СПИСОК ЦИТИРОВАННЫХ ИСТОЧНИКОВ

1. Найденов В.И., Швейкина В.И. Нелинейные модели колебаний речного стока // Водные ресурсы. Том 29, № 1. – М., 2002. – С. 62 – 67.
2. Пантелеев А.В., Якимова А.С., Босов А.В. Обыкновенные дифференциальные уравнения в примерах и задачах: Учеб. пособие. – М.: Высш. шк., 2001. – 376 с.
3. Прокопеня А.Н., Чичурин А.В. Применение системы Mathematica к решению обыкновенных дифференциальных уравнений: Учеб. пособие. – Мн.: БГУ, 1999. – 265 с.
4. Сванидзе Г.Г. Математическое моделирование гидрологических рядов. – Л.: Гидрометеоиздат, 1977. – 296 с.

Статья поступила в редакцию 10.10.2005

УДК 004.8.032.6

Крапивин Ю.Б., Алькатауна Х.А., Страчук И.В.

ИСПОЛЬЗОВАНИЕ НЕЙРОННЫХ СЕТЕЙ ДЛЯ ПОИСКА ИНФОРМАЦИИ В СЕТИ INTERNET

Введение

Наиболее быстро развивающейся технологией является Интернет, которая объединяет множество разных сетей, миллионы компьютеров, около 300 миллионов пользователей со всех континентов, число которых по разным оценкам, увеличивается на 15-80% ежегодно.

Основная часть информации в Интернете - это неструктурированное хранилище информации огромного объёма, которая характеризуется высокой динамичностью. В связи с чем, обеспечение поиска в Интернете становится критически важной задачей, которая не разрешима без соответствующих поисковых средств.

В статье рассматриваются современные поисковые системы в Интернете и основные технологии их построения, статистические закономерности естественно-языковых текстов, модели индексирования и поиска документов, а также использование нейронных сетей в поисковых системах.

Методы информационного поиска

Субъективно понимаемая цель идеального информационного поиска - найти все пертинентные и только пертинентные документы (найти «только то, что надо, и ничего больше»), при этом запрос должен максимально точно выражать информационную потребность, а документ должен быть максимально релевантным.

Информационный поиск производится при помощи систем информационного поиска как специальных комплексов программного, информационного и технического обеспечения.

Выполнение основных функций систем информационного поиска, обеспечивается её различными структурными элементами: информационно-поисковым языком, поисковой машиной, инструментами метапоиска, тематическими рубрикаторами.

Как показывает проведенный анализ существующих поисковых инструментов в Интернете, все они имеют свои достоинства и недостатки.

По виду выдаваемой информации системы информационного поиска делят на документальные и фактографические. К документальным системам относятся поисковые машины, средства метапоиска, рубрикаторы.

Поисковая машина представляет собой, с одной стороны Web-сервер, главная страница которого обеспечивает пользователю возможность формирования запроса. С другой стороны, она обеспечивает создание и ведение каталога Web-страниц, который позволяет выбрать адреса нужных страниц по данным, содержащимся в запросе.

Схема, поясняющая организацию работы типичной поисковой машины, представлена на рисунке 1.

Каждая из основных универсальных поисковых машин покрывает ограниченное Web-пространство Интернет. По различным оценкам, покрытие не превышает 30-40% доступных Web-страниц. При этом языковые возможности для записи поискового выражения также ограничены. Они не выходят за пределы ключевых слов и фраз, связанных операторами Буля (AND, OR, NOT) [1,2].

Метапоисковые средства позволяют расширить область поиска практически на всё Web-пространство Интернет, используя одновременно 6-12 основных универсальных поисковых машин. Однако выразительные средства языка формирования поискового выражения остаются теми же [3].

Наиболее совершенными поисковыми инструментами на сегодняшний день являются поисковые утилиты, так как они позволяют получить результаты поиска непосредственно на компьютер пользователя и самому пользователю выполнять их дополнительный анализ в режиме off-line. При этом возможно применение более мощного языка формирования поискового выражения [4].

Еще одним средством поиска информации в сети Интернет являются иерархические классификаторы (директории). Они обеспечивают рубрикацию по заданным тематикам как целых сайтов, так и других электронных ресурсов.

Классификацию текстов на естественном языке называют рубрицированием [5].

В настоящее время практическое применение получили следующие группы классификаторов:

- статистические классификаторы, на основе вероятностных методов.
- классификаторы, основанные на функциях подобия.
- классификаторы, использующие методы на основе искусственных нейронных сетей.

