На правах рукописи

ШАБАНОВ Владислав Игоревич

МОДЕЛИ И МЕТОДЫ АВТОМАТИЧЕСКОЙ КЛАССИФИКАЦИИ ТЕКСТОВЫХ ДОКУМЕНТОВ

Специальность 05.13.11 - Математическое и программное обеспечение вычислительных машин, комплексов и компьютерных сетей.

АВТОРЕФЕРАТ
диссертации на соискание ученой степени кандидата технических наук

Москва – 2003 г.

Работа выполнена в Московском Государственном Техническом Университете им. Н.Э. Баумана.

Научный руководитель:	доктор технических наук, профессор В.В. Сюзев
Официальные оппоненты:	доктор физико-математических наук, профессор Г.С. Осипов
	кандидат технических наук, доцент В.Б. Тарасов
Ведущее предприятие:	ОАО Всероссийский научно-исследовательский институт автоматизации и управления в непромышленной сфере (ОАО ВНИИНС)

Защита диссертации состоится "__" ____ 2004 года на заседании диссертационного совета Д 212.141.10 в Московском Государственном Техническом Университете им. Н.Э. Баумана по адресу 107005, г. Москва, 2-я Бауманская ул. д. 5.

С Диссертацией можно ознакомиться в библиотеке МГТУ им. Н. Э. Баумана.

Отзыв на автореферат в одном экземпляре, заверенный печатью, просим направлять в адрес совета университета.

Автореферат разослан "__" ____ 2004 г.

Ученый секретарь диссертационного совета к.т.н., доцент С.Р. Иванов

ОБЩАЯ ХАРАКТЕРИСТИКА РАБОТЫ

Актуальность проблемы

Актуальность проблемы автоматического анализа текстов на естественном языке связана с бурным развитием современных информационных систем и ростом объемов хранимой в них информации. Для работы с такими объемами данных нужны эффективные способы поиска и тематической классификации.

Классифицирующие системы могут решать задачи фильтрации и маршрутизации почтовых сообщений, уточнения результатов поиска в информационно-поисковых системах или оптимизации рекламных кампаний в сети Интернет.

В диссертационной работе рассматривается метод классификации, пригодный для обработки больших объемов текстовых документов, а также метод автоматического обучения классификатора, который обеспечивает высокую эффективность классификации и допускает интуитивно понятный способ ручной корректировки и настройки системы. Также в работе рассматриваются два метода повышения эффективности классификации: учет ассоциативных связей между терминами и распознавание в текстах фрагментов с легко формализуемым синтаксисом (дат, номеров телефонов и т. д.).

Цели и основные задачи работы

Целью работы является разработка метода тематической классификации текстовых документов с автоматическим обучением на заданных пользователем образцах, отличающихся более высокой эффективностью, по сравнению с известными методами, и при этом способного работать с большими объемами текстов. Для достижения поставленной цели в работе поставлены следующие задачи:

- Выбор способа оценки эффективности классификации. Анализ жизненного цикла системы классификации. Сравнительный анализ известных методов автоматической классификации текстов.
- Разработка метода классификации и соответствующего метода обучения классификатора, использующего в своей работе алгоритмы полнотекстового поиска и ранжирования. Разработка моделей представления документов, которые учитывают большее, по сравнению с традиционными методами, количество характеристик текста, и используют преимущества алгоритмов полнотекстового поиска при решении задачи классификации.
- Разработка метода повышения эффективности классификации, учитывающего ассоциативные связи между терминами. Разработка метода выявления ассоциативных связей, основанного на анализе потока запросов к ИПС.

- Разработка метода повышения эффективности классификации, основанного на выделении из текста метатекстовых конструкций (МТК), таких как даты, номера телефонов, фамилии и т. д. Разработка алгоритма выделения таких объектов.

Методы исследований

При решении поставленных задач в работе использован метод обучения классификатора, основанный на вычислении различительных сил, методы квазиреферирования, полнотекстового поиска и ранжирования, а также морфологического анализа.

Эксперименты проводились на реальных базах данных, поставляемых в составе коммерческой ИПС. В частности, исследовались базы нормативных актов Российского законодательства, коллекции сообщений, поступающих от служб новостей, а также аннотации к ресурсам сети Интернет, участвующих в рейтинге Rambler`s Top100. Для оценки качества работы предложенных методов было проведено сравнительное тестирование разработанного механизма классификации и контрольного механизма, использующего традиционный метод классификации, основанный на однословных терминах.

Научная новизна

В работе получены следующие научные результаты:

- Предложен и разработан оригинальный метод автоматической классификации, использующий алгоритмы полнотекстового поиска и ранжирования. Данный метод отличается более высокой точностью, по сравнению с традиционными методами, благодаря использованию словосочетаний в качестве классификационных признаков.
- Предложен и разработан метод обучения классификатора, базирующийся на методе различительных сил. Отличительной особенностью метода является использование во время обучения обратной связи от процедуры полнотекстового поиска, которое позволяет повысить эффективность классификации.
- Разработана модель представления документов в виде взвешенных векторов словосочетаний, в которой учитываются лексические, контекстные и статистические характеристики. В отличие от традиционных моделей, при расчете веса учитывается большее количество факторов, влияющих на точность классификации.
- Предложен метод повышения эффективности классификации, основанный на учете ассоциативных связей между терминами. Разработан оригинальный метод построения словарей ассоциативных связей при помощи анализа потока запросов к поисковой машине.
- Разработан метод повышения эффективности классификации, основанный на автоматическом выделении МТК. Шаблоны правил распознава-

ния МТК состоят из декларативной части, в которой задается синтаксис объектов, и процедурной части, задающей алгоритм проверки корректности и приведения объекта к канонической форме.

Достоверность научных положений и выводов диссертационной работы подтверждена практической реализацией и опытной эксплуатацией системы автоматической классификации, основанной на методах полнотекстового поиска и ранжирования.

Практическая ценность и реализация результатов

Практическая значимость работы определяется возможностью использования рассматриваемых в ней моделей и методов в системе автоматической классификации текстов, которая может применяться в составе системы документооборота или информационно-поисковой системы (ИПС). Разработанные методы автоматизируют процесс построения тематических рубрикаторов и каталогов, а также повышают качество поиска в ИПС.

Апробация работы

Содержание отдельных разделов и диссертации в целом было доложено:

- На заседаниях аттестационной комиссии при ежегодной аттестации аспирантов кафедры «Компьютерные системы и сети» МГТУ им. Н.Э. Баумана.
- На семинарах и заседаниях кафедры «Компьютерные системы и сети» МГТУ им. Н.Э. Баумана.
- На межвузовской юбилейной научно-технической конференции аспирантов и студентов «Современные информационные технологии» 15 ноября 2000 г.
- На межвузовской научно-технической конференции аспирантов и студентов «Современные информационные технологии» 22 ноября 2001 г.
- На международной конференции Диалог`2003 по компьютерной лингвистике и интеллектуальным технологиям 15 июня 2003 г.
- На российском семинаре по оценке методов информационного поиска РОМИП-2003, проходившем в рамках пятой всероссийской научной конференции по электронным библиотекам RCDL-2003, 31 октября 2003 г.

Публикации

Основные результаты работы опубликованы в 7 печатных работах.

Содержание работы

Диссертационная работа состоит из введения, шести глав, заключения, списка литературы и приложений. Общий объем диссертации 228 страниц, включая 25 рисунков, список литературы и приложения. Библиография содержит 98 наименований, из них 65 из иностранных источников.

Во введении обосновывается актуальность темы диссертации, рассматривается место систем автоматической классификации текстов в современных информационных технологиях, дается краткая характеристика таких систем и методов их взаимодействия с пользователем. Формулируется цель работы и ее связь с другими аспектами информационных систем и систем обработки текстов на естественном языке. Приводятся примеры применения системы классификации.

В главе Постановка задачи дается формальное описание задачи автоматической классификации текстовых документов, задачи обучения системы классификации, а также выбираются критерии, по которым оценивается качество работы системы классификации..

Автоматическая классификация множества документов $\mathbf{D}=\{d_i\}$ на множестве рубрик $\mathbf{Rub}=\{r_j\}$, организованных в иерархию $\mathbf{H}=\{<r_j,r_p> : r_j,r_p \in \mathbf{Rub}\}$, заключается в вычислении следующей функции:

$$Classify\ (\mathbf{Rub},\mathbf{H},\mathbf{Sem},\mathbf{D})\text{->}\{<d_i,r_j,f_{ij}> : d_i \in \mathbf{D},\ r_j \in \mathbf{Rub},\ 0 \quad f_{ij} \quad 1\},$$

где:

$\mathbf{Sem} = \{\ \mathbf{Sem}(r_j),\ r_j \in \mathbf{Rub}\ \}$ – множество классификационных признаков;

$\mathbf{Sem}\ (r_j)$ – классификационные признаки рубрики r_j;

f_{ij} – мера подобия документа d_i рубрике r_j.

Задача обучения классификатора на заданных пользователем образцах формулируется следующим образом:

Пусть мы имеем обучающую выборку $\mathbf{T_{teach}}=\{<d_i,r_j> : d_i \in \mathbf{D},\ r_j \in \mathbf{Rub}\}$, состоящую из документов d_i, которые пользователи-эксперты просмотрели и вручную приписали к соответствующим рубрикам r_j. Тогда обучение классификатора состоит в формировании классификационных признаков **Sem** на основе анализа обучающей выборки:

$Teach\ (\mathbf{Rub},\mathbf{H},\mathbf{T_{teach}})\text{->}\mathbf{Sem}$

Для оценки эффективности системы предлагается сравнивать ее работу с данными, получаемыми от человека-эксперта. Для численной оценки эффективности в работе применяются традиционные для задач обработки текстовой информации меры: полноту, точность и комбинированную меру F_β (меру Ван Ризбергена).

Полнота U классификации определяется как отношение количества правильно распознанных связей документ-рубрика к истинному количеству таких связей:

$$U = \frac{a}{a+c}$$

где:

a – количество правильно распознанных связей рубрика-документ;

c – количество связей, которые система не смогла обнаружить.

Точность V определяется как отношение количества правильно распознанных связей к общему количеству распознанных классификатором связей:

$$V = \frac{a}{a+b}$$

где:

b – количество неверно распознанных связей рубрика – документ.

На практике оказывается, что не составляет никакого труда построить систему, обладающую высокой точностью при низкой полноте или высокой полнотой при низкой точности. Например, полнота системы, в которой все документы приписываются ко всем рубрикам, равна единице. Поэтому для адекватной оценки необходимо использовать комбинированную меру, которая учитывала бы одновременно и полноту и точность. В качестве такой меры а работе используется мера F_β:

$$F_\beta(V,U) = \frac{(\beta^2+1)VU}{\beta^2 V + U}$$

где:

β – параметр, задающий приоритет точности над полнотой.

Глава Обзор методов распознавания образов и классификации состоит из трех разделов. В первом разделе приводится общее сравнение методов автоматической классификации и обучения классификатора по следующим параметрам:

- способ предъявления обучающей выборки;
- варианты описания классифицируемых объектов;
- виды процедур классификации.

Существуют два основных способа предъявления выборки: обучение с использованием единственной выборки, и итерационное обучение на последовательности выборок. Недостатком итерационного обучения является накопление ошибок и неустойчивость к погрешностям обучающей выборки. По этой причине для классификации текстов обычно используют первый способ.

Классифицируемые объекты могут быть представлены точками в N-мерном евклидовом пространстве, списками признаков, или структурными описаниями. Известны методы классификации, основанные на каждом из трех подходов, однако наиболее распространенным является представление документов в виде списков признаков.

Сопоставление классифицируемого документа с признаками рубрик может идти параллельно или последовательно. В связи с тем, что большинство тематических рубрикаторов имеют иерархическую структуру, при классификации обычно идет параллельная обработка рубрик одного уровня иерархии и последовательная – для рубрик разных уровней.

Во втором разделе рассматриваются традиционные методы классификации текстов:

- вероятностные методы, основанные на правиле Байеса;
- методы, основанные на поиске ближайших соседей (kNN);
- методы, основанные на кластерном анализе и построении центроид;
- нейросетевые методы, а также ассоциативные семантические сети.

Обучение классификатора на многоуровневых рубрикаторах может проводиться индивидуально для каждого уровня иерархии, или комплексно, на основе анализа статистических характеристик документов всего дерева рубрик.

В третьем разделе приводится анализ и сравнение методов, выявляются факторы, оказывающие наибольшее влияние на эффективность классификации. Большинство методов ориентировано на применение однословных терминов в качестве признаков, что ограничивает точность классификации. С другой стороны, при использовании комбинаций слов (n-грамм) или словосочетаний, отмечается снижение полноты классификации за счет многообразия формулировок терминов.

Глава Математическая модель автоматического классификатора текстовых документов содержит описание модели автоматического классификатора, основанной на полнотекстовом поиске, морфологическом, статистическом и контекстном анализе. Подробно рассматриваются алгоритмы классификации и обучения, а также архитектура программного комплекса, реализующего эти алгоритмы.

Глава состоит из шести разделов.

В первом разделе определяется структура классификационных признаков: виды терминов, процедуры их сравнения на эквивалентность и способы преобразования к канонической форме (нормализации). Рассматриваются следующие варианты:

- **словоформы** – слова в том виде, в котором они содержатся в тексте;
- **лексемы** – нормальные формы слов, полученные при помощи морфологического анализа. Если слово имеет омонимы, генерируется несколько нормальных форм;
- **псевдоосновы** – нормальные формы, полученные путем приближенного отсечения окончаний и суффиксов;
- **словосочетания** – комбинации слов, выделенные из текста при помощи синтаксического анализа, или на основании анализа статистики совместной встречаемости слов в тексте.

В диссертационной работе анализируются достоинства и недостатки, а также проводится экспериментальная оценка эффективности каждого из вариантов.

Второй раздел главы посвящен оценке значимости терминов - численной меры, характеризующей способность термина отражать смысл документа, которая учитывается при формировании классификационных признаков рубрик. Предлагается метод оценки значимости, комбинирующий морфологические, контекстные и статистические характеристики. Значимость W_i для термина t_i вычисляется следующим образом:

$$W_i = C\, W_{lex\,i}^{\,a} * W_{context\,i}^{\,b} * W_{stat\,i}^{\,c}$$

где:

$W_{lex\,i}$ – собственная (лексическая) значимость термина t_i, которая зависит только от его текста;

$W_{context\,i}$ – контекстная значимость термина t_i, зависящая от информативности фрагментов документа, в которые от входит;

$W_{stat\,i}$ – статистическая значимость термина t_i, зависящая от характеристик его распределения в анализируемом массиве документов;

a, b, c, C – настроечные параметры.

Лексическая значимость определяется как функция от частей речи, признаков отглагольности и числа омонимов каждого из слов, входящих в термин.

Контекстная значимость определяется методами автоматического квазиреферирования. Анализируемый текст разбивается на предложения, которые разделяются на *автосемантичные* (не связанные с другими предложениями) и *синсемантичные* (связанные с другими предложениями). Формальным признаком синсемантичности является наличие в предложении *коннектора* – фрагмента, указывающего на наличие связи с предыдущим текстом. Для различения коннекторов составляются специальные шаблоны, в которых задается текст коннектора и возможное его контекстное окружение.

Синтаксические связи между соседними предложениями также выявляются на основе *лексического повтора*. Экспериментальным путем установлено, что при повторении существительного связь между предложениями существует в 92% случаев. Для прилагательных и глаголов данный параметр соответственно равен 65% и 34%. Дополнительно при оценке значимости учитываются *индикаторы* – выражения, указывающие на повышенную информативность предложения, в которое они входят.

Контекстная значимость термина рассчитываются на основе следующих характеристик:

- Распределение по автосемантичным и различным типам синсемантичных предложений. Например, при прочих равных условиях термин, входящий в 5 автосемантичных предложений, имеет более высокий вес

по сравнению с термином, входящим в 5 синсемантичных предложений.

- Количество связей между предложениями, образованных за счет лексического повтора анализируемого термина (чем больше связей, тем выше вес).
- Размеры связных фрагментов текста, содержащие термин (чем больше средний размер, тем выше вес). Под связным фрагментом понимается несколько последовательных предложений текста, первое из которых автосемантичное, а остальные синсемантичные, причем связи между предложениями зафиксированы не только путем учета коннекторов, но и на уровне лексического повтора.
- Наличие в предложениях, содержащих термин, индикативного выражения, указывающего на его важность.
- Принадлежность термина к предложениям, входящим в различные структурные части текста. Так, появление термина в структурной части "выводы" более важно, чем в структурной части "введение".

Третья составляющая значимости – статистическая. В диссертации предлагается использовать для оценки статистической значимости термина его различительную силу, величину которой можно оценить по следующей формуле:

$$W_{stat\,i} = \frac{1}{\sum_{i=1}^{N} \overline{f_i^2}} \left[\frac{\sum_{i=1}^{N} \overline{f_i^2} \cdot \overline{f_i}^2}{\sum_{i=1}^{N} \overline{f_i}^2} - \overline{f_i}^2 \right]$$

где

$\overline{f_i}$ – среднее число появлений термина t_i в документах обучающей выборки;

$\overline{f_i^2}$ – средний квадрат числа появлений термина t_i;

N – число терминов.

