информацию о местонахождении этих единиц в документах, а также о самих документах.

Подсистема поиска – подсистема, обеспечивающая обработку запроса пользователя, поиск в базе данных и выдачу результатов поиска пользователю. Поисковая система общается с пользователем через пользовательские интерфейсы, предоставляющие экранные формы для ввода запроса и вывода результата.

Индексная база представляет собой набор связанных между собой файлов, ориентированных на быстрый поиск данных по запросу. Основной файл базы строится по инвертированной схеме, при которой доступ к документам обеспечивается через их идентификаторы содержания. Каждая запись файла идентифицирована соответствующим идентификатором содержания (идентификатор термина, имени автора, названия организации и т.п.) и содержит адреса всех документов, в которых он содержится.

Так как для пользователя портала важно видеть связь поискового термина из запроса с содержанием текста, то для каждого идентификатора содержания в инвертированном массиве вместе с адресом документа хранятся контексты поискового термина.

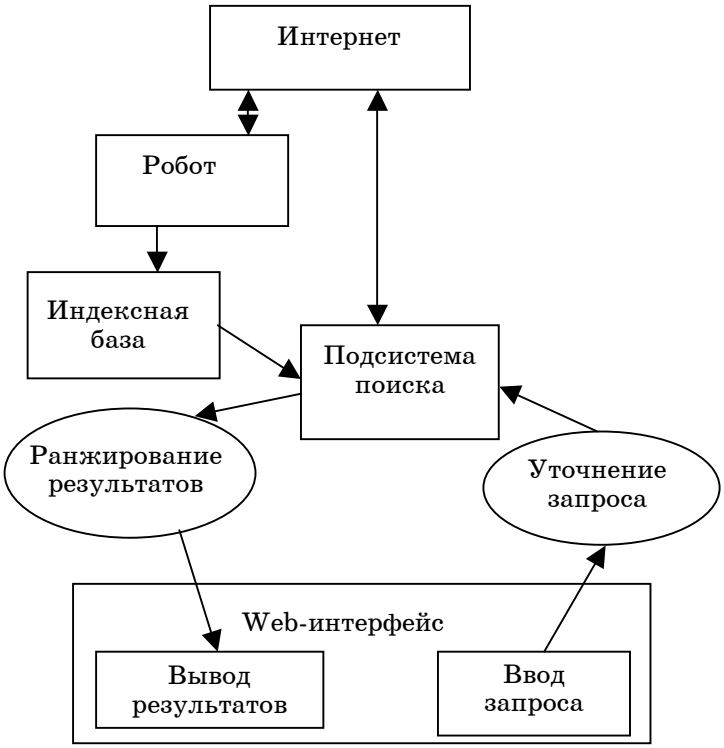

Рисунок 2 – Общая схема работы портала

МОДЕЛЬ ПОИСКА ДАННЫХ

Для представления массива документов в базе данных используется векторно-пространственная модель [3]. В рамках этой модели i-й документ описывается вектором

$$d_i=\{w_{i1}, \dots , w_{in}\},$$

где w_{ij} представляет собой вес j-го слова в данном документе. Весовой коэффициент w_{ij} будем вычислять по формуле

$$w_{ij} = a_{ij} \log \frac{N}{n_j},$$

где a_{ij} – частота появления j-го слова в i-м документе, n_j – количество документов, в которых встречается j-е слово, N – общее количество документов в коллекции. Запрос q будем представлять вектором

$$q_i=\{q_1, \dots, q_n\},$$

где q_i=1, если j-е слово присутствует в запросе q, и q_i=0 в противном случае.

Тематическую близость $r(d_i,q)$ запроса к документу будем оценивать как нормированное скалярное произведение

$$r(d_i, q) = \frac{(d_i, q)}{\| d_i \| \bullet \| q \|}.$$

Рассматриваемая модель обеспечивает простую реализацию поиска независимо от длины запроса и возможность ранжирования результатов поиска на основе близости документа к запросу.

Однако следует отметить тот факт, что терминология в области информационных технологий пока не является установившейся. Одни и те же объекты разные разработчики называют по-разному, часто используются англоязычные синонимы. В то же время термины часто бывают многозначными, например: сеть (компьютерная, нейронная, мобильная, информационная), кластер, интерфейс, порт. Следовательно для организации эффективного поиска нужно автоматически распознавать смысловые оттенки слов в зависимости от контекста их использования. В этом случае запрос может быть уточнен терминами из найденного контекста.

