На основании анализа ряда публикаций (например, [6]), в которых и ранее обращалось внимание на неэффективность критерия χ^2 Пирсона с группированием данных наблюдений в интервалы равной длины, можно выделить следующие основные причины такой неэффективности:

1.Потеря информации при группировании данных наблюдений, которая, при разбиении на интервалы равной длины, является абсолютно непредсказуемой (неуправляемой). «Хорошее» разбиение, по-видимому, сильно зависит от распределения случайной величины.

2.Большинство результатов математической статистики имеют асимптотический характер, а на практике всегда имеют дело с ограниченными выборками.

3.Классические предположения математической статистики опираются, как правило, на нормальный закон распределения (например, ошибок наблюдений). Это не всегда справедливо в приложениях, а теоретическое обоснование подобных ситуаций связано с аналитически трудноразрешимыми задачами и потому отсутствует.

Таким образом, разбиение области определения случайной величины на интервалы равной длины применимо лишь для построения гистограмм или вычисления оценок параметров распределения методом минимизации расстояния χ^2, для проверки же гипотез с помощью критерия Пирсона таковое разбиение следует признать неэффективным.

ЛИТЕРАТУРА

1. Бочаров П.П., Печинкин А.В. Теория вероятностей. Математическая статистика. – М.: Гардарика, 1998. – 328 с.

2. Sturgess H.A. The choice of classic intervals // J. Am. Statist. Assoc. – March 1926. – P. 47.

3. Шторм Р. Теория вероятностей. Математическая статистика. Статистический контроль качества. – М.: Мир, 1970. – 368 с.

4. Heinhold I., Gaede K.W. Ingeniur statistic. – München; Wien, Springer Verlag, 1964. – 352 s.

5. Ртищева М.В., Разумейчик В.С., Дереченник С.С. Анализ топологических характеристик неплотных неупорядоченных монодисперсных структур / Проблемы проектирования и производства радиоэлектронных средств: Сб. материалов IV Международной НТК (25-26 мая 2006, г. Новополоцк). – Т.2. – Новополоцк: ПГУ, 2006. – С. 214-217

6. Лемешко Ю.Б., Чимитова Е.В. Об ошибках и неверных действиях, совершаемых при использовании критериев согласия типа χ^2 //Измерительная техника. – 2002. – №6. – С. 5-11.

УДК 693.22.004.18

Крапивин Ю.Б.

Научный руководитель: д.т.н. проф. Головко В.А.

НЕЙРОННЫЕ СЕТИ И ИХ ПРИМЕНЕНИЕ ДЛЯ ПОИСКА ИНФОРМАЦИИ В СЕТИ INTERNET

ВВЕДЕНИЕ

Наиболее быстро развивающейся технологией является Интернет, которая объединяет миллионы компьютеров, около сотни миллионов пользователей, число которых увеличивается на 50% 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 системы информационного поиска в сети Интернет, основные технологии их построения, статистические закономерности существующие в текстах естественного языка, 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.

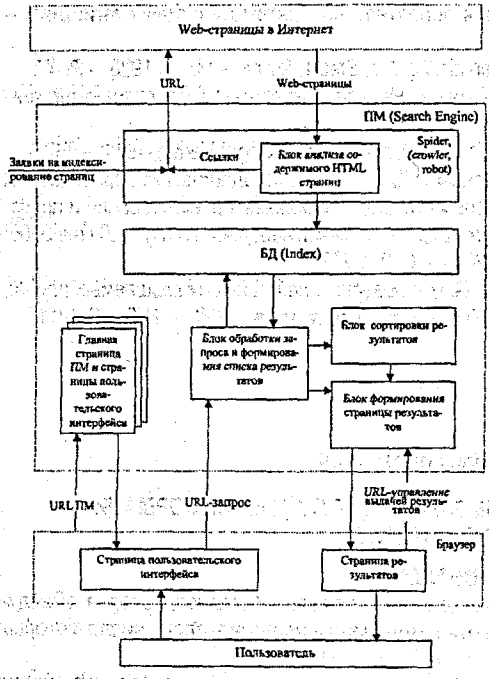

Рисунок 1 - Схема функционирования поисковой машины

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].

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 [2].

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.

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.

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

Система поиска и автоматической классификации документов в сети Интернет на основе машинного обучения реализована в виде программы на языке Java с использованием реляционной СУБД Microsoft SQL Server 7.0.

Для проведения экспериментов коллекция документов предварительно анализируется программой-индексатором. Каждому документу, расположенному по указанному адресу URL, присваивается уникальный идентификатор, осуществляется приведение слов к нормальной форме и выделение наиболее значимых слов, являющихся ключевыми для данного документа, в соответствии со статистическими закономерностями в текстах на естественном языке, определенных Дж. Зипфом, и алгоритмом индексирования Дж. Сол-тона [3]. При этом происходит пополнение документами базы данных под управлением СУБД.

Экспертом определяются наборы ключевых слов из словаря, выражающих ее смысловое содержание и однозначно описывающих рубрику. Затем, в соответствии с моделью «терм-документ» [3], эти наборы представляются в виде лексических векторов.

В описываемой системе функции автоматической классификации и поиска осуществляет нейронная сеть Хэмминга, обеспечивающая классификацию и поиск информации, представленной в виде лексических векторов запроса, рубрик и документов, и относящаяся к классу *релаксационных* [4, 5, 6] нейронных сетей. 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 ней-

ронную сеть лексический вектор рубрицируемого документа. Поиск документов происходит подобным образом. Принятый запрос 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.