Классификаторы, использующие методы на основе искусственных нейронных сетей хорошо зарекомендовали себя в задачах распознавания изображений, и в данной статье рассматривается возможность их использования в обработке текстов на естественном языке.

Общие принципы функционирования системы

Предлагаемую поисковую систему можно логически разделить на следующие четыре части:

1. Интерфейс пользователя;
2. Индексирующий агент;
3. Модули системы;
4. База данных.

Укрупнённое взаимодействие этих частей показано на рисунке 2.

В основе функционирования описываемой системы лежит нейросетевой аппарат, обеспечивающий поиск и рубрикацию проиндексированных документов. Описания классов документов, представляют собой векторы действительных чисел, заложенные в синаптических весах искусственных нейронов, а

Крапивин Ю.Б., *магистрант кафедры интеллектуальных информационных технологий БрГТУ.*
Алькатауна Х.А., *аспирант БГУИР.*
Страчук Игорь Васильевич, *ассистент кафедры технической эксплуатации автомобилей БрГТУ.*
Беларусь, Брестский государственный технический университет, 224017, г. Брест, ул. Московская 267.

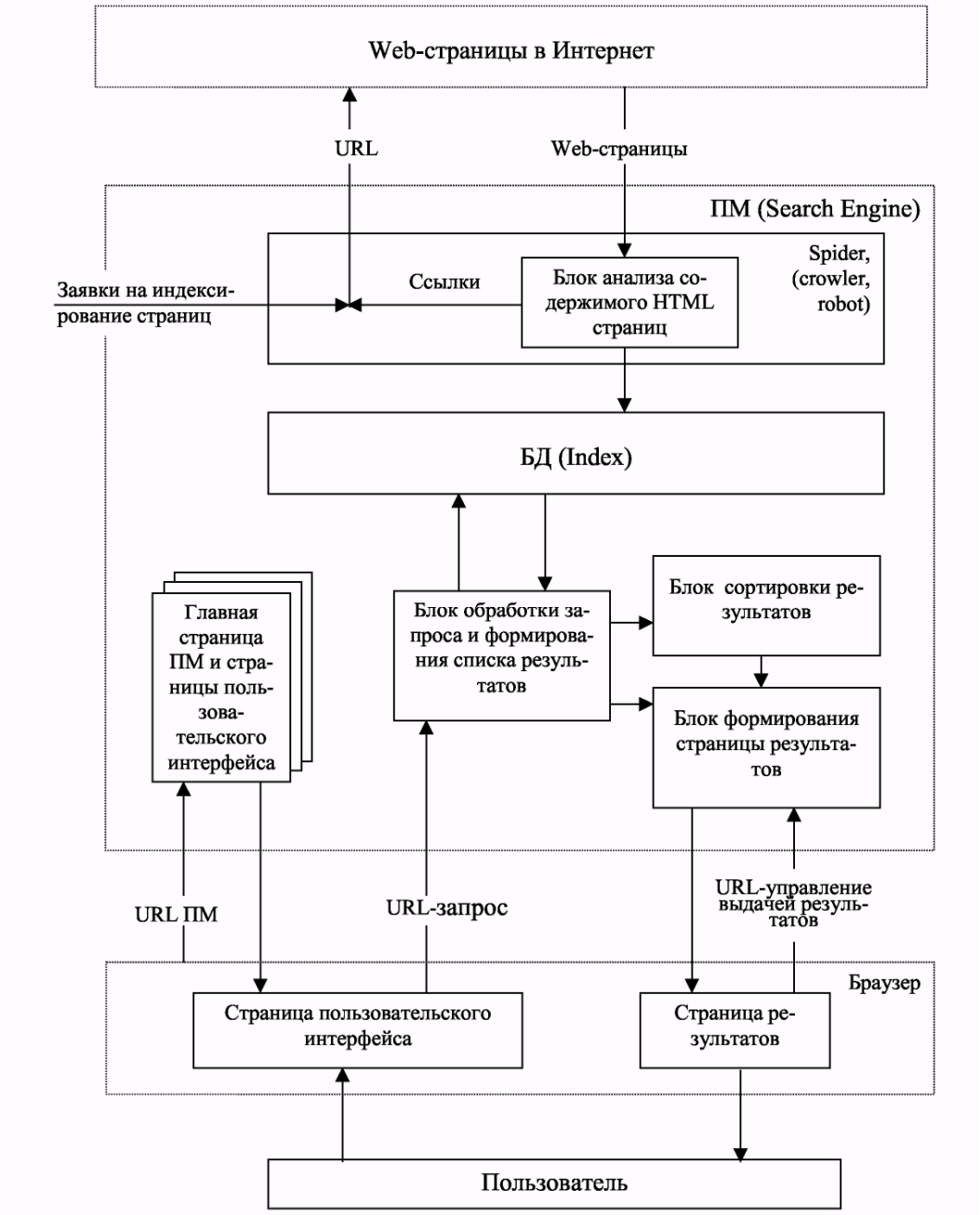

Рис. 1. *Схема функционирования поисковой машины*

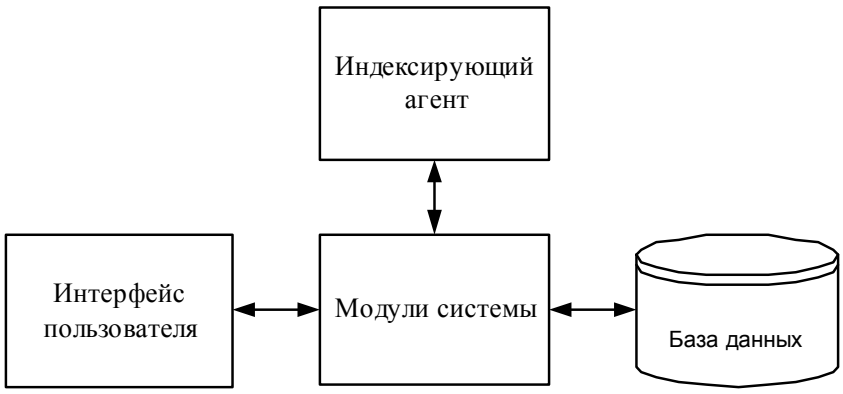

Рис. 2. *Общая структура системы*

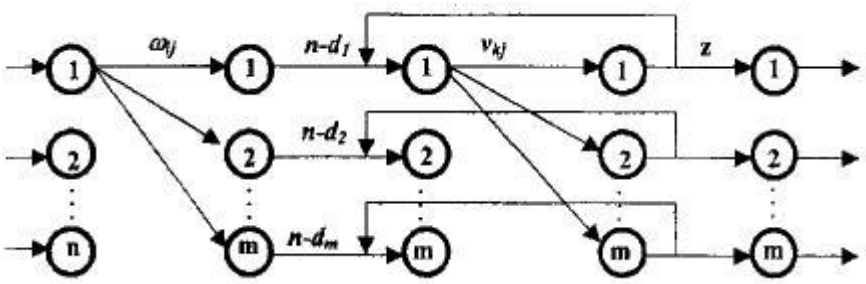

Рис. 3. *Архитектура сети Хэмминга*

процедура классификации характеризуется способом преобразования анализируемого текста к аналогичному вектору, видом функции активации нейронов, а также топологией сети. Процесс обучения системы, в данном случае, совпадает с процедурой обучения сети и зависит от выбранной топологии.

В системе реализована нейронная сеть Хэмминга, архитектура которой представлена на рисунке 3 [6].

Т.к. нейронные сети приспособлены обрабатывать только информацию, представленную числовыми векторами, то для их применения в обработке текстов на естественном языке, последние необходимо представлять в векторном виде [7]. Для этого наилучшим образом подходит модель «терм-документ». В модели «терм-документ», используемой в системе, текст описывается лексическим вектором:

$$\{\tau_i\}, i = 1...N_W ,$$

где τ_i – важность (информативный вес) термина w_i в документе;

N_w – полное количество терминов в документальной базе (словаре).

Вес термина, отсутствующего в документе, принимается равным 0. Для удобства веса нормируются, так что $\tau_i \in [0,1]$. В работе использовались дискретные значения, так что присутствующий термин в тексте имеет вес 1, а отсутствующий – вес 0.

В описываемой системе нейронная сеть используются в качестве классификаторов двоичных векторов: нейронная сеть документов – для поиска документов по ключевым словам и поиска похожих документов; нейронная сеть рубрик – для автоматического определения рубрик документов.

В процессе функционирования нейронная сеть проходит две фазы: обучение (задание классов); классификация (определение класса, к которому относится вектор). Классы нейронной сети документов представляют собой лексические векторы проиндексированных документов сайта, а классы нейронной сети рубрик – лексические векторы рубрик сайта.

Предлагаемая поисковая система функционирует следующим образом.

Определяется словарь ключевых слов рубрик. Далее для каждой рубрики определяются наборы ключевых слов из словаря, выражающих ее смысловое содержание и однозначно описывающих рубрику. Затем, в соответствии с моделью «терм-документ», эти наборы представляются в виде лексических векторов. Рубрицируемые документы описываются с помощью набора ключевых слов на этапе индексирования автоматически, и представляются в виде лексических векторов. Сформированные векторы рубрик нейронная сеть рубрик запоминает на этапе обучения. Определение рубрики документа производится сетью на этапе классификации. Число нейронов выходного слоя сети определяется числом хранимых ей рубрик. Выходное значение нейронной сети определяется номером нейрона с максимальным значением выхода. Такой подход дает возможность получать на выходах нейронной сети (по номеру активизировавшегося нейрона выходного слоя) номер рубрики в соответствующих индексных таблицах базы данных рубрик и индексированных документов, к которой относится поданный на нейронную сеть лексический вектор рубрицируемого документа.

Поиск документов происходит подобным образом. Запрос, принятый с помощью интерфейса пользователя, разделяется на отдельные слова и представляется в виде лексического вектора. Затем, вектор подается на нейронную сеть, определяющую номер наиболее релевантного документа в соответствующих индексных таблицах базы данных рубрик и индексированных документов. Далее вступают в действие традиционные алгоритмы по выборке и выводу конечному пользователю результатов поиска.

На этапе индексирования индексирующий агент осуществляет обработку и пополнение базы данных документами, расположенными по указанному адресу URL. В основе метода выделения ключевых слов документа на этом этапе лежат статистические закономерности формирования текстов на естественном языке, определенные Дж. Зипфом и оформленные законами «ранг-частота», «количество-частота», а так же алгоритм индексирования Дж. Солтона [8].

В качестве средства хранения базы данных была выбрана система управления базой данных Microsoft SQL Server 7.0. Такой выбор был сделан по ряду причин: высокая производительность, обеспечение секретности, удобство проектирования сложных баз данных, а так же доступность данного продукта.

Интерфейс пользователя, реализованный в виде скриптов с помощью технологии JSP, предоставляет авторизированный доступ как пользователю, так и администратору системы.

Заключение

Таким образом, в статье проведен сравнительный анализ функционирования поисковых машин, средств метапоиска, тематических рубрикаторов и методов их построения, рассмотрена система информационного поиска в Интернете, использующая теорию нейронных сетей, что подтверждает основную мысль статьи о возможности применения нейронных сетей для организации поиска информации во всемирном Web-пространстве.

СПИСОК ЦИТИРОВАННЫХ ИСТОЧНИКОВ

1. Voorhees Ellen M. Overview of the TREC-9 question answering track. - Тр. 9-й международной конференции по проблемам поиска информации (TREC-9), 2000.
2. John H. Hine, A Survey of Information Discovery and Retrieval Tools // Department of Computer Science Victoria University of Wellington, 1996 – http://www.csvuw.com
3. Lan Huang, A Survey on Web Information Retrieval Technologies // Computer Science Department State University of New York, 2000 - http://www.dsuny.com
4. Robert J. Kuhns, A Survey of Information Retrieval Vendors // Sun Microsystems Laboratories 2550 Garcia Avenue Mountain View, CA 94043 - http://www.sun.com

5. Гаврилова Т.А., Червинская К.Р. Извлечение и структурирование знаний для экспертных систем. – М.: Радио и связь, 1992.
6. Головко В.А. Нейроинтеллект: Теория и применения. Книга 1. Организация и обучение нейронных сетей с прямыми и обратными связями - Брест:БПИ, 1999, - 260с.
7. Ермаков А.Е., Плешко В.В. Синтаксический разбор в системах статистического анализа текста. // Информационные технологии. – 2002. – N7.
8. Солтон Дж. Динамические библиотечно-информационные системы. – М.: Мир, 1979 – 230с.

Статья поступила в редакцию 06.03.2007

УДК 801.73:681.3

Антонов С.Г.

ИЕРАРХИЧЕСКАЯ МОДЕЛЬ ТЕКСТА ДЛЯ ЗАДАЧИ КОРРЕКЦИИ ОШИБОК

Введение

Традиционно основными показателями эффективности систем автоматической переработки текста являются показатели полноты и точности. Поскольку такая переработка в общем случае является многостадийной, т.е. предполагает прохождение всех уровней глубины языка, то, учитывая его природу, итоговые значения указанных показателей в реальных промышленных системах, как правило, невелики. Анализ показал, что одной из наиболее существенных причин этого является наличие различного рода ошибок в самом входном тексте, т.е. нарушение его лингвистической адекватности.

Процедура устранения нарушений лингвистической адекватности текста в общепринятом понимании включает: поиск ошибок, построение вариантов коррекции и устранение неоднозначности. Наиболее сложным этапом является устранение неоднозначности вариантов коррекции, поскольку их построение, как и поиск ошибок, являются в большей степени техническими задачами.

Устранение неоднозначности вариантов коррекции, очевидно должно базироваться на формализованных лингвистических знаниях, выраженных посредством модели текста и ее содержания. В основу такой модели, как оказалось, целесообразно положить естественную иерархичность графического и лингвистического представления текстов любых алфавитных естественных языков: буква (графический уровень) – морфема – словоформа - сегмент предложения – предложение [1].

Общая схема построения иерархической модели

Рассмотрим общую схему построения иерархической модели текста до синтаксического уровня включительно.

При построении модели сначала выделяются элементы текста каждого уровня иерархии и принципы перехода с уровня на уровень. В результате определяется тип модели, например, n-граммная, морфологическая, синтаксическая и т.п., и далее строится непосредственно сама модель [2].

Основной составной частью модели является система простых элементов модели или элементов первого уровня. Обозначим их множеством $Z^{(1)} = \left\{ z_i^{(1)} \right\}_{i=\overline{1,r_1}}$. Это элементы, не имеющие деления на более мелкие в пределах данного уровня. Выбор системы простых элементов определяет основу всей модели, являясь при этом связующим звеном между моделью и реальными объектами окружающей действительности.

Другой составной частью являются правила построения элементов следующего уровня $Z^{(2)} = \left\{ z_i^{(2)} \right\}_{i=\overline{1,r_2}}$ на языке простых элементов, заданные или списочно, или посредством некоторого другого описания. Дальнейшее наращивание модели происходит по такой же схеме.

Поскольку система простых элементов текста может быть выделена на любом уровне его иерархического представления, уровнем модели будем считать уровень иерархии текста, на котором базируется система простых элементов. Например, модель на уровне букв означает, что простыми элементами модели являются буквы, из которых строятся другие элементы текста, а в модели на уровне слов, соответственно, текст строится из словоформ.

Наибольшую сложность представляет формализация закономерностей построения элементов текста каждого уровня. Если ввести в состав сложных элементов любого уровня выше первого так называемый пустой элемент $\varnothing$, т.е. элемент, не соответствующий никакому объекту текста, то тогда можно описать правило построения элементов h-го уровня из элементов (h-1)-го в виде некоторого отображения

$$\pi_{h-1} : \bigcup_j \left(Z^{(h-1)} \right)^j \to Z^{(h)},$$

где $\left(Z^{(h-1)} \right)^j$ есть j-я декартова степень (т.е. все последовательности длины j) множества элементов (h-1)-го уровня. Вообще говоря, j есть любое натуральное число. Однако текстам естественных языков всегда присущи ограничения на существующие длины последовательностей, т.е. $\exists J_{h-1} : \forall \bar{z} \in \left(Z^{(h-1)} \right)^j, j > J_{h-1}, \pi_{h-1}\left(\bar{z}\right) = \varnothing$. Тогда нам достаточно рассматривать ограничение π_{h-1} на $\bigcup_{j=1}^{J_{h-1}} \left(Z^{(h-1)} \right)^j$. Указанное ограничение будем называть правилами построения элементов следующего уровня

$$\pi_{h-1}^* : \bigcup_{j=1}^{J_{h-1}} \left(Z^{(h-1)} \right)^j \to Z^{(h)}.$$

Если модель имеет p уровней, то тогда должны быть описаны правила $\pi_h^*, h = \overline{1, p-1}$. Набор $\pi^* = \left(\pi_h^* \right)_{h=\overline{1,p-1}}$ можно назвать набором правил модели, а суперпозицию $\pi' = \pi_{p-1}\left(\pi_{p-2}\left(\ldots \left(\pi_1 \right) \ldots \right) \right)$ правилами построения $Z^{(p)}$ из $Z^{(1)}$.

Очевидно, что для построения модели необходимо описать элементы всех уровней и правила их построения, основанные на неких признаках этих элементов.

Множество элементов модели h-го уровня должно содержать все объекты формализованного текста, которые могут встретиться на соответствующем уровне. Эти объекты можно назвать «максимальными объектами», соответствующими данному уровню модели. Иерархическое представление уровней модели предопределяет, что максимальные объекты надо

***Антонов Сергей Георгиевич**, к.т.н., старший научный сотрудник, научный консультант ООО «Крос 2000», Россия, г. Москва.*