Московский государственный университет
имени М.В. Ломоносова

На правах рукописи

Глазкова Валентина Владимировна

ИССЛЕДОВАНИЕ И РАЗРАБОТКА МЕТОДОВ ПОСТРОЕНИЯ ПРОГРАММНЫХ СРЕДСТВ КЛАССИФИКАЦИИ МНОГОТЕМНЫХ ГИПЕРТЕКСТОВЫХ ДОКУМЕНТОВ

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 2008

Работа выполнена на кафедре автоматизации систем вычислительных комплексов факультета вычислительной математики и кибернетики Московского государственного университета имени М.В. Ломоносова.

Научные руководители:	доктор физико-математических наук, профессор Машечкин Игорь Валерьевич; кандидат физико-математических наук, доцент Петровский Михаил Игоревич
Официальные оппоненты:	доктор технических наук, кандидат физико-математических наук, профессор Кузнецов Сергей Дмитриевич; доктор технических наук, кандидат физико-математических наук, профессор Рыжов Александр Павлович.
Ведущая организация:	Межведомственный суперкомпьютерный центр РАН

Защита диссертации состоится "_28_" _ноября_ 2008 г. в 11:00 часов на заседании диссертационного совета Д 501.001.44 в Московском государственном университете им. М.В. Ломоносова по адресу: 119991, ГСП-1, Москва, Ленинские горы, МГУ, 2-й учебный корпус, факультет ВМиК, аудитория 685.

С диссертацией можно ознакомиться в библиотеке факультета ВМиК МГУ. С текстом автореферата можно ознакомиться на официальном сайте факультета ВМиК Московского государственного университета имени М.В. Ломоносова http://www.cs.msu.su в разделе «Наука» - «Работа диссертационных советов» - «Д 501.001.44».

Автореферат разослан "____" _октября_ 2008 г.

Ученый секретарь
диссертационного совета
профессор Н.П. Трифонов

Общая характеристика работы

Актуальность темы

Настоящая работа посвящена исследованию и разработке методов построения программных средств классификации многотемных гипертекстовых документов. *Задача классификации многотемных документов (multi-label classification)*, заключается в определении принадлежности документа к одному или нескольким классам (из предопределённого набора классов) на основании анализа совокупности признаков, характеризующих данный документ. Классы, к которым принадлежит документ, называются *релевантными* для данного документа. Классы в рассматриваемой задаче не являются взаимоисключающими (как в традиционной постановке задачи классификации), а могут пересекаться и быть вложенными.

Разработка подходов и алгоритмов решения задачи классификации многотемных документов – это относительно новое направление исследований, которое в настоящее время активно развивается за рубежом и в России. Большинство существующих подходов является альтернативой непосредственного сведения задачи классификации многотемных документов к традиционной задаче классификации, характеризующейся тем, что классифицируемый объект может принадлежать только к одному классу (*multi-class classification*).

На сегодняшний день существует ряд актуальных прикладных задач, при решении которых возникает необходимость разработки программных средств классификации многотемных документов. К числу таких задач относятся: категоризация электронной почты; мониторинг документооборота пользователей и предотвращение утечек конфиденциальной информации; анализ и фильтрация Интернет-трафика; категоризация документов в электронных библиотеках и другие. Во всех перечисленных приложениях возникает необходимость решения задачи классификации, причем *классифицируемый документ имеет многотемную природу*, и для принятия решения необходимо знать набор всех классов, релевантных для документа.

Для решения перечисленных прикладных задач требуется применение методов классификации *на основе машинного обучения*, поскольку состав и содержимое анализируемых документов постоянно изменяется, и одним из путей адаптации к этой динамике является использование таких методов. Цель методов машинного обучения для задачи классификации многотемных документов заключается в построении модели классификации на основе обучающего набора и применении построенной модели для предсказания набора классов, релевантных для нового документа. *Обучающий набор* для рассматриваемой задачи классификации состоит из документов, каждому из которых

сопоставлено множество релевантных классов. В рассматриваемых прикладных задачах обучающие наборы имеют достаточно большой размер, ввиду чего при решении этих задач необходимо применение методов классификации с возможностью *дообучения* без необходимости хранения обучающего набора (*incremental learning*, *пошаговое обучение*). При пошаговом обучении обучающие данные подаются алгоритму последовательно (по одному примеру на каждом шаге обучения), и на последующих шагах алгоритм использует *только новые* обучающие примеры. При традиционном пакетном обучении (batch learning), в отличие от пошагового, для обучения алгоритма классификации весь обучающий набор должен быть задан целиком.

Возможность пошагового обучения может включать в себя различные сценарии функционирования, связанные как с уточнением модели классификации на новых обучающих примерах в рамках предопределённого набора тематик, так и с модификацией модели классификации при *удалении существующих и добавлении новых тематик* классификации с новыми обучающими примерами. Специфика перечисленных прикладных задач такова, что набор интересующих тематик классификации может динамически изменяться. Например, в задачах предотвращения утечек конфиденциальной информации и фильтрации Интрнет-трафика список запрещённых и разрешённых тематик может пополняться, в то время как некоторые из существующих тематик могут становиться неактуальными в процессе работы прикладной системы. В задаче категоризации электронной почты пользователь может создавать новые тематические папки и удалять существующие, и в этом случае необходимо, чтобы алгоритм классификации дообучался с учётом этих изменений. Модификации набора категорий могут быть достаточно частыми и должны оперативно отражаться в модели классификации, чтобы анализ новых сообщений осуществлялся относительно наиболее актуального набора категорий. Поэтому важно, чтобы для учёта изменений набора тематик не приходилось заново обучать модель классификации.

Таким образом, актуальна разработка программных средств классификации многотемных документов, обеспечивающих следующие *сценарии функционирования*: обучение на основе обучающего набора, классификация документов, дообучение на новых документах без необходимости хранения предыдущего обучающего набора, добавление и удаление тематик классификации. Разрабатываемые программные средства должны удовлетворять требованиям к производительности, предъявляемым современными прикладными задачами. Качественной неформальной оценкой этих требований будем считать, что скорость классификация документов должна соответствовать интерактивному режиму работы пользователей.

Ещё одной важной подзадачей при создании программных средств классификации многотемных гипертекстовых документов является разработка модели представления гипертекстовых документов. Это обусловлено тем, что для алгоритма классификации на основе машинного обучения выбор модели представления документов влияет на большинство важных критериев оценки алгоритма, таких как скорость обучения и классификации, точность, размер модели. Формальной моделью описания электронных документов, с которыми работают обозначенные прикладные проблемы, является гипертекст. *Гипертекстовая модель представления* определяется ориентированным графом, в вершинах которого располагаются блоки содержательной информации. Эти блоки имеют смысловую связь, фиксируемую дугами и ребрами графа. Благодаря этому гипертекст отличается от обычного линейного текста, который имеет последовательную структуру. Учёт гиперссылок в документе может позволить получить более точное (для классификации) представление, по сравнению с учётом только локального содержимого (контента) классифицируемого документа [1].

Таким образом, на сегодняшний день является актуальным проведение исследований и разработка программных средств, осуществляющих классификацию многотемных гипертекстовых документов на основе методов машинного обучения с возможностью дообучения и добавления (удаления) категорий классификации.

Цель работы

Цель настоящей работы – исследование и разработка методов и построение программных средств классификации многотемных гипертекстовых документов, учитывающих специфику актуальных прикладных задач и обеспечивающих б*о*льшую эффективность по сравнению с существующими методами.

В рамках данной работы необходимо решение следующих задач:

1. исследование и разработка методов для классификации многотемных гипертекстовых документов на основе машинного обучения с возможностью дообучения и добавления (удаления) категорий классификации;
2. анализ эффективности разработанных методов по сравнению с существующими;
3. разработка программного модуля классификации многотемных гипертекстовых документов на основе предложенных методов.

Методы исследования

При разработке программного модуля классификации использовались методы объектно-ориентированного анализа и проектирования. При разработке методов

[1] Soumen Chakrabarti, Byron E. Dom, Piotr Indyk. Enhanced hypertext categorization using hyperlinks // Proceedings of the ACM International Conference on Management of Data, SIGMOD, 1998. pp. 307-318

классификации многотемных гипертекстовых документов использовались методы теории машинного обучения, математической статистики и анализа текстов на естественном языке, а также численные эксперименты на ЭВМ.

Научная новизна

В настоящей работе предложен новый метод классификации многотемных документов, основанный на подходе попарных сравнений с помощью набора бинарных классификаторов, где результирующие степени принадлежности документа классам (релевантности классов) вычисляются с помощью обобщенной модели Брэдли-Терри, а нерелевантные классы отсекаются с помощью пороговой функции, заданной в пространстве релевантностей классов.

Разработана модель представления гипертекстовых документов основанная на учёте гиперссылок посредством анализа структуры адресов документов и на расширении традиционной векторной модели представления за счёт добавления частых комбинаций признаков.

Практическая ценность

На основе разработанных методов спроектирован и реализован программный модуль классификации многотемных гипертекстовых документов. Разработанный модуль поддерживает следующие сценарии функционирования: обучение на основе обучающего набора, классификация документов, дообучение на новых документах без необходимости хранения обучающего набора, добавление и удаление тематик классификации.

Разработанный модуль может быть применён для широкого спектра прикладных задач, таких как категоризация электронной почты; анализ и фильтрация Интернет-трафика; мониторинг документооборота пользователей; категоризация документов в электронных хранилищах данных. Разработанный модуль апробирован в системе анализа и фильтрации Интернет-трафика на факультете ВМиК МГУ им. М.В.Ломоносова.

Апробация работы и публикации

Результаты, представленные в работе, докладывались на объединённом научно-исследовательском семинаре кафедр Автоматизации систем вычислительных комплексов, Системного программирования и Алгоритмических языков факультета ВМиК МГУ под руководством профессора М.Р.Шура-Бура, на научных семинарах лаборатории Технологий программирования факультета ВМиК МГУ под руководством профессора И.В.Машечкина, а также на следующих конференциях:

- XIII Международная конференция студентов, аспирантов и молодых учёных «ЛОМОНОСОВ» (Москва, 2006).
- Научная конференция «Ломоносовские чтения» (Москва, 2006).