РЕЗУЛЬТАТЫ ЭКСПЕРИМЕНТОВ

Оценка качества поиска предлагаемой системы осуществлялась с помощью метрик, наиболее часто использующихся для оценки качества работы систем информационного поиска и являвшихся официальными метриками РОМИП'2004 [7]: полнота (recall), точность (precision), аккуратность (accuracy), ошибка (error), F-мера (F-measure). В качестве тестового набора документов использовалась 20 Newsgroups data set [8] - коллекция документов, предоставленных в свободный доступ для проведения исследований в области автоматической обработки текстов. Для экспериментов случайным образом были выбраны 157 документов, которые необходимо было автоматически отнести к одной из 9 рубрик.

При решении задачи поиска документов, релевантных запросам пользователя, средняя точность соответствия результата запросу составила 37,6%, при этом значение полноты поиска - 56,6%. Среднее значение F-меры - единой метрики, объединяющей метрики полноты и точности, составило 0,44039. Нейронная сеть, применяемая в задаче поиска, генерирует наряду с точным решением (документом, содержащим все слова запроса), если такое можно найти, неточные решения – документы, содержащие одно или несколько слов. То есть она реализует в себе принцип работы поисковой машины, позволяющей формировать запрос на основании ключевых слов, связанных булевыми операторами И и ИЛИ одновременно.

Результаты работы системы для решения задачи классификации документов по заданным рубрикам представлены в таблице 1. Средняя точность правильного рубрицирования документов составила 83,8%, при этом значение полноты рубрицирования - 54,8%. Среднее значение F-меры для решения задачи классификации составило 0,662572.

Таблица 1 – Результаты работы системы для решения задачи классификации документов

Рубрики	Точность	Полнота	Аккуратность	Ошибка	F-мера
alt.atheism;	1,000	0,125	0,95541401	0,044586	0,222222
computers	0,875	0,583	0,96178344	0,038217	0,7
comp.os.ms-windows.misc	0,857	0,429	0,94267516	0,057325	0,571429
comp.sys.ibm.pc.hardware	0,652	0,714	0,91082803	0,089172	0,681818
misc.forsale	0,857	0,400	0,93630573	0,063694	0,545455
talk.politics.guns	1,000	1,000	1	0	1
talk.politics.mideast	1,000	0,833	0,99363057	0,006369	0,909091
rec.autos	0,875	0,700	0,97452229	0,025478	0,777778
rec.motorcycles	1,000	0,167	0,93630573	0,063694	0,285714
science	0,250	0,500	0,97452229	0,025478	0,333333
sci.crypt	0,875	0,700	0,97452229	0,025478	0,777778
sci.electronics	0,333	0,091	0,92356688	0,076433	0,142857
sci.med	1,000	0,429	0,97452229	0,025478	0,6
sci.space	1,000	0,545	0,96815287	0,031847	0,705882
rec.sports	1,000	1,000	1	0	1

ЗАКЛЮЧЕНИЕ

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

ЛИТЕРАТУРА

1. Электронный ресурс: http://www.searchenginewatch.com/webmaster/work.html.
2. Электронный ресурс: http://www.searchenginewatch.com/links/Metacrawlers.html.
3. Солтон Дж. Динамические библиотечно-информационные системы. – М.: Мир, 1979 – 230с.
4. Головко В.А. Нейроинтеллект: Теория и применения. Книга 1. Организация и обучение нейронных сетей с прямыми и обратными связями - Брест:БПИ, 1999, - 260с.
5. Головко В.А. Нейроинтеллект: Теория и применения. Книга 2. Самоорганизация, отказоустойчивость и применение нейронных сетей - Брест:БПИ, 1999, - 228с.
6. Горбань А.Н., Дунин-Барковский В.Л., Кирдин А.Н. и др. Нейроинформатика. Новосибирск: Наука. Сибирское предприятие РАН, 1998. – 296 с.
7. Кураленок И., Некрестьянов И., Оценка систем текстового поиска. // Программирование. 28(4), 2002, 226-242.
8. Электронный ресурс: http://people.csail.mit.edu/jrennie/20Newsgroups.

УДК 004.896

Калюхович Д.О.

Научный руководитель: проф., доктор техн. наук Головко В.А.

УПРАВЛЕНИЕ ДВИЖЕНИЕМ РОБОТА В ЗАДАЧЕ СЛЕДОВАНИЯ ЗА ЛИНИЕЙ

Целью данной работы является разработка алгоритма управления движением робота в задаче следования за линией. Разработанный алгоритм может использоваться в качестве одной из подсистем распределенной системы управления транспортировкой грузов как между различными участками одного производства, так и между различными организационными структурами предприятия, что позволяет повысить эффективность и, как следствие, рентабельность производства. В настоящее время для слежения за направлением движения используются дорогостоящие инфракрасные датчики и CCD-камеры, что ограничивает область применения таких систем. Ввиду этого разработка алгоритма, использующего для слежения за линией относительно дешевой веб-камеры, является востребованной и актуальной задачей.

Во время реализации данного алгоритма были выполнены следующие этапы:

1. Разработка и программная реализация алгоритма детектирования линии.
2. Создание новой команды для мобильного робота MAX (рис. 1), позволяющей во время движения поворачивать его передние колеса на определенный угол.
3. Разработка и программная реализация алгоритма управления движением робота в задаче следования за линией.
4. Тестирование.