В третьем разделе главы рассматривается математическая модель документов и рубрик, а также метод классификации, основанный на полнотекстовом поиске. Семантический образ рубрики состоит из взвешенного списка терминов (поисковых запросов), по которым может быть найден соответствующий рубрике документ:

$$\mathbf{Sem}(r) = <\theta(r), l(r)>$$

где:

$\theta(r) = \{<t_i, f_i, w_i>\}$, $i = 1 \ldots n$ - список терминов (поисковых запросов) рубрики r ($r \in \mathbf{Rub}$);

$l(r)$ – пороговое значение классифицирующей функции (см. ниже), комбинирующей поисковые ранги в общую меру подобия документа рубрике.

t_i – термин (поисковый запрос);

f_i – вес поискового запроса t_{ji} для рубрики r;

w_j – пороговое значение поискового веса (ранга), начиная с которого поиск по данному запросу начинает учитываться при классификации;

Процедура обучения классификатора заключается в обнаружении поисковых запросов, специфических для каждой из рубрик, и вычислении их веса.

Процедура классификации состоит в индексировании анализируемого документа d (построении для него инверсного индекса $\mathbf{I}^{-1}(d)$, см. ниже), и рекурсивном обходе дерева рубрик с выполнением на каждом уровне иерархии полнотекстового поиска по терминам, содержащимся во множествах $\mathbf{Sem}(r)$ всех рубрик данного уровня. Найденные в $\mathbf{I}^{-1}(d)$ документы ранжируются, на основании полученных рангов вычисляются итоговые меры подобия документов рубрикам.

Решение о соответствии или несоответствии документа рубрике принимается после сравнения итоговой меры с вычисленными во время обучения пороговыми значениями $l(r_i)$. Рекурсивный обход иерархии выполняется только для нескольких рубрик, мера подобия которым максимальна.

Преимущества метода:

1) **высокая производительность** – вычисление поисковых запросов на инверсном индексе можно делать очень быстро.
2) **возможность быстрой пакетной классификации** – при построении общего инверсного индекса для группы документов, становится возможной одновременная классификация всех проиндексированных документов.
3) **высокая точность** – пользуясь данным методом можно поместить в семантический образ рубрики слова, словосочетания и даже целые фразы, при наличии которых в тексте можно «почти наверняка» утверждать, что документ соответствует рубрике.
4) **полнота классификации** – поиск может найти словосочетание в тексте, даже если его формулировка изменена.
5) **возможность ручной настройки классификатора** – структура семантических образов рубрик проста для понимания, добавление или удаление терминов (поисковых запросов) приводит ко вполне предсказуемым последствиям.
6) **возможность учета гиперссылок** – полнотекстовый поиск легко можно модифицировать так, чтобы он учитывал гипертекстовую структуру документов.

Документ, подаваемый на вход процедуре классификации, представляется инвертированным индексом со следующей структурой:

$$\mathbf{\Gamma}^{-1}(d) = <\tau, \nu(\tau_i), \mathbf{c}(\tau_i), \mathbf{z}(\tau_i)>,$$

где

$\tau = \{\tau_i\}$, $i=1\ldots n$ – множество слов документа d;

n – количество слов в документе (размер словаря документа);

$v(\tau_i) \in \mathbf{R}$ – вес (значимость) слова τ_i в документе;

$\mathbf{c}(\tau_i) = \{c_{ij}\}, j=1\ldots n(\tau_i), c_{ij} \in \mathbf{N}$ – вектор координат (номеров позиций), в которых встречается слово τ_i.

$\mathbf{z}(\tau_i) = \{z_{ij}\}, j=1\ldots n(\tau_i), z_{ij} \in \mathbf{R}$ – вектор весов зон документа (название, заголовок и т. д.), в которых встречается слово τ_i.

$\mathbf{g}(\tau_i) = \{g_{ij}\}, j=1\ldots n(\tau_i), g_{ij} \in \mathbf{R}$ – вектор информации о грамматических формах слова τ_i в тексте документа.

Данное представление позволяет вычислять следующие виды поисковых запросов:

- поиск одного слова;
- поиск группы слов, находящихся в тексте на расстоянии не более l слов;
- то же самое, но с фиксацией порядка слов в тексте;
- буквальный поиск (поиск слов, записанных в точности так, как они перечислены в запросе);
- поиск альтернатив (поиск одного из заданных слов);
- любая комбинация из перечисленных выше видов запросов

При ранжировании результатов учитываются следующие факторы:

- расстояние между словами в тексте – чем ближе слова находятся друг от друга, тем выше вес. Для вычисления этой составляющей веса используются вектора $\mathbf{c}(\tau_i)$;
- вхождение слов в важные зоны документа (заголовки, названия разделов и т. д.). Для учета этого фактора используются вектора $\mathbf{z}(\tau_i)$;
- совпадение грамматической формы слова с той, которая указана в запросе – вес повышается, если в тексте слово находится в такой же или близкой форме (другой падеж того же существительного) и понижается, если слово в тексте отличается.

В диссертационной работе предложен метод классификации, основанный на полнотекстовом поиске всех запросов из семантического образа рубрики и вычислении мер подобия документов рубрикам. Учитываются следующие характеристики:

- поисковый ранг – документ соответствует рубрике, если запросы, по которым он был найден, имеют высокий поисковый ранг;
- доля запросов рубрики – документ скорее соответствует рубрике, если он найден по большому количеству запросов из ее семантического образа;
- доля запросов документа – количество запросов рубрики, по которым был найден документ, среди всех запросов, по которым он был найден, должно быть велико.

В четвертом разделе рассматривается метод обучения классификатора, основанного на полнотекстовом поиске. Метод основан на обнаружении терминов, специфичных для каждой из рубрик, вычислении мер их значимости и порогового значения поискового веса. Также в процессе обучения определяется пороговое значение итоговой меры подобия документа рубрике.

Документ обучающей выборки представлен в виде взвешенного множества терминов:

$$d = \{<t_k, \tau_k, w_k>\}, k = 1 \ldots n_{terms}$$

где

t_k – термин (нормализованное представление);

τ_k – текст (список слов) самой частой реализации термина;

w_k – значимость термина для документа d.

Процедура обучения обладает следующими характеристиками:

- Классификация документов, на которых выполнялось обучение, выполняется с точностью, близкой к 1.
- Баланс между точностью классификации и полнотой обеспечивается за счет формирования множеств терминов с n-кратной избыточностью и требования m-кратного покрытия классифицируемого документа терминами (m n).
- Выбор терминов в семантические образы рубрик происходит на основе анализа их различительной силы относительно рубрик.

Отличительной особенностью предлагаемого метода является учет обратной связи от оценки результатов автоматической классификации обучающей выборки. При этом выполняется полнотекстовый поиск и ранжирование всех выбранных терминов (поисковых запросов), вычисляются пороговые значения поискового веса каждого термина и пороговое значение итоговой меры подобия рубрика-документ.

Обратная связь позволяет удалить из рассмотрения термины, ошибочно попавшие в семантические образы рубрик, и таким образом повысить эффективность классификации.

В пятом разделе главы дается детальное описание алгоритма обучения классификатора, а в шестом – структура программного комплекса Классификатор 2.0, в котором реализованы предлагаемые в диссертационной работе модули и методы.

Глава Автоматическое выявление ассоциативных связей между словами и словосочетаниями содержит описание метода автоматического построения словарей ассоциативных связей между терминами, а также применение их в автоматическом классификаторе.

Для формирования словарей ассоциативных связей предлагается анализировать поток запросов, которые пользователи подают на вход ИПС. Метод построен на следующем предположении: запросы, поданные одним и тем же пользователем в течение короткого промежутка времени,

скорее всего имеют одну и ту же тематику. Поэтому если большое количество пользователей, дав запрос *x*, сразу после этого дает еще и запрос *y*, можно сделать вывод о том, что запросы *x* и *y* между собой связаны.

Экспериментальная оценка метода на потоке запросов, подаваемых поисковой машине Rambler показала, что указанный выше критерий не является достаточным, так как в списки ассоциаций большого количества попадают такие запросы, как *реферат*, *Москва*, *знакомства* и т. д. Для решения этой проблемы предлагается выполнять ранжирование и фильтрацию ассоциативных связей. Ранг ассоциативной связи вычисляется следующим образом:

$$rank(x, y_i, f_i) = S(\mathbf{assoc}(x), \mathbf{assoc}(y_i))^{\alpha} \cdot \left(\frac{f_i}{\max_{i=1\ldots n} f_i} \right)^{\beta} \cdot W(y_i) \cdot Z(x, y_i)$$

где:

- x – исходный запрос;
- y_i –запрос, ассоциативно связанный с запросом x;
- f_i - количество пользователей ИПС, которые подали запросы x и y_i в течение короткого промежутка времени;
- n – количество разных запросов, которые пользователи подавали вместе с запросом x (количестве элементов в статье ассоциативного словаря);
- $S(x, y_i)$ – мера похожести списков ассоциаций запроса x и списка ассоциаций запроса y_i (косинусообразный коэффициент корреляции);
- $W(y_i)$ – функция от числа слов в запросе y_i, дающая небольшой приоритет запросам из двух или трех слов и уменьшающая ранг длинных запросов;
- $Z(x, y_i)$ – функция, повышающая вес ассоциациям, текст которых целиком включает в себя запрос x;
- α, β – постоянные коэффициенты.