- Artificial Intelligence: 19th ACS Australian Joint Conference on Artificial Intelligence, Tasmania, Australia, 2006.
- First Spring Young Researches' Colloquium on Software Engineering (SYRCoSE'2007), Moscow, Russia, May 31- June 1, 2007.
- Научная конференция «Тихоновские чтения» (Москва, 2007).
- Вторая международная конференция «Системный анализ и информационные технологии» САИТ-2007, г. Обнинск, Россия, 10-14 сентября 2007.
- Математические методы распознавания образов: 13-я Всероссийская конференция. Ленинградская обл., г. Зеленогорск, 30 сентября – 6 октября 2007.
- УкрПРОГ'2008: шестая международная конференция по программированию, Украина, г. Киев, 27-29 мая 2008.

По теме диссертации автором опубликовано 14 печатных работ, в том числе две – в изданиях, рекомендованных ВАК. Список работ приводится в конце автореферата.

Структура и объём диссертации

Диссертация состоит из введения, четырёх глав, заключения и библиографии. Общий объём диссертации – 103 страницы. Библиография включает 74 наименования.

Краткое содержание диссертации.

Во **введении** кратко охарактеризована задача классификации многотемных гипертекстовых документов, обоснована актуальность задачи, сформулирована постановка задачи и приведён краткий обзор содержания диссертации.

Первая глава посвящена обзору существующих методов классификации многотемных документов, представляющих три основных подхода: оптимизационный подход; подход на основе декомпозиции в набор независимых бинарных проблем и подход на основе ранжирования. Для каждого из подходов описываются основные принципы и особенности, приводятся наиболее известные алгоритмы. Целью обзора является исследование существующих методов с точки зрения их применимости при решении современных прикладных задач.

В **разделе 1.1** сформулирована постановка задачи классификации многотемных документов и требования к решению. В задаче классификации многотемных документов в обучающей совокупности $Z = \{x_i, y_i\}_{i=1}^m$ для каждого примера $x_i \in X$ задано множество релевантных классов $y_i \subseteq \{1,...,q\}$, и целью алгоритма машинного обучения является построение на основе обучающей совокупности классификатора $f_Z : X \to 2^q$, предсказывающего для заданного примера все релевантные классы (X – исходное

пространство признаков, q - число классов). В разделе сформулированы требования к решению данной задачи, связанные как с оценкой качества и скорости классификации, так и с наличием возможности дообучения и добавления (удаления) категорий классификации.

Раздел 1.2 посвящён обзору и сравнительному анализу методов классификации многотемных документов.

К методам классификации многотемных документов, основанным на *оптимизационном подходе*, можно отнести следующие методы: AdaBoost.MH, ADTBoost.MH (минимизируется функция Hamming Loss для оценки потерь multi-label классификации), метод Multi-Label-kNN (максимизируются апостериорные вероятности принадлежности классам) и метод на основе модели смешивания, обученной с помощью метода EM (параметры модели оцениваются на основе принципа максимизации математического ожидания).

Методы, основанные на *декомпозиции multi-label проблемы в набор независимых бинарных проблем* (*«каждый-против-остальных»*), создают одну бинарную проблему для каждого из q классов. В бинарной проблеме для класса $l \in \{1,...,q\}$ все обучающие примеры, помеченные этим классом, считаются положительными, а все остальные обучающие примеры считаются отрицательными. При классификации на основе декомпозиции «каждый-против-остальных» решение о принадлежности объекта конкретному классу принимается независимо от остальных классов. Ввиду такого подхода декомпозиции методы этой группы имеют возможность добавления и удаления классов без необходимости обучения модели классификации «с нуля». На сегодняшний день декомпозиция «каждый-против-остальных» является наиболее популярным подходом при решении задачи классификации многотемных документов в современных практических приложениях. Основным поводом для критики методов этой группы является то, что строятся независимые классификаторы, которые не учитывают корреляции между классами.

Решение задачи multi-label классификации *на основе ранжирования* включает два этапа. Первый этап состоит в обучении алгоритма ранжирования, который упорядочивает все классы по степени их релевантности для заданного классифицируемого объекта. Для ранжирования классов многотемных объектов применяются следующие алгоритмы: MMP,

k-NN, RankSVM. Второй этап заключается в построении пороговой функции [2], отделяющей релевантные классы от нерелевантных.

В **разделе 1.3** сформулированы выводы из проведённого исследования существующих методов классификации многотемных (multi-label) документов. Исследование показало, что обозначенным требованиям к сценариям функционирования программных средств классификации удовлетворяет только подход на основе декомпозиции «каждый-против-оставльных» с применением дообучаемых нейросетевых алгоритмов для осуществления двуклассовой классификации; однако методы, основанные на таком подходе, как правило, обеспечивают качество классификации, недостаточное для их применения в современных прикладных системах.

Вторая глава содержит описание предложенного решения задачи классификации многотемных документов на основе подхода попарных сравнений и результаты экспериментальной оценки характеристик предложенного решения.

В **разделе 2.1** приведена общая структура и обоснование выбранных подходов при разработке предложенного решения. Учитывая требования к методу классификации, связанные с наличием возможности дообучения и добавления (удаления) классов, а также недостатки большинства существующих методов, связанные с ресурсоёмкостью; в качестве базового подхода при разработке метода классификации был выбран подход на основе декомпозиции multi-label проблемы на бинарные подпроблемы. С целью повышения качества классификации, по сравнению с традиционным подходом декомпозиции типа «каждый-против-остальных», в качестве метода декомпозиции была предложена модификация для случая существенно пересекающихся классов метода попарных сравнений с помощью набора бинарных классификаторов (декомпозиция типа «каждый-против-каждого»). Для объединения результатов бинарных классификаторов при декомпозиции типа «каждый-против-каждого» было предложено оценивать степени принадлежности документа классам (ранжирование классов) на основе обобщённой модели Бредли-Терри «с ничьёй»[3]. После оценки степеней принадлежности классам возникает подзадача выделения релевантных для документа классов из отранжированного списка всех классов. Для решения этой подзадачи предложено находить пороговую функцию в пространстве релевантностей классов.

В **разделе 2.2** дан краткий обзор традиционного подхода попарных сравнений для решения задачи классификации, в которой каждый классифицируемый объект может

[2] Elisseeff A., Weston J. A kernel method for multi-labelled classification // Proceedings of the 14th Neural Information Processing Systems (NIPS) Conference, Cambridge, 2002.
[3] T.-K. Huang, R. Weng and C.-J. Lin. A Generalized Bradley-Terry Model: from Group Competition to Individual Skill, Proc. of NIPS'04, 2004.

принадлежать только к одному классу. Для такой задачи классификации при декомпозиции на основе подхода попарных сравнений достигаются лучшие результаты по точности классификации, по сравнению с декомпозицией на основе подхода «каждый-против-остальных».

Раздел 2.3 посвящён описанию предложенного метода ранжирования, основанного на модифицированном для случая существенно пересекающихся классов методе попарных сравнений с помощью набора бинарных классификаторов и вычислении степеней принадлежности документа классам с использованием обобщенной модели Брэдли-Терри с «ничьей»[4].

Рассмотрим задачу ранжирования на основе попарных сравнений для случая существенно перекрывающихся классов. В предложенном методе каждая пара возможно перекрывающихся (и даже вложенных) классов j и k разделяется с помощью *двух бинарных классификаторов*, которые отделяют *пересекающиеся* и *непересекающиеся* области. Используя их, можно выделить четыре области: область "только класса j"; область "только класса k"; перекрывающаяся область (j и k) и область, не принадлежащую ни классу j, ни классу k (рис.1). Для построения двух разделяющих поверхностей для каждой пары классов j и k формулируются две задачи обучения, которые включают только примеры, помеченные в обучающем наборе Z либо j, либо k, либо обоими классами одновременно (число таких примеров, как правило, существенно меньше размера m всего обучающего набора). В первой подзадаче примеры, помеченные только классом k, рассматриваются как "положительные", а все остальные - как "отрицательные". Важно отметить, что при такой формулировке оба бинарных классификатора разделяют взаимно исключающие суперклассы, и для решения и оценки попарных вероятностей могут быть использованы стандартные алгоритмы бинарной классификации[5]. Формулируя и решая таким образом $q(q-1)$ задач бинарной классификации (каждая задача имеет размер меньший, чем m), оцениваем вероятности принадлежности объекта каждой из выделенных областей при попарных сравнениях.

[4] T.-K. Huang, R. Weng and C.-J. Lin. A Generalized Bradley-Terry Model: from Group Competition to Individual Skill // Proc. of NIPS'04, 2004

[5] J. Platt. Probabilistic Outputs for Support Vector Machines and Comparison to Regularized Likelihood Methods // Adv. in Large Margin Classifiers. MIT Press, 1999. pp. 61–74

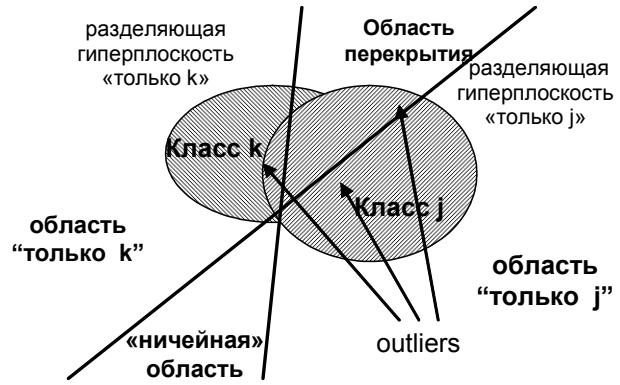

Рисунок 1. Разделение существенно перекрывающихся классов.

Затем попарные вероятности принадлежности, предсказанные всеми бинарными классификаторами, объединяются вместе, чтобы оценить результирующие значения ранжирования классов для документа. Для этого предлагается использовать обобщённую модель ранжирования Бредли-Терри с «ничьёй», которую сформулировали Рао и Купер[6]. Для оценки результирующих релевантностей классов на основе этой модели можно использовать итерационный Minorization-Maximization (MM) алгоритм [7].

Раздел 2.4 посвящён описанию предложенного алгоритма построения пороговой функции для отсечения нерелевантных классов (т.е. для сведения задачи ранжирования к задаче multi-label классификации). Предложено строить пороговую функцию не в исходном пространстве признаков (как это делает большинство традиционных алгоритмов[8]), а в пространстве релевантностей классов, что позволяет упростить вид пороговой функции, уменьшить вычислительную сложность и в большинстве случаев повысить точность. Пространство релевантностей классов – это пространство векторов, координатами которых являются значения степеней принадлежности документа классам (размерность этого пространства равна числу классов и значительно меньше размерности пространства признаков). Предложенная пороговая функции в пространстве релевантностей классов использует результат работы алгоритма ранжирования как новое множество признаков анализируемого электронного документа. Основная мотивация предложенного подхода состоит в упрощении вида решающей функции, сокращении времени обучения, времени классификации и объёма памяти, необходимого для работы алгоритма.

Раздел 2.5 посвящён описанию методики дообучения для разработанного алгоритма многотемной (multi-label) классификации. При дообучении осуществляется

[6] P.V. Rao, L.L. Kupper. Ties in paired-comparison experiments: A generalization of the Bradley–Terry model // Amer. Statist. Assoc, 62, 1967. pp. 194–204

[7] D. R. Hunter. MM-algorithms for generalized Bradley-Terry models // Annals of Statistics, Inst. of Math. Stat., 32 (1), 2004. pp. 384–406

[8] Elisseeff A., Weston J. A kernel method for multi-labelled classification // Proceedings of the 14th Neural Information Processing Systems (NIPS) Conference, Cambridge, 2002

модификация модели классификации (моделей бинарных классификаторов и модели пороговой функции) на основе характеристик новых документов. При дообучении с добавлением новой тематики классификации в модель классификации добавляется набор новых бинарных классификаторов. Для начального обучения бинарных классификаторов в предложенном методе используется метод опорных векторов SVM [9]. Дообучение бинарных классификаторов осуществляется на основе алгоритма персептрона (Kernel Perceptron) [10], который имеет возможность дообучения, но, как правило, обеспечивает менее высокую точность, по сравнению с методом опорных векторов. Поэтому в предложенном методе для повышения качества классификации перед началом процесса дообучения алгоритм персептрона инициализируется коэффициентами, вычисленными на основе обученной методом опорных векторов модели классификации.

Раздел 2.6 посвящён экспериментальному исследованию характеристик предложенного решения. В разделе приведено описание методики проведения экспериментов для оценки методов классификации многотемных документов. Целью экспериментов является анализ и сравнение эффективности разработанного метода ранжирования и классификации многотемных документов с существующими методами. Для экспериментального сравнения эффективности методов использовались следующие критерии: HammingLoss, Coverage и AveragePrecision[11]; статистическая достоверность полученных результатов оценивалась на основе k-раздельного t-теста перекрёстной проверки[12]. Сравнение проводилось с существующими методами, обладающими возможностью дообучения и динамического изменения набора категорий классификации. Как показали результаты экспериментального исследования, на эталонном наборе многотемных документов Reuters-2000[13] разработанный метод обеспечивает более высокое качество классификации по всем перечисленным критериям сравнения.

Третья глава посвящена исследованию и разработке методов построения модели представления гипертекстовых документов. В главе приведён обзор существующих методов решения рассматриваемой задачи, описание предложенной модели представления гипертекстовых документов и результаты экспериментального исследования предложенного решения.

---

[9] J. Platt. Fast Training of Support Vector Machines using Sequential Minimal Optimization, in Advances in Kernel Methods - Support Vector Learning, B. Schölkopf, C. Burges, and A. Smola, eds., MIT Press, 1998

[10] J. Kivinen, A. Smola, R. C. Williamson. Online Learning with kernels. Advances in Neural Information Processing Systems 14, Cambridge, MA: MIT Press, 2002. pp. 785-793

[11] Zhang M.-L., Zhou Z.-H. A k-nearest neighbor based algorithm for multi-label classification // Proceedings of the 1st IEEE International Conference on Granular Computing (GrC'05). Beijing, China, 2005. pp. 718-721

[12] Snedecor G.W., Cochran W.G.. Statistical Methods // 8th ed. Ames, Iowa, Iowa State University Press, 1989.

[13] D.D. Lewis, Y. Yang, T. G. Rose, F. Li. RCV1: A new benchmark collection for text categorization research // Machine Learning, 5, 2004. pp. 361-397

В **разделе 3.1** дана характеристика задачи представления гипертекстовых документов и сформулированы требования к её решению. Модель представления должна удовлетворять следующим требованиям: обучение модели с возможностью дообучения без необходимости хранения обучающего набора; скорость построения модели должна удовлетворять требованиям интерактивного режима; использование данной модели представления должно обеспечивать качество классификации, не уступающее существующим методам.

Раздел 3.2 посвящён обзору и сравнительному анализу методов построения моделей представления гипертекстовых документов. В разделе приводятся критерии сравнения моделей представления и рассматриваются существующие методы выделения признаков в документах (метод ключевых слов со стеммингом [14] и метод N-грамм[15]), меры сходства между документами (частотная[16] и k-spectrum[17]) и метод учёта ссылочной структуры при представлении гипертекстовых документов; обсуждаются основные характеристики и недостатки существующих методов. Существующий метод [18] учёта ссылочной структуры документов, основанный на загрузке и классификации содержимого документов-соседей, позволяет получить более точное (для классификации) представление, по сравнению с учётом только локального текста документа. Однако данный метод имеет высокую вычислительную сложность (в связи с необходимостью загрузки содержимого документов-соседей), что делает его неприменимым для представления документов в интерактивном режиме. Основным же недостатком существующих методов выделения признаков является то, что не учитывается контекст вхождения признаков в текст документа. В данном разделе показана актуальность задачи разработки вычислительно более эффективной модели представления гипертекстовых документов.

В **разделе 3.3** содержится описание разработанной модели представления гипертекстовых документов и приведены результаты экспериментальной оценки характеристик предложенного решения на эталонных наборах данных. Разработанная модель представления гипертекстовых данных является расширением традиционного векторного представления документов. Она учитывает базовые текстовые признаки

[14] Vector Theory and Keyword Weights [электронный ресурс] : выделение признаков из документов / Garcia E. - Режим доступа: http://www.miislita.com/information-retrieval-tutorial/indexing.html
[15] William B. Cavnar, John M. Trenkle. NGram-Based Text Categorization // In Proceedings of SDAIR-94, 3rd Annual Symposium on Document Analysis and Information Retrieval, Las Vegas, US, 1994. pp. 161—175
[16] Vector Theory and Keyword Weights [электронный ресурс] : выделение признаков из документов / Garcia E. - Режим доступа: http://www.miislita.com/information-retrieval-tutorial/indexing.html
[17] H. Lodhi, C. Saunders, J. Shawe-Taylor, N. Cristianini, C. Watkins. Text Classification using String Kernels // Journal of Machine Learning Research, 2, 2002. pp. 419-444
[18] Soumen Chakrabarti, Byron E. Dom, Piotr Indyk. Enhanced hypertext categorization using hyperlinks // Proceedings of the ACM International Conference on Management of Data, SIGMOD, 1998. pp. 307-318

(лексемы или N-граммы), базовые нетекстовые признаки (идентификаторы классов гиперссылок) и составные признаки, являющиеся частыми комбинациями базовых.

В настоящей работе предлагается *метод учёта гиперссылок* при построении модели представления, основанный на том, что гиперссылки анализируются как текстовые строки и на базе этого определяются их классы, которые учитываются при представлении текущего документа. При формировании представления документа, каждую гиперссылку заменяем набором специальных идентификаторов, соответствующих идентификаторам предсказанных классов. Предложенный метод, в отличие от существующего, позволяет учитывать встречающиеся в документах гиперссылки без необходимости получения и анализа содержимого документов-соседей, что очень важно для осуществления классификации в интерактивном режиме. Для преодоления недостатков традиционных подходов, связанных с тем, что не учитывается контекст вхождения признаков, в настоящем исследовании предложен метод построения *модели представления, основанный на выделении частых комбинаций (эпизодов) признаков.* В этом случае множество частых эпизодов лексем (или N-грамм) формирует новое пространство признаков, где каждому эпизоду, в который входит одна или более лексема (или N-грамма), соответствует отдельная координата в пространстве признаков.

В разделе приведено описание *методики проведения экспериментов* для оценки эффективности моделей представления гипертекстовых документов. Для экспериментального сравнения эффективности моделей представления в качестве базового метода классификации был выбран метод k-ближайших соседей (поскольку этот метод наиболее зависим от модели представления); использовались следующие критерии оценки качества: HammingLoss, Coverage и AveragePrecision; оценивалась статистическая достоверность полученных результатов. Экспериментальное исследование моделей представления гипертекстовых документов включало:

- сравнение предложенного метода формирования пространства признаков на основе частых комбинаций базовых признаков (ключевых слов, N-грамм) и традиционного метода, учитывающего только базовые признаки;
- сравнение предложенного метода учёта гиперссылок при представлении документов и метода представления, основанного на учёте только локального текстового содержимого документов.

Как показали результаты экспериментальной оценки на эталонных наборах данных Reuters-2000 и BankResearch[19], предложенная модель представления превосходит

[19] Bank Research Dataset [электронный ресурс]: Набор данных BankResearch. - Режим доступа: http://lib.stat.cmu.edu/datasets/bankresearch.zip

традиционные модели по всем перечисленным критериям. Предложенный метод включения информации о гиперссылках в модель представления позволяет получить улучшение точности практически для любой базовой модели и удовлетворяет требованиям интерактивного режима работы. Результаты экспериментальной оценки эффективности предложенного метода классификации с предложенной моделью представления показали, что разработанное решение обеспечивает более высокое качество классификации по всем основным критериям сравнения.

Четвёртая глава посвящена описанию программного модуля классификации многотемных гипертекстовых документов. Работа модуля основывается на предложенном методе классификации многотемных документов на основе попарных сравнений и предложенной модели представления гипертекстовых документов. Для программной реализации модуля был выбран язык C++. Глава содержит описание архитектуры, сценариев функционирования и результаты экспериментального исследования производительности разработанного программного модуля классификации.

В главе сформулированы требования к программной реализации средств классификации многотемных гипертекстовых документов. Программные средства классификации многотемных гипертекстовых документов должны удовлетворять ряду специфических требований, предъявляемых как к алгоритму классификации и модели представления документов, так и к архитектуре и программной реализации выбранных алгоритмов.

Требования к алгоритму классификации и модели представления документов, главным образом, связаны с тем, что программные средства классификации должны обеспечивать определённые сценарии функционирования: обучение на основе обучающего набора, дообучение с возможностью добавления новых тематик, классификация и удаление темы. Все эти требования были достигнуты при разработке алгоритма классификации многотемных гипертекстовых документов.

Требования к архитектуре и программной реализации модуля связаны преимущественно с тем, что методы классификации многотемных документов достаточно ресурсоёмки, и одна из традиционных проблем применения этих методов на практике – это скорость классификации. Поэтому одним из требований к архитектуре разрабатываемого модуля является масштабируемость и возможность распараллеливания вычислений. Кроме того, методы классификации документов постоянно развиваются и совершенствуются, поэтому архитектура модуля должна обладать свойством расширяемости.

В главе дано описание архитектуры разработанного модуля и сформулированы основные свойства разработанного программного решения. На рисунке 2 представлена общая архитектура разработанного модуля классификации многотемных гипертекстовых документов. Модуль классификации состоит из трёх основных компонент:

- *компонент лексического анализа (парсер)* – осуществляет разбор, выделение признаков и преобразование гипертекстовых документов во внутреннее представление;
- *компонент вычисления меры сходства* – определяет значения близости между документами на основе выданного парсером представления и осуществляет кэширование этих значений;
- *классификатор* – строит дообучаемую модель классификации и на её основе осуществляет классификацию многотемных гипертекстовых документов.

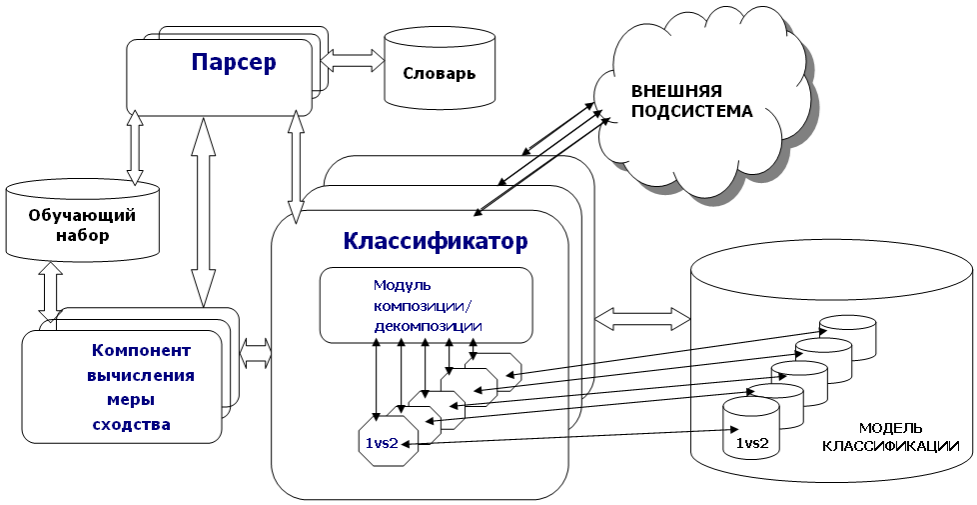

Рисунок 2. Архитектура модуля классификации.

Приводятся результаты экспериментального исследования характеристик разработанного модуля. Для проведения экспериментов по оценке эффективности разработанного модуля использовались обучающие и тестовые поднаборы эталонных наборов BankResearch и Reuters-2000. Как показали результаты экспериментов по оценке производительности разработанного модуля, скорость классификации документов удовлетворяет требованиям интерактивного режима работы.

Разработанный программный модуль классификации многотемных гипертекстовых документов удовлетворяет требованиям современных практических приложений, в которых возникают задачи классификации такого типа.

Заключение содержит основные результаты диссертации.

Основные результаты работы.

1. Разработан новый метод многотемной классификации многотемных документов (на основе попарных сравнений с отсечением нерелевантных классов при помощи пороговой функции), обеспечивающий: возможность дообучения; возможность динамического удаления и добавления классов; более высокое качество классификации, по сравнению с существующими методами.
2. Разработана модель представления гипертекстовых документов, основанная на учёте гиперссылок посредством анализа структуры адресов документов и на расширении традиционной векторной модели представления за счёт добавления частых комбинаций признаков. Разработанное решение позволяет повысить эффективность представления гипертекстовых документов по сравнению с существующими методами.
3. На основе предложенных решений разработан программный модуль классификации многотемных гипертекстовых документов. Разработанный модуль поддерживает следующие основные сценарии функционирования: обучение на основе обучающего набора, классификация документов, дообучение на новых документах без необходимости хранения обучающего набора, добавление и удаление тематик классификации.

Разработанный модуль апробирован в системе анализа и фильтрации Интернет-трафика в рамках Государственного контракта №02.514.11.4026 (федеральная целевая программа «Исследование и разработка по приоритетным направлениям развития научно-технологического комплекса России на 2007-2012 годы»). Система анализа и фильтрации Интернет-трафика зарегистрирована в реестре программ для ЭВМ (свидетельство о государственной регистрации №2008614494).

ПУБЛИКАЦИИ ПО ТЕМЕ ДИССЕРТАЦИИ

1. Глазкова В.В., Петровский М.И. Методы классификации многотемных документов // Сборник тезисов XIII Международной конференции студентов, аспирантов и молодых учёных «ЛОМОНОСОВ», секция ВМиК, 2006, стр. 16-17.
2. Глазкова В.В. Исследование и разработка методов классификации многотемных документов // Сборник тезисов лучших дипломных работ 2006 года, М.: Изд-во факультета ВМиК МГУ, 2006, стр. 75-76.
3. Mikhail Petrovskiy, Valentina Glazkova. Linear Methods for Reduction from Ranking to Multilabel Classification // Springer-Verlag, Lecture Notes in Artificial Intelligence, 2006, vol. 4304, pp. 1152-1156.
4. Глазкова В. В., Петровский М.И. Дообучаемый метод классификации многотемных документов для анализа и фильтрации Интернет информации // Программные

системы и инструменты. Тематический сборник № 7, М.: Изд-во факультета ВМиК МГУ, 2006, стр. 71-82.
5. Глазкова В.В., Петровский М.И. Метод быстрой классификации многотемных текстовых документов // Сборник статей молодых учёных факультета ВМиК МГУ, №3, М., 2006, стр. 55-64.
6. Глазкова В.В., Масляков В.А., Машечкин И.В., Петровский М.И. Интеллектуальная система анализа и фильтрации Интернет-информации // Сборник статей молодых учёных факультета ВМиК МГУ, №4, М., 2007, стр. 18-26.
7. Valentina Glazkova and Mikhail Petrovskiy. Multi-topic Text Categorization Based on Ranking Approach // Proceedings of the First Spring Young Researches' Colloquium on Software Engineering (SYRCoSE'2007), Volume. 1. May 31- June 1, 2007. – Moscow, Russia, pp. 49-55.
8. Valentina Glazkova, Vladimir Maslyakov, Igor Mashechkin and Mikhail Petrovskiy. Internet Traffic Filtering System Based on Data Mining Approach // Proceedings of the First Spring Young Researches' Colloquium on Software Engineering (SYRCoSE'2007), Volume. 1. May 31- June 1, 2007. – Moscow, Russia, pp. 57-62.
9. Петровский М.И., Глазкова В.В. Метод многотемной (multi-label) классификации на основе попарных сравнений с отсечением наименее релевантных классов // Математические методы распознавания образов: 13-я Всероссийская конференция. Ленинградская обл., г. Зеленогорск, 30 сентября – 6 октября 2007 г.: Сборник докладов. – М.: МАКС Пресс, 2007, стр. 197-200.
10. Машечкин И.В., Петровский М.И., Глазкова В.В., Масляков В.А. Концепция построения систем анализа и фильтрации Интернет-трафика на основе методов интеллектуального анализа данных // Математические методы распознавания образов: 13-я Всероссийская конференция. Ленинградская обл., г. Зеленогорск, 30 сентября – 6 октября 2007 г.: Сборник докладов. – М.: МАКС Пресс, 2007, стр. 494-496.
11. Петровский М.И., Глазкова В.В., Царёв Д.В. О выборе модели представления текстовой информации для задачи анализа и фильтрации Интернет-трафика //Математические методы распознавания образов: 13-я Всероссийская конференция. Ленинградская обл., г. Зеленогорск, 30 сентября – 6 октября 2007 г.: Сборник докладов. – М.: МАКС Пресс, 2007, стр. 519-522.
12. Петровский М.И., Глазкова В.В. Алгоритмы машинного обучения для задачи анализа и рубрикации электронных документов. Вычислительные методы и программирование, №8, 2007, стр. 57-69. (www.num-meth.srcc.su/zhurnal/tom8r207.html).
13. Глазкова В.В., Масляков В.А., Машечкин И.В., Петровский М.И. Система фильтрации Интернет-трафика на основе методов data mining // Программные продукты и системы (приложение к международному журналу «Проблемы теории и практики управления»), №2(82), 2008, стр. 22-25.
14. Глазкова В.В., Масляков В.А., Машечкин И.В., Петровский М.И. Система фильтрации Интернет-трафика на основе методов машинного обучения // Вопросы современной науки и практики. Университет им. В.И.Вернадского, серия «Технические науки», ассоциация «Объединённый университет им. В.И.Вернадского, №2(12)/2008, том 2, 2008, стр. 155-168.