На правах рукописи

Астраханцев Никита Александрович

Методы и программные средства извлечения терминов из коллекции текстовых документов предметной области

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

Автореферат
диссертации на соискание учёной степени
кандидата физико-математических наук

Москва — 2014

Работа выполнена в Федеральном государственном бюджетном учреждении науки Институт системного программирования Российской академии наук

Научный руководитель: **Иванников Виктор Петрович**
доктор физико-математических наук,
профессор, академик РАН

Официальные оппоненты: **Соловьев Валерий Дмитриевич,**
доктор физико-математических наук, профессор,
главный научный сотрудник научно-образовательного центра по лингвистике им. И.А. Бодуэна де Куртене Федерального государственного автономного учреждения высшего профессионального образования Казанский (Приволжский) федеральный университет

Браславский Павел Исаакович,
кандидат технических наук, с. н. с. лаборатории комбинаторной алгебры института математики и компьютерных наук Федерального государственного автономного образовательного учреждения высшего профессионального образования Уральский федеральный университет имени первого Президента России Б.Н. Ельцина

Ведущая организация: Федеральное государственное бюджетное учреждение науки Вычислительный центр им. А. А. Дородницына Российской академии наук (ВЦ РАН)

Защита состоится «05» марта 2015 года, в 17 часов на заседании диссертационного совета Д 002.087.01 при Федеральном государственном бюджетном учреждении науки Институт системного программирования Российской академии наук по адресу: 109004, Москва, ул. Александра Солженицина, д. 25.

С диссертацией и авторефератом можно ознакомиться в библиотеке Федерального государственного бюджетного учреждения науки Институт системного программирования Российской академии наук.

Автореферат разослан «03» февраля 2015 года.

Ученый секретарь
диссертационного совета
кандидат физ.-мат. наук Зеленов С. В.

Общая характеристика работы

Актуальность темы. Термин — это слово или словосочетание, обозначающее понятие заданной предметной области. Автоматическое извлечение терминов является важным этапом решения многих задач, связанных с обработкой текстов предметной области. К таким задачам относятся построение глоссариев, тезаурусов или онтологий, информационный поиск, машинный перевод, классификация и кластеризация документов.

К настоящему времени разработано множество методов автоматического извлечения терминов, однако их эффективность остается достаточно низкой: как правило, их точность и полнота не превышают 50%[1]. Кроме того, многие методы требуют размеченных вручную данных, что сужает их практическую применимость.

Одна из причин низкой эффективности методов заключается в том, что они недостаточно полным образом используют возможные источники данных. Большая часть методов ограничивается текстами предметной области, которые зачастую не содержат в себе необходимого объема информации для автоматического извлечения терминов. Некоторые методы используют и внешние ресурсы, такие как корпуса текстов других предметных областей, поисковые машины или созданные экспертами онтологии, однако все эти ресурсы обладают заметными недостатками. Так, созданные вручную онтологии обычно обладают малым объемом и покрывают лишь самые общие понятия предметных областей или только одну предметную область, а текстовые документы, в том числе принадлежащие внешним корпусам или найденные поисковыми машинами, не имеют структуры, которой обладают онтологии, и позволяют использовать только статистическую информацию о встречаемости слов и словосочетаний вне рассматриваемой предметной области.

Указанных недостатков во многом лишена многоязычная интернет-энциклопедия Википедия. Ее статьи описывают понятия реального мира — как универсальные, так и специфичные для узких предметных областей. Она содержит структурную информацию в виде гиперссылок между статьями; обладает очень большим размером и ежедневно пополняется сообществом пользователей.

Вследствие продолжающегося развития Википедии и, в частности, все большего покрытия ею предметных областей, использование ее в настоящее время может значительно повысить эффективность методов автоматического извлечения терминов.

[1]Определения точности и полноты приводятся в разделе 1.4 диссертационной работы.

Целью настоящей диссертационной работы является разработка методов и программных средств извлечения терминов из коллекции текстовых документов предметной области с использованием структуры гиперссылок Википедии.

