МОСКОВСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ
имени М.В. Ломоносова
Факультет Вычислительной Математики и Кибернетики

На правах рукописи

Максаков Алексей Владимирович

ПОВЫШЕНИЕ РЕЛЕВАНТНОСТИ ПЕРИОДИЧЕСКОГО ТЕМАТИЧЕСКОГО ПОИСКА ИНФОРМАЦИИ В WEB

Специальность 05.13.11 – математическое и программное обеспечение вычислительных машин, комплексов и компьютерных сетей

АВТОРЕФЕРАТ
диссертации на соискание ученой степени
кандидата физико-математических наук

МОСКВА
2007

Работа выполнена на кафедре автоматизации систем вычислительных комплексов факультета вычислительной математики и кибернетики МГУ им М.В. Ломоносова

Научный руководитель:	доктор физико-математических наук, профессор Руслан Леонидович Смелянский
Официальные оппоненты:	доктор физико-математических наук, заведующий сектором Института прикладной математики РАН Михаил Михайлович Горбунов-Посадов
	кандидат физико-математических наук, старший научный сотрудник ВЦ РАН Виктор Иванович Филиппов
Ведущая организация:	Институт системного программирования Российской академии наук

Защита диссертации состоится 26 октября 2007 г. в 11:00 на заседании диссертационного совета Д 501.001.44 при Московском государственном университете им. М.В, Ломоносова по адресу: 119992, ГСП-2, Москва, Воробьевы горы, МГУ, 2-ой учебный корпус, факультет ВМиК, аудитория 685.

С диссертацией можно ознакомиться в библиотеке факультета ВМиК МГУ. С текстом автореферата можно ознакомиться на официальном сайте ВМиК МГУ им. М.В. Ломоносова http://www.cmc.msu.ru в разделе «Наука» - «Работа диссертационных советов» - «Д 501.001.44»

Автореферат разослан « » сентября 2007 года.

Ученый секретарь

диссертационного совета

профессор Н.П. Трифонов

Общая характеристика работы

Актуальность темы

Развитие сетевых технологий, в том числе и сети Интернет, привело к значительному увеличению числа доступных информационных ресурсов и объемов передаваемой информации. Зачастую это разнородная и слабо структурированная информация, обладающая высокой динамикой обновления. Необходимость эффективного использования этого колоссального и динамично изменяющегося объема информации обуславливает актуальность и значимость исследований в области информационного поиска. Согласно опросу пользователей Интернет[1] более 63% от общего числа пользователей используют системы поиска в World Wide Web практически каждый день.

В области информационного поиска отдельно выделяется задача тематического поиска, то есть целенаправленного поиска документов, относящихся с той или иной степенью релевантности к определенной теме, заявленной пользователем. При проведении исследований, обучении и профессиональной деятельности, в связи с высокой скоростью появления новой информации возникает потребность не только в нахождении сведений, соответствующих одной или нескольким темам, но и в постоянном получении новых данных. Одним из возможных вариантов удовлетворения этой потребности является периодическое обновление ранее полученных сведений, по аналогии с подпиской на тематические издания, такие как специализированные газеты и журналы. Для обеспечения такого рода доставки информации из Web необходимо решить задачу периодического тематического поиска, то есть такого поиска, который ведется систематически, через определенные промежутки времени. Причем ищутся не только обновления на уже известных Web-сайтах, но и новые сайты.

[1] Rainie L. Search Engine use November 2005.
http://www.pewinternet.org/pdfs/PIP_SearchData_1105.pdf

Следует отметить, что сервис периодической доставки релевантных документов имеет практический смысл только в том случае, если тематическая потребность остается актуальной и слабо изменяется в течение большого промежутка времени. В этих условиях можно сформулировать следующие особенности задачи периодического тематического поиска в Web:

- Высокая динамичность и объем пространства поиска (согласно оценкам ежемесячно изменяется до 40%[2] общего объема доступной информации, составляющего более чем 11 млрд. web-страниц)
- Информационная потребность пользователя слабо изменяется со временем и остается актуальной в течение большого промежутка времени (носит долговременный характер).
- Результаты поиска необходимо предоставить пользователю в течение ограниченного, вполне определенного интервала времени.

Высокая динамичность пространства поиска и слабая изменчивость информационной потребности позволяют отнести задачу периодического тематического поиска к классу задач фильтрации информации[3]. В традиционных методах фильтрации для описания информационной потребности используются как наборы ключевых слов, так и обучающие коллекции документов. Отличие задачи фильтрации на всем Web от традиционной задачи фильтрации состоит в том, что протокол передачи данных в Web HTTP[4] реализует модель “запрос-ответ” и не позволяет оповещать об изменениях в данных. Это приводит к тому, что обнаружить все изменения в Web, можно только проанализировав всю доступную информацию, объем которой очень велик. В связи с этим в существующих методах фильтрации информации на всем Web для описания

[2] Kahle B. Preserving the Internet// Scientific American: March 1997. p. 82-83
[3] Belkin N., Croft W. Information filtering and information retrieval: two sides of the same coin?// Communications of the ACM, Volume 35 , Issue 12. New York: ACM Press, 1992. p. 29-38.
[4] Liu L. Query routing in large-scale digital library systems// In proc. of the 15th conference on Data Engineering. IEEE press, 1999. p. 154-163.

информационной потребности используются только наборы ключевых слов. К недостаткам методов поиска, использующих запросы по ключевым словам для представления информационной потребности, относят слабую выразительность языка запросов и высокую трудоемкость составления оптимального запроса[5], что приводит к низкому качеству тематического поиска в Web. С другой стороны, существует множество успешно применяемых методов определения тематической принадлежности документов, в том числе и с использованием алгоритмов классификации (или *методов машинного обучения*[6]), использующих обучающие коллекции документов. Однако высокая вычислительная сложность задач обучения и классификации ограничивает практическую применимость таких методов для Web.

В этих условиях разработка метода периодического тематического поиска в Web в условиях долговременности информационной потребности пользователя и динамичности пространства поиска, повышающего качество поиска по сравнению с традиционными методами, представляется актуальной.

Цель работы

Целью диссертационного исследования является создание метода периодического тематического поиска информации в Web, обладающего более высоким качеством поиска по сравнению с другими известными методами.

В рамках данной работы исследуются решения следующих задач:

1. Разработка метода периодического тематического поиска, обладающего свойствами методов поиска по ключевым словам в части полноты охвата информационных источников в Web и свойствами методов тематической фильтрации, основанных на

[5] Kobayashi M., Takeda K. Information retrieval on the Web// ACM Computing Surveys, vol.32, 2. New York: ACM Press, 2000. p. 144-173.

[6] Агеев М. С. Методы автоматической рубрикации текстов, основанные на машинном обучении и знаниях экспертов. Дис. канд. физ-мат. наук: 05.13.11. Московский гос. унив. - Москва, 2005.

машинном обучении, в части высокой точности результатов поиска.

2. Исследование существующих методов машинного обучения с целью повышения их эффективности для использования в создаваемом методе периодического тематического поиска.
3. Анализ эффективности предложенного метода в сравнении с существующими методами периодического тематического поиска информации в Web.

Научная новизна

В работе предложен новый метод решения задачи периодического тематического поиска в Web, основанный на комбинации поиска по ключевым словам и тематической фильтрации с использованием классификаторов текстов, применяемой в ограниченных по объему коллекциях документов.

Разработаны алгоритмы классификации, зависимость вычислительной сложности обучения которых от числа примеров в обучающей выборке близка к линейной, при этом качество классификации близко к качеству лучших известных алгоритмов, обладающих более высокой вычислительной сложностью обучения.

Предложен способ оценки весов признаков, применение которого совместно с методом опорных векторов позволяет повысить качество классификации по сравнению с другими известными способами.

Практическая ценность

В работе показано, что предложенный метод обеспечивает более высокое качество поиска по сравнению с существующими. Проведена экспериментальная реализация предложенного метода.

Результаты данной работы могут быть использованы в следующих областях:

1. Для организации поиска новой информации в системах непрерывного обучения.

2. При создании сервиса периодической доставки новой релевантной информации для информационной поддержки различного рода профессиональной деятельности, в том числе и научной.

Методы исследования

При разработке метода периодического тематического поиска в Web использовались методы статистического и лингвистического анализа текстов на естественном языке, методы математической статистики, методы экспериментальной проверки.

Апробация работы и публикации

По теме диссертационного исследования опубликовано 7 печатных работ, в том числе одна – в издании из списка рекомендованных ВАК. Результаты работы докладывались на объединенном научно-исследовательском семинаре кафедр Автоматизации систем вычислительных комплексов, Алгоритмических языков и Системного программирования факультета ВМиК МГУ, на научных семинарах лаборатории Вычислительных комплексов кафедры Автоматизации систем вычислительных комплексов факультета ВМиК МГУ, Российском семинаре по Оценке Методов Информационного поиска (РОМИП), а также на следующих конференциях:

- Международная научная конференция “Интеллектуализация обработки информации” (Алушта, 2002);
- Всероссийская научная конференция "Математические методы распознавания образов" (Звенигород, 2005);
- Всероссийская научная конференция "Электронные библиотеки: перспективные методы и технологии, электронные коллекции" (Суздаль, 2006).

Структура и объем диссертации

Диссертация состоит из введения, четырех глав, заключения и списка литературы. Объем диссертации – 117 страниц, список литературы содержит 116 наименований.

Краткое содержание диссертации

Во **введении** дано краткое описание задачи периодического тематического поиска информации в Web и ее отличительных особенностей. Обоснована актуальность задачи и приводится краткий обзор содержания диссертации.

В **первой главе** приведен обзор методов информационного поиска в Web и тематической фильтрации с точки зрения эффективности их применения для решения задачи периодического тематического поиска.

В **разделе 1.1** приведены основные особенности Web, как источника информации. Введено понятие индекса поисковой системы. Показана необходимость учета показателя степени соответствия локального индекса реальному состоянию Web при общей оценке метода поиска в Web.

Далее, в **разделе 1.2**, приведены традиционные показатели качества поиска и общепринятая методология оценки качества поиска, применяемая в рамках конференции Text Retrieval Evaluation Conference (TREC) и Российского семинара по Оценке Методов Информационного Поиска (РОМИП).

В **разделе 1.3** вводятся критерии эффективности методов периодического тематического поиска, которые включают в себя как критерии, связанные с качеством поиска (*качество поиска на статической коллекции документов*, *доля упущенной в процессе поиска информации, возможность обнаружения новых источников информации в Web)*, так и критерии, оценивающие возможности и удобство практического применения

реализации метода (*трудоемкость описания информационной потребности, универсальность и практическая применимость*).

Раздел 1.4 посвящен анализу существующих решений в области периодического поиска с точки зрения их соответствия указным ранее критериям.

Существующие решения в области периодического поиска можно разделить на 4 класса: периодический поиск с использованием систем поиска по ключевым словам (СПКС), периодический поиск с использованием мета-информационнных поисковых систем, периодический поиск новой информации на подмножестве источников информации Web и поиск обновлений в тематических каталогах. Показана низкая эффективность применения СПКС для решения задачи периодического тематического поиска. Показано, что для СПКС также характерна высокая степень несоответствия между данными, хранящимися в поисковом индексе и реальным состоянием Web. Частично эту проблему несоответствия можно решить путем использования при периодическом поиске мета-информационных поисковых систем. Полнота периодического поиска новой информации на подмножестве источников информации Web оценивается как низкая[7], и падает со временем, поскольку не предусмотрена возможность обнаружения новых источников в Web. Тематические каталоги строятся вручную, что приводит к очень низкой полноте периодического поиска. Также ограничение набора тем в каталогах приводит к нарушению требования универсальности поиска.

В **разделе 1.5** рассматривается возможность применения методов тематической фильтрации, основанных на машинном обучении для решения задачи периодического тематического поиска информации в Web. Показано, что так называемые *линейные классификаторы* (алгоритмы построения разделяющей гиперплоскости в линейном пространстве признаков)

[7] Bun K., Ishizuka M. Emerging Topic Tracking System// Proceedings of Web Intelligence Conference. London: Springer-Verlag, 2001. p. 125-130.

превосходят по качеству классификации *алгоритмы индукции правил*. Показано, что практическое применение методов тематической фильтрации с использованием линейных классификаторов ограничено высокой вычислительной сложностью классификации.

Проведенный в **разделе 1.6** сравнительный анализ различных методов периодического тематического поиска в Web позволил выбрать в качестве направления дальнейшей работы разработку метода периодического тематического поиска, который бы обладал приемлемой вычислительной сложностью нахождения релевантных документов и высоким качеством тематической фильтрации, характерным для методов фильтрации с использованием линейных классификаторов.