Мера похожести между списками ассоциаций **assoc**(x) и **assoc**(y) определяется следующим образом:

$$S(assoc(x), assoc(y)) = \frac{\sum_{i,j,\ a_i=b_j} f_i \cdot f_j}{\sqrt{\sum_i (f_i^2) \cdot \sum_j (f_j^2)}}$$

где

- a_i – запрос, ассоциативно связанный с запросом x, i = 1 … n;
- n – число ассоциаций запроса x;
- b_j – запрос, ассоциативно связанный с запросом y; j = 1 … m;
- m – число ассоциаций запроса y;

f_i – количество пользователей ИПС, которые одновременно подали запрос x и запрос a_i в течение короткого промежутка времени;

f_j – то же самое для y и b_j.

Благодаря учету степени похожести из списков ассоциаций отсеиваются запросы, мало похожие на исходный запрос. Это означает, что слова *реферат*, *Москва* и др. получают высокий вес только среди подобных запросов.

Во втором разделе главы рассматривается метод повышения эффективности классификации за счет расширения семантических образов рубрик ассоциативными терминами. В ходе расширения выполняется вычисление различительных сил добавляемых терминов (запросов), а затем производится отсев тех ассоциаций, которые уменьшают эффективность классификации. Отсев выполняется за счет обратной связи от оценки результатов автоматической классификации обучающей выборки.

Глава Автоматическое распознавание текстовых метаконструкций содержит описание разработанной в рамках диссертации системы распознавания текстовых конструкций, которая позволяет выявлять в тексте документа объекты с формализуемым синтаксисом (даты, номера телефонов, адреса, и т.д.).

Необходимость разработки системы распознавания обусловлена тем, что в семантические образы рубрик часто попадают фрагменты дат, куски номеров телефонов и т. д., что неблагоприятно сказывается на точности классификации. Полнота классификации снижается из-за того, что для каждого такого объекта существует несколько эквивалентных способов записи, которые с точки зрения процедуры выделения терминов являются разными.

Для повышения полноты и точности необходимо распознавать такие конструкции и обрабатывать их как единое целое, объединяя при этом все эквивалентные формы написания в одну. Механизм распознавания МТК предназначен для решения этой задачи.

Система распознавания управляется шаблонами, состоящими из двух частей. Первая часть описывает на специальном декларативном языке синтаксис распознаваемых конструкций. Вторая часть шаблона содержит программный код, выполняющий проверку корректности, нормализацию и оценку значимости.

Специальный транслятор разделяет шаблоны на декларативную и процедурную части, компилирует декларативные части во внутреннее представление, формирует исходный текст из всех процедурных частей и передает его соответствующему компилятору или интерпретатору. Данный подход обладает следующими достоинствами:

- **простота** – система использует уже существующие языки программирования в процедурных частях шаблонов, язык описания декларатив-

ных частей не содержит громоздких элементов, задающих детальные правила проверки и нормализации.

- **гибкость** – процедурная часть может содержать код любой сложности.

В первом разделе рассматривается структура системы распознавания, которая состоит из транслятора шаблонов, библиотеки времени исполнения, лингвистических модулей и самих шаблонов с описаниями.

Во втором разделе главы предложен алгоритм работы системы распознавания. Текст, поступающий на вход системе, разбивается на атомарные единицы – фрагменты. Затем из этих фрагментов организуется очередь, с элементами которой сопоставляются декларативные части шаблонов. В случае успешного распознавания все фрагменты передаются процедурной части соответствующего шаблона. Если процедурная часть подтвердила корректность распознавания, фрагменты удаляются из очереди, а на их место помещается специальный нетерминальный фрагмент.

В третьем разделе описывается синтаксис декларативной части языка описания шаблонов и приводятся примеры.

Глава Результаты экспериментов содержит результаты экспериментальной проверки эффективности предлагаемых в диссертационной работе методов, а также оценка вклада каждого из них в итоговые показатели.

В частности, получены данные, свидетельствующие о повышении точности классификации при использовании словосочетаний вместо однословных терминов. Полнотекстовый поиск позволяет избежать низких показателей полноты, характерных для традиционных методов классификации, использующих словосочетания.

Отмечена четкая тенденция к повышению эффективности классификации по сравнению с контрольным методом, не использующим предлагаемый в работе способ оценки контекстной значимости терминов. Для реальных тестовых наборов и обучающих выборок, составленных тестировщиками, повышение точности составило 10-20% при неизменных значениях полноты.

На всех тестовых наборах наблюдалось повышение эффективности классификации при увеличении объема обучающей выборки. Также на одном из наборов была специально исследована возможность дальнейшего повышения эффективности за счет ручной корректировки семантических образов рубрик. Ручная фильтрация терминов позволила повысить точность на 9%, а расширение образов рубрик подобранными вручную терминами повысило полноту на 43% при сохранении показателей точности.

Экспериментально оценивался вклад в эффективность классификации, который дает учет ассоциативных связей и распознавание МТК. Учет ассоциативных связей повышает точность классификации на 5-20% при незначительном увеличении полноты (3-5%). Учет метаконструкций по-

ложительно сказывается и на полноте и на точности классификации, общий прирост эффективности (мера F_β) составляет 10-15%.

В заключении сформулированы основные выводы и результаты данной работы, а также намечены пути дальнейшего развития систем автоматической классификации на основе механизмов полнотекстового поиска.

В приложениях приводятся фрагменты словарей коннекторов и индикаторов, а также примеры автоматически сформированных семантических образов рубрик.

Основные результаты работы

1. Предложен и разработан новый метод автоматической классификации, использующий алгоритмы полнотекстового поиска и ранжирования. Экспериментальная оценка показала, что по совокупности полноты и точности классификации данный метод заметно превосходит традиционные методы.
2. Предложен и разработан метод обучения классификатора, основанный на известном методе различительных сил, и отличающийся использованием обратной связи от процедуры полнотекстового поиска.
3. Разработана модель представления текстовых документов в виде взвешенных векторов словосочетаний. Веса словосочетаний определяются на основе лексических, контекстных и статистических характеристик, учет которых позволяет повысить точность на 10-20%.
4. Разработан метод повышения эффективности классификации, основанный на учете ассоциативных связей между терминами. Разработан оригинальный метод построения словарей ассоциативных связей при помощи анализа потока запросов к поисковой машине. Экспериментальная оценка метода показала 20-50% увеличение полноты или 4-5% увеличение точности в зависимости от способа учета ассоциативных связей.
5. Разработан метод повышения эффективности классификации за счет автоматического выделения в тексте метатекстовых конструкций. Экспериментальная оценка метода показала повышение полноты классификации на 5-15% при увеличении точности также на 5-15%/
6. Предложенные в данной работе методы, модели и алгоритмы реализованы в программном комплексе Классификатор компании МедиаЛингва, доведены до состояния коммерческого программного продукта, внедрены на нескольких предприятиях. Часть алгоритмов и методов внедрена в поисковую машину компании Рамблер.

Работы по теме диссертации

1. Применение статистических методов для интеллектуальной компьютерной обработки текстов / *И.С. Ашманов, С.П. Григорьев, В.М. Гусев и др.* // Диалог'97: Труды Международного семинара по компьютерной лингвистике и ее приложениям. - Ясная Поляна, 1997. - С. 33-37.
2. *Шабанов В. И.* Автоматическое индексирование запросов в документальной ИПС, основанное на статистической и морфологической информации. // КомпьюЛог.- 1997. - №3. - С. 20-24.
3. *Березкин Д.В., Шабанов В.И., Андреев А.М.* Методы выделения терминов из текста. // Современные информационные технологии: Межвузовская юбилейная научно-техническая конференция аспирантов и студентов. – М.: МГТУ им Н. Э. Баумана, 2001. - С. 117-127.
4. *Шабанов В.И., Андреев А.М., Сюзев В.В.* Построение ассоциативных связей в системах обработки текстов. // Современные информационные технологии: Сборник трудов кафедры ИУ6. – М.: МГТУ им. Н.Э. Баумана, 2002. – С. 191-196.
5. *Шабанов В.И., Власова А.Е.* Алгоритм формирования ассоциативных связей и его применение в поисковых системах. // Труды международной конференции Диалог`2003 по компьютерной лингвистике и интеллектуальным технологиям. – Протвино, 2003. - С. 603-609.
6. *Андреев А.М. Шабанов В.И.* Метод классификации текстовых документов, основанный на полнотекстовом поиске // Труды российского семинара по оценке методов информационного поиска – Санкт-Петербург, 2003 – С. 52-71
7. Модели и методы автоматической классификации текстовых документов. *А.М. Андреев, Д.В. Березкин, В.В Сюзев., В.И. Шабанов* // Вестн. МГТУ. Приборостроение. М.: - 2003. - №4. - С. 64-92

Подписано к печати 20.02.04г. Объем 1,0 п.л. Тираж 100 экз.
Заказ № 31г Типография МГТУ им. Н.Э. Баумана