Задача нахождения латентных тем часто решается методами факторного анализа [4], при этом из пространства элементарных факторов методом сингулярного разложения матриц выделяются главные факторы. Более гибким представляется вероятностное латентно-семантическое индексирование PLSI [5, 6]. Применительно к рассматриваемой ситуации с каждой парой (d_i, w_j) может быть сопоставлен латентный класс z_k. Вероятностное моделирование позволяет оценить вероятности $P(z| d, w)$, $P(w| z)$, $P(d| z)$. В соответствии с принципом максимального правдоподобия упомянутые вероятности определяются путем максимизации функции правдоподобия

$$L = \sum_{d\in D}\sum_{w\in W} n(d,w)\log P(d,w),$$

где $n(d,w)$ – число вхождений слова w в документ d.

Максимизация функции L выполняется методом *ожидания-максимизации* (EM) [7]. На каждой итерации выполняются шаг *ожидания*

$$P(z \mid d,w) = \frac{P(z)P(d \mid z)P(w \mid z)}{\sum_{z'} P(z')P(d \mid z')P(w \mid z')}$$

и шаг *максимизации*

$$P(w \mid z) = \frac{\sum_d n(d,w)P(z \mid d,w)}{\sum_{d,w'} n(d,w')P(z \mid d,w')},$$

$$P(d \mid z) = \frac{\sum_w n(d,w)P(z \mid d,w)}{\sum_{d',w} n(d',w)P(z \mid d',w)},$$

$$P(z) = \frac{1}{R}\sum_{d,w} n(d,w)P(z \mid d,w), \quad R = \sum_{d,w} n(d,w).$$

Уточнение запроса основывается на выявлении тематической принадлежности документов, отмеченных пользователем как релевантные на первом этапе поиска. На втором этапе запрос расширяется словами из отмеченных документов.

Процесс расширения выполняется следующим образом. Пусть пользователь выделил множество документов S. Для слов w из документов множества S вычисляется их вес

$$Q(w) = \sum_{d \in S, z \in Z} P(z)P(d \mid z)P(w \mid z).$$

После этого множество слов w упорядочивается по убыванию весов и из построенного списка выбираются первые несколько слов.

Процесс ранжирования результатов, который основывается на выявлении близости документа к запросу, модернизируется следующим образом. Вместо эмпирической вероятности появления слова в документе $P'(w \mid d)$ используется линейная комбинация ее же и вероятности, полученной методом PLSI

$$(1-\lambda)P'(w \mid d) + \lambda\sum_z P(w \mid z)P(z \mid d),$$

где $(0 < \lambda < 1)$. Интуитивно понятно, что хотя слово-синоним из запроса может не присутствовать в документе, тем не менее условная вероятность его появления в документе, полученная методом PLSI, может быть отлична от нуля.

ВЫВОДЫ

1 Разработан проект портала по информационным технологиям и искусственному интеллекту с возможностями обработки ресурсов Интернета.

2 Предложена схема индексирования и поиска документов на основании пространственно-векторной модели и вероятностного латентно- семантического подхода.

Внедрение портала знаний позволит сконцентрировать информацию о последних достижениях в области информационных технологий и выполнять эффективный поиск необходимой информации

SUMMARY

DESIGNING OF THE KNOWLEDGE PORTAL ON INFORMATION TECHNOLOGY

Babiy M.S., Chekalov O.P.,Shevchenko S.S.

There is offered project of the portal of the knowledges on information technology and artificial intelligence. It is considered scheme of the indexing and searching for data with possibility of elaborating of the request and ranking result searching for.

СПИСОК ЛИТЕРАТУРЫ

1. Grammer. The Enterprise Knowledge Portal // DM Review Magazine. – March 2000.
2. Загорулько Ю.А. и др. Подход к построению предметной онтологии для портала знаний по компьютерной лингвистике // Компьютерная лингвистика и интеллектуальные технологии: Труды международной конференции «Диалог 2006». – Москва: Изд-во РГГУ, 2006. -С. 148-151.
3. Salton, Gerard. Introduction to Modern Information Retrieval. – McGraw-Hill, 1983.
4. Deerwester S., Dumais S.T., Furnas G.W., Landauer T.K., Harshman R. Indexing by latent semantic analysis // Journal of the American Society for Information Science, 41(6), 1990. – 3.391-407.
5. Hofmann T. and Puzicha J. Unsupervised learning from dyadic data // Technical report, UC, Berkeley, Berkeley, CA, 1998.
6. Hofmann T. Probabilistic latent semantic indexing // In. Proceedings of SIGIR'99, 1999.
7. Dempster A.P., Laird N.M., and Rubin D.B. Maximum likelihood from incomplete data via the EM algorithm // Journal of the Royal Statistical Society. Series B. – 39(1):13, 1977.

Бабий М.С., канд. техн. наук;
Чекалов А.П.,канд. техн. наук;
Шевченко С.С., соискатель

Поступила в редакцию 26 марта 2008 г.