Во **второй главе** приводится описание и исследование предложенного автором метода периодического тематического поиска, основанного на комбинации метода поиска по ключевым словам и метода тематической фильтрации с помощью классификаторов текстов.

В предложенном методе поиска информационная потребность представляется в виде пары

$\{q, D\}$, где q – запрос по ключевым словам, использующийся для первичного отбора документов из Web

$D = \{D+, D-\}$ – обучающая выборка, описывающая тему, интересующую пользователя. Данная обучающая выборка содержит примеры релевантных теме документов ($D+$) и нерелевантных документов ($D-$).

Процесс поиска разделяется на два этапа:

1. Отбор документов из Web, соответствующих запросу по ключевым словам q с помощью глобальных систем поиска по ключевым словам, таких как *Google*, *Яндекс* и т.п. Данный этап позволяет с одной стороны обеспечить высокую полноту

поиска, а с другой – существенно сократить объем обрабатываемой на следующем этапе информации.

2. Уточнение результатов поиска с помощью классификатора, обученного на предоставленной пользователем обучающей выборке *D*. Этот этап позволяет обеспечить высокую точность результатов поиска.

Ряд исследований[8,9] показали, что применение тематической классификации результатов поиска позволяет существенно сократить время поиска нужной информации. Разбиение линейного списка результатов поиска на тематически связанные подгруппы позволяет быстрее ориентироваться в полученных результатах и, как следствие, повысить удобство использования поисковой системы.

Для реализации классификации результатов поиска пользователю необходимо разбить в обучающей выборке множество релевантных документов *D*+ на подмножества, описывающие интересующие пользователя подтемы. В этом случае обучающая выборка будет представлять собой множество

$D=\{D_1+, D_2+, D_3+,\ldots, D_n+, D-\}$, где D_i+ - обучающая выборка *i*-ой подтемы, *n* – общее количество подтем.

Таким образом, в рамках предложенного метода классификатор применяется для решения задачи тематической фильтрации (*бинарной классификации*) и задачи разбиения множества релевантных теме документов на подтемы (задачи классификации с большим количеством классов в обучающей выборке).

Пользователь также может осуществить обратную связь с системой, изменив обучающую выборку, что приводит к необходимости дообучения

[8] Chen H., Dumais S. Bringing Order to the Web: Automatically Categorizing Search // Proceedings of ACM SIGCHI Conference on Human Factors in Computing Systems, Vol. 1. New York: ACM Press, 2000. p. 145-152.

[9] Diao Y., Lu H., Wu D. A comparative study of classification-based personal E-Mail filtering// In proceedings of 4th Pacific-Asia Conference on Knowledge Discovery and Data Mining. Kyoto: Springer Verlag, 2000. p.408-419.

классификатора. Для того чтобы пользователь оперативно смог оценить влияние обратной связи на качество поиска, дообучение классификатора должно занимать как можно меньше времени.

В **разделе 2.4** приводится аналитическое исследование предложенного метода. В первой главе был сделан вывод о том, что требованиям универсальности, практической применимости и высокой полноты поиска, одновременно удовлетворяют методы, основанные на представлении информационной потребности пользователя в виде запроса по ключевым словам.

Поскольку на этапе отбора документов из Web используется тот же метод поиска по ключевым словам, то *доля упущенной в процессе поиска информации* будет не больше, чем при прямом применении этого метода. Таким образом, можно утверждать, что предложенный метод (обозначим его $M_{гибр}$) не будет уступать методу поиска по ключевым словам $M_{КС}$, при условии, что он не будет уступать ему по *качеству поиска на статической коллекции документов*. Для сравнения качества поиска на статической коллекции необходимо сравнить оценки полноты и точности этих методов.
Согласно разделу 2.2 процесс поиска состоит из двух этапов:

1. отбор соответствующих запросу q' документов из Web с помощью метода поиска по ключевым словам $M_{кс}(q')$;
2. уточнение результатов поиска с помощью метода тематической фильтрации $M_{тф}$, реализованного в виде линейного классификатора, обученного на предоставленной пользователем обучающей выборке D.

Выразим показатели полноты и точности предложенного метода через показатели полноты и точности отбора документа из Web с помощью поиска по ключевым словам $M_{кс}(q')$ и метода тематической фильтрации $M_{тф}$. Обозначим общее количество релевантных документов как N_p, количество отобранных на первом этапе – N_o, количество релевантных из отобранных -

$N_{ро}$ и количество релевантных документов из итогового списка найденных – $N_{рн}$:

$$P(M_{гибр}) = \frac{N_{рн}}{N_o} = P(M_{тф})$$

$$R(M_{гибр}) = \frac{N_{рн}}{N_p} = \frac{N_{ро}}{N_p} \cdot \frac{N_{рн}}{N_{ро}} = R(M_{кс}(q')) \cdot R(M_{тф})$$

Таким образом, точность предложенного метода равна точности тематической фильтрации $M_{тф}$, используемой на втором этапе поиска. Полнота же определяется произведением полноты отбора документов Web по ключевым словам q' на полноту тематической фильтрации.

Покажем, что при выполнении определенных условий предложенный метод будет превосходить по качеству поиска, выраженному мерой F1, метод поиска по ключевым словам.

Запрос по ключевым словам, результаты поиска по которому обладают наилучшим среди данного множества запросов Q показателем меры F1, будем называть F1-оптимальным запросом на этом множестве.

$$q = \arg\max_{q \in Q} F1(q)$$

Полноту поиска с помощью F1-оптимального запроса обозначим $R(q)$, точность - $P(q)$.

Предположим, что на практике верны следующие утверждения:

1. Применяемый классификатор превосходит по полноте F1-оптимальный запрос на множестве запросов Q, которые может предложить пользователь, т.е. $\exists \alpha > 1 : R(M_{тф}) \geq \alpha \cdot R(q)$ (1),
2. F1-оптимальный запрос обладает полнотой меньше единицы $R(q) < 1$ (2)

Введем дополнительное условие:

3. Возможно подобрать запрос по ключевым словам q', уменьшающий ошибку, связанную с полнотой в произвольное количество раз

$$\exists q' : R(q') > R(q) \quad (3)$$

Лемма: Предположим, что утверждения (1), (2) верны, а также выполняется условие (3) , причем

$$R(q') \geq R(q) + b \cdot (1 - R(q))$$

Тогда $\forall \alpha > 1 \quad \exists b < 1 : R(M_{гибр}) = R(q') \cdot R(M_{тф}) \geq R(q)$

Доказательство:

Получим оценку значений параметра *b*, при которых выполняется условие $R(M_{гибр}) = R(q') \cdot R(M_{тф}) \geq R(q)$:

$$\alpha \cdot R(q) \cdot (R(q) + b \cdot (1 - R(q)) \geq R(q),$$

$$R(q) + b \cdot (1 - R(q)) \geq 1/\alpha,$$

$$b \geq \frac{1/\alpha - R(q)}{1 - R(q)}.$$

Поскольку $\alpha > 1$, то $\frac{1/\alpha - R(q)}{1 - R(q)} < 1$.

Обозначим $\lambda = \frac{1/\alpha - R(q)}{1 - R(q)}$.

Таким образом, требуемое условие выполняется при $b \in [\lambda, 1)$. Т.к. данное множество непустое, то при выполнении условия (3):

$\forall \alpha > 1 \quad \exists b < 1 : R(M_{гибр}) = R(q') \cdot R(M_{тф}) \geq R(q)$ - что и требовалось доказать.

Следствие. При выполнении указанных в лемме условий можно подобрать такой запрос по ключевым словам, что полнота гибридного подхода будет превосходить полноту поиска по F1-оптимальному запросу.

Множество *Q* может быть составлено двумя способами:

- Путем включения множества ad-hoc запросов, предложенных пользователями для описания заданной темы
- Путем получения запроса с помощью *алгоритма индукции правил*, обученного на представленной пользователем обучающей выборке.

Показано, что для запросов полученных с помощью алгоритмов индукции правил утверждение (1) верно на практике в большинстве случаев. При этом *линейные классификаторы* превосходят алгоритмы индукции

правил и по точности классификации. Результаты апробации, приведенные в разделе 4.3, подтверждают, что качество поиска по F-мере с помощью предложенного метода выше качества поиска с помощью F-оптимального запроса на множестве запросов, предложенных пользователями. Таким образом, если используемый классификатор превосходит по точности F1-оптимальный запрос, то качество результатов гибридного подхода, выраженное мерой F1 в среднем будет выше, чем качество поиска по ключевым словам, при условии подбора для отбора документов из Web запроса по ключевым словам, обладающего достаточно высокой полнотой. Теоретически возможно составление запроса, обладающего полнотой, равной единице, путем включения в него всех слов, которые могут встречаться в текстах, относящихся к заданной теме (таким образом, условие (3) выполнимо).

С практической точки зрения преимущество предложенного метода состоит в том, что для получения высокого качества поиска необходимо составить запрос по ключевым словам, обеспечивающий высокую полноту поиска, в отличие от традиционного поиска по ключевым словам, в котором требуется составление запроса, обеспечивающего одновременно высокую полноту и точность.

В **третьей главе** приводится обзор и сравнительный анализ алгоритмов классификации, эффективность которых в значительной степени определяет эффективность предложенного метода в целом. Рассматриваются предложенные автором масштабируемые алгоритмы классификации.

В **разделах 3.1 – 3.3** приводятся общие требования к алгоритмам классификации в рамках рассматриваемой задачи, метрики качества классификации, основные этапы классификации текстов.

В **разделе 3.4** описаны основные способы предварительной обработки текстов: применение морфологического, синтаксического анализа, способы сокращения пространства признаков, отбора фраз, а также оценки весов

признаков. Показано, что прямое применение методов кластеризации признаков неэффективно при решении задачи периодического тематического поиска. Описан метод отбора фраз, основанный на композиции синтаксического и статистического подходов. Предложен новый способ оценки значимости признака в коллекции, учитывающий, в отличие от традиционной оценки в виде инверсной частоты признака, характер распределения признака по классам в обучающей выборке.

В **разделах 3.5-3.6** приведено описание методики оценки алгоритмов классификации, введены критерии сравнения эффективности алгоритмов классификации в рамках решения задачи периодического тематического поиска.

В **разделах 3.7-3.8** приводится обзор широко известных алгоритмов классификации и их сравнительный анализ. На основе ряда независимых публикаций[10,11] делается вывод о преимуществе по качеству классификации алгоритма SVM (метода опорных векторов) над другими известными алгоритмами. Однако к недостаткам этого алгоритма следует отнести высокую вычислительную сложность обучения ($O(N^a)$, где $a > 1,7$[12]). Исходя из наличия требований к низкой вычислительной сложности обучения, требуется разработка алгоритмов, которые бы обладали меньшей вычислительной сложностью обучения при качестве классификации, близком к качеству алгоритма SVM.

Описание таких алгоритмов, предложенных автором, приводится в **разделе 3.9**. Предложена модификация метода Байеса, использующая парадигму класса-дополнения и нормализацию весов признаков для устранения систематических недостатков, присущих исходному алгоритму.

---

[10] Sebastiani F., Machine Learning in Automated Text Categorization// ACM Computing Surveys, vol.1, 2002. p. 1-47.

[11] Yang Y., Liu X. A re-examination of text categorization methods// Proc. of International ACM Conf. on Research and Development in Information Retrieval (SIGIR-99). New York: ACM Press, 1999. p. 42-49.

[12] Chakrabarti S. Mining The Web Discovering Knowledge From Hypertext Data. San Francisco: Morgan Kaufmann Publishers, 2004.

Поскольку применение парадигмы класса-дополнения эффективно только при наличии большого количества классов в обучающей выборке, для решения задачи бинарной классификации был предложен другой алгоритм – алгоритм построения нескольких разделяющих гиперплоскостей, основанный на последовательном нахождении нескольких дискриминантов Фишера. Все предложенные алгоритмы обладают вычислительной сложностью обучения, близкой к $O(N)$, где N – число документов в обучающей выборке.

В **разделе 3.10** рассмотрен предложенный автором способ оценки весов признаков, основанный на описанном в разделе 3.4 новом способе оценки значимости признака.

Экспериментальные исследования, проведенные на четырех широко применяемых тестовых коллекциях документов, показали следующее:

- При решении задачи бинарной классификации качество результатов предложенного алгоритма построения нескольких разделяющих гиперплоскостей (ModFisher) сопоставимо с качеством алгоритма SVM.
- Модифицированный метод Байеса при решении задачи бинарной классификации существенно проигрывает по качеству и алгоритму SVM и алгоритму ModFisher.
- При решении задачи классификации с большим количеством классов модифицированный метод Байеса превосходит алгоритм ModFisher по качеству классификации, а также и алгоритм SVM без применения предложенного в работе модификатора весов признаков на большинстве тестовых коллекций.
- На всех тестовых коллекциях метод опорных векторов с применением предложенного в работе способа оценки весов признаков превосходит по качеству классификации все рассмотренные алгоритмы.
- Использование фраз в качестве признаков позволяется повысить качество классификации на 1-2% для всех рассмотренных алгоритмов,

за исключением алгоритма SVM, для которого улучшение было незначительно и составило 0,4%.

- Сокращение пространства признаков с использованием критерия хи-квадрат или критерия соотношения порядков позволяет сократить размерность пространства в 5-10 раз без существенной потери качества классификации

В **четвертой главе** приведено описание архитектуры прототипа системы периодического тематического поиска в Web. Описан механизм получения множества анализируемых классификатором документов из Web, который включает в себя реализацию мета-поискового подхода, рекурсивного обхода выбранного множества ресурсов и тематически ориентированного обхода Web.

Приводятся результаты апробации предложенного метода на реальных данных, в том числе и с учетом периодической составляющей поиска. Исследовано влияние обратной связи пользователя с системой на качество поиска. При проведении экспериментального исследования использовался метод экспертной оценки. Оценка производилась на реальных данных Web.

Эксперименты показали, что предложенный подход обеспечивает более высокое качество поиска по сравнению с методом поиска по ключевым словам, как по точности, так и по мере F1. Учет обратной связи пользователя с системой, согласно результатам экспериментов, также позволяет увеличить качество периодического тематического поиска.

В **заключении** сформулированы основные результаты диссертации и направления дальнейшего развития.

Основные результаты работы

1. Предложен новый метод периодического тематического поиска информации в Web, созданный на основе композиции

метода поиска по ключевым словам и метода тематической фильтрации с помощью классификаторов текстов. Данный метод учитывает долговременный характер информационной потребности и динамичность пространства поиска и позволяет повысить релевантность результатов поиска.

2. Разработаны оригинальные масштабируемые алгоритмы классификации, обладающие меньшей вычислительной сложностью обучения и сопоставимым качеством классификации по сравнению с одним из лучших известных алгоритмов – методом опорных векторов:
   - алгоритм на основе построения нескольких разделяющих гиперплоскостей для решения задачи бинарной классификации
   - модифицированный алгоритм Байеса для случая большого количества классов в обучающей выборке
3. Сформулированы условия эффективного совместного применения алгоритмов классификации и способов предварительного анализа текста при построении систем периодического тематического поиска.
4. Реализован прототип системы периодического тематического поиска в Web и получены экспериментальные оценки полноты и точности предложенного метода, показывающие его преимущество перед существующими методами.

Представленные результаты получены автором самостоятельно и изложены в следующих работах:

1. Максаков А.В. Исследование способов уменьшения набора характеристик в алгоритмах классификации текстов// Труды Всероссийской научной конференции "Методы и средства

обработки информации". М.: Издательский отдел факультета ВМиК МГУ, 2003, стр. 234-240.

2. Максаков А.В. Масштабируемые алгоритмы классификации текстов// Труды 12-й конференции "Математические методы распознавания образов" (ММРО-12), Москва, 2005.
3. Максаков А.В. Об одном методе периодического тематического поиска информации в Web// Труды восьмой всероссийской научной конференции "Электронные библиотеки: перспективные методы и технологии, электронные коллекции". М.: МАКС Пресс, 2006, стр. 101-107
4. Максаков А.В. Об одном методе повышения качества периодического тематического поиска в Web// Вестн. Моск. ун-та. Сер.15. Вычислительная математика и кибернетика, 2007. № 2, стр. 35-44.
5. Максаков А.В. Оценка эффективности масштабируемых алгоритмов классификации текстов// Труды четвертого российского семинара по оценке методов информационного поиска. Санкт-Петербург: НУ ЦСИ, 200, стр. 92-100
6. Максаков А.В. Обеспечение контекстного поиска информации для баз знаний// Искусственный интеллект (Донецк), 2002 № 2, стр. 493-500
7. Максаков А.В. Сравнительный анализ алгоритмов классификации и способов представления Web-документов// Труды третьего Российского семинара по Оценке Методов Информационного Поиска (РОМИП 2005). Санкт-Петербург: НИИ Химии СПбГУ, 2005, стр. 63-73.