Разрабатываемые методы должны обладать следующими свойствами:

1. полная автоматичность, в том числе отсутствие требований к наличию размеченных вручную данных;
2. точность и полнота выше соответствующих показателей существующих методов извлечения терминов

Для достижения цели были поставлены и решены следующие задачи:

1. Исследовать существующие методы извлечения терминов.
2. Разработать метод автоматического извлечения терминов, использующий структуру гиперссылок Википедии.
3. Реализовать разработанный метод в виде программной системы и провести экспериментальное исследование его применения с целью определения эффективности разработанного метода.

Основные положения, выносимые на защиту:

1. Предложен подход к использованию информации Википедии для задачи извлечения терминов, основанный на структуре гиперссылок Википедии.
2. Предложен подход к извлечению терминов на основе алгоритма частичного обучения, не требующий размеченных данных.
3. В рамках предложенных подходов разработан метод автоматического извлечения терминов.
4. Разработана программная система извлечения терминов и проведено экспериментальное исследование, доказывающее повышение эффективности разработанного метода по сравнению с существующими методами.

Научная новизна. В настоящей работе предложен новый метод извлечения терминов из коллекции текстов предметной области, основанный на алгоритме частичного обучения и использовании структурной информации Википедии. Математически доказана оценка вычислительной сложности разработанного метода. Экспериментально подтверждено повышение эффективности разработанного метода по сравнению с существующими методами.

Разработанный метод не зависит от предметной области, не требует размеченных вручную данных, может применяться в различных задачах обработки текстов предметной области.

Практическая значимость. Предложенный подход к извлечению терминов и разработанные в его рамках методы могут быть использованы при решении прикладных задач автоматической и полуавтоматической обработки текстов, в том числе информационного поиска, определения ключевых фраз, классификации и кластеризации документов, машинного перевода, построения и обогащения словарей, тезаурусов, онтологий. Созданная на основе разработанного метода программная система была включена в систему Texterra, разрабатываемую в Институте системного программирования РАН.

Апробация работы. Основные результаты работы докладывались на следующих конференциях и семинарах:

- на десятом весеннем коллоквиуме молодых исследователей в области баз данных и информационных систем (SYRCoDIS) (2013г.);
- на сто шестьдесят первом заседании Московской Секции ACM SIGMOD (2013г.);
- на двадцатой Международной конференции по компьютерной лингвистике «Диалог» (2014г.);
- на научном семинаре «Управление данными и информационные системы» Института системного программирования РАН (2014г.)
- на научном семинаре «Интернет, распределенные информационные системы и цифровые библиотеки» ВЦ РАН (2014г.)

Личный вклад. Автором проведено исследование предметной области и существующих методов, разработаны все описанные в диссертации методы, подготовлена спецификация для программной системы на основе разработанных методов, проведено экспериментальное исследование. Программная система разработана совместно с Д.Г. Федоренко.

Публикации. Основные результаты по теме диссертации изложены в 6 печатных работах [1-6], 4 из которых опубликованы в журналах, рекомендованных ВАК [1-4].

Объем и структура работы. Диссертация состоит из введения, четырех глав, заключения и двух приложений. Полный объем диссертации составляет 133 страниц с 26 рисунками и 14 таблицами. Объем приложений составляет 15 страниц. Список литературы содержит 117 наименований.

Содержание работы

Во **введении** обосновывается актуальность исследований, проводимых в рамках данной диссертационной работы, формулируются цель, постановка задачи, научная новизна и практическая значимость работы.

Первая глава содержит обзор предметной области, существующих методов извлечения терминов и методологии оценки эффективности.

Как отмечается многими исследователями в области лингвистики (классической и компьютерной), в настоящее время не существует общепринятого универсального определения понятия «термин». Тем не менее, в большинстве работ по автоматическому извлечению терминов используется следующее определение, которое принято и в данной работе: *термин* — это слово или словосочетание, обозначающее понятие заданной предметной области.

Можно выделить две постановки задачи, или два сценария, извлечения терминов:

1. для каждого слова и словосочетания (уникального с точностью до лемматизации) определить, является ли оно термином предметной области;
2. для каждого вхождения слова и словосочетания определить, является ли оно термином предметной области: при этом разные вхождения одного и того же словосочетания могут быть по-разному классифицированы в термины и не-термины в зависимости от контекста.

В данной работе рассматривается первая постановка задачи. Она подразумевает следующую методологию оценки эффективности: для коллекции текстов предметной области формируется эталонное множество терминов, с которым и сравнивается результат работы системы извлечения терминов с помощью стандартных метрик полноты, точности и средней точности. Следует отдельно отметить, что на практике зачастую невозможно получить эталонное множество терминов, в точности соответствующее коллекции текстов. В таких случаях точность и полнота оцениваются приближенно.Многие работы также ограничиваются извлечением только однословных, или только двухсловных, или только многословных терминов; в данной работе рассматриваются термины любой длины.

Методы извлечения терминов, как правило, состоят из трех последовательных этапов:

1. **Сбор кандидатов**: фильтрация слов и словосочетаний, извлеченных из коллекции документов, по статистическим и лингвистическим критериям.
2. **Подсчет признаков**: перевод каждого кандидата в вектор признакового пространства.

3. **Вывод на основе признаков**: классификация кандидатов на термины и не термины либо сортировка всех кандидатов по вероятности быть термином и взятие заранее определенного числа кандидатов.

Методы сбора кандидатов, в свою очередь, также состоят из нескольких шагов. На первом шаге применяются лингвистические фильтры, цель которых — оставить только существительные и именные группы, то есть словосочетания с существительным в роли главного слов. Для этого применяется поверхностный синтаксический разбор (shallow parsing, chunking) или фильтрация N-грамм по шаблонам частей речи. На последующих шагах сбора кандидатов с целью снижения шума производится дополнительная фильтрация по частоте либо содержанию стоп-слов из заранее составленного списка.

Большинство *признаков для извлечения терминов* основано на частоте вхождения кандидатов в рассматриваемую коллекцию текстовых документов. К таковым относятся, например, частота вхождений термина (Term Frequency, TF), TF-IDF, Domain Consensus, C-Value. Одними из первых методов извлечения многословных терминов можно считать меры ассоциации, измеряющие, насколько случайно совместное появление слов в составе термина: взаимная информация (Mutual Information, MI), критерии Стьюдента (TTest), хи-квадрат, логарифмическое правдоподобие (Loglikelihood Ratio), LexicalCohesion и др. В отдельную подгруппу можно вынести методы на основе тематического моделирования: Term Score, Maximum Term Frequency, Novel Topic Model и др.

Некоторые признаки также учитывают контекст вхождений, например NC-Value и PostRankDC (или DomainModel). В других признаках — Weirdness, Domain Pertinence, Domain Relevance, Relevance — используется частота вхождений во внешнюю коллекцию документов, не принадлежащую какой-либо определенной предметной области, например корпус новостей или художественной литературы. Иногда для извлечения терминов — как правило, двухсловных, реже многословных — применяются и другие внешние ресурсы, такие как поисковые машины интернета или существующие тезаурусы.

В последние годы стали появляться методы, основанные на интернет-энциклопедии Википедия. Как правило, они используют алгоритмы поиска путей в графе категорий или случайного блуждания по этому графу и требуют вручную выбрать несколько категорий, которые соответствуют интересующей предметной области. При этом большая часть этих методов не использует коллекцию документов предметной области, опираясь исключительно на информацию Википедии; исключением можно назвать работу Вивальди и др.[2], в которой свойства путей в графе категорий (количество и длина путей) применяется для

[2]Vivaldi J. et al. Using Wikipedia to Validate the Terminology found in a Corpus of Basic Textbooks //LREC. 2012. P. 3820-3827.

проверки терминов, определенных с помощью другого метода, однако здесь также требуется вручную задать категории Википедии, описывающие предметную область.

На этапе *вывода на основе признаков*, в случае использования нескольких признаков, возникает задача преобразования вектора признаков в число, показывающее уверенность метода в том, что данный кандидат является термином.

Наиболее простым способом является линейная комбинация, применяемая, например, в методе TermExtractor. Также используется метод на основе алгоритма голосования:

$$V(t) = \sum_{i=1}^{n} \frac{1}{r(f_i(t))}$$

где t – кандидат в термины, n – количество признаков, $r(f_i(t))$ – позиция кандидата t среди всех кандидатов, отсортированных по значению признака $f_i(t)$. Данный метод не требует нормализации признаков и показывает в среднем лучшие результаты.

При наличии размеченных данных становится возможным применять алгоритмы машинного обучения с учителем, в частности AdaBoost, Ripper, метод опорных векторов (SVM). Как было показано в работе [2], классификаторы на основе машинного обучения достигают лучшей средней точности.

В завершение обзора делается вывод, что существующие методы извлечения терминов недостаточно полным образом используют структурированные внешние ресурсы, в частности — интернет-энциклопедию Википедию.

Во **второй главе** приводится обзор Википедии и ее структурных элементов, полезных для задач обработки текстов, а также приводятся описание и экспериментальное исследование предложенных в работе двух новых методов извлечения терминов: «Вероятность быть гиперссылкой» и «Близость к ключевым концептам».

Метод «Вероятность быть гиперссылкой» (LinkProbability) представляет собой нормализованную частоту, с которой кандидат в термины является гиперссылкой в статьях Википедии:

$$LinkProb_T(t) = \begin{cases} 0, & \text{если } t \text{ не содержится в Википедии или } \frac{H(t)}{W(t)} < T; \\ \frac{H(t)}{W(t)}, & \text{иначе,} \end{cases}$$

где t — кандидат в термины (слово или словосочетание, прошедшее фильтрацию по частям речи, частоте и списку стоп-слов); $H(t)$ показывает, сколько раз кандидат t встречался в статьях Википедии в виде названия гиперссылки (hyperlink

caption), $W(t)$ — сколько раз t встречался в статьях Википедии всего T — параметр метода, предназначенный для фильтрации слишком малых значений, так как они, как правило, являются ошибками разметки. Экспериментально было подобрано значение $T = 0.018$.

Значение этого признака будет близко к нулю для слов и словосочетаний, являющихся частью общей лексики, то есть не принадлежащих какой-либо предметной области. Таким образом, мотивация использования этого метода заключается в фильтрации таких слов и словосочетаний, поскольку они, скорее всего, не принадлежат и к предметной области, для которой извлекаются термины.

Например, слово *Gene* (ген) встречается 85972 раза в статьях Википедии и 14569 раза в виде гиперссылки на статью, описывающую соответствующее понятие. Таким образом, значение признака составит 0.16946. Для словосочетания *Last card* (последняя карта) значение признака равняется 0.012 (332 раза в статьях и всего лишь 4 раза в виде гиперссылки на статью про карточную игру с одноименным названием). Слово *Size* (размер) встречается всего 261607 раз, из которых 58 раз в виде гиперссылки, что приводит к значению 0,0002. Таким образом, при прочих равных условиях вероятность быть отнесенным к терминам для слово *Gene* выше, чем для словосочетания *Last card*, которое, в свою очередь, является более вероятным термином, чем *Size*.

Метод «Близость к ключевым концептам» (KeyConceptRelatedness) основан на следующей интерпретации определения термина: «Термин — слово или словосочетание, обозначающее *понятие*, которое *принадлежит* заданной *предметной области*», где «понятие» интерпретируется как концепт, присутствующий в Википедии в виде статьи; «предметная область» — набор близких по смыслу понятий; «принадлежность понятия к предметной области» — близость по смыслу понятия к предметной области; «близость по смыслу» — семантическая близость, которая представляет собой функцию, определенную для любой пары концептов и имеющую значения от 0 до 1: чем ближе значение функции к 1, тем больше общего между концептами. В данной работе семантическая близость вычисляется на основе Википедии по модифицированной формуле Дайса, где соседними считаются статьи, связанные хотя бы одной гиперссылкой, и при этом учитывается тип гиперссылок.

Кроме того, фраза «обозначающее понятие» интерпретируется как «обозначающее хотя бы одно понятие»: это позволяет решить проблему лексической многозначности кандидата в термины, то есть ситуации, когда кандидат в термины может иметь несколько значений (концептов Википедии), путем выбора максимально близкого понятия из обозначаемых термином. Заметим, что такая

интерпретация применима только для сценария, не различающего различные вхождения кандидатов в термины.

При вычислении данного признака в качестве набора понятий, образующих собой предметную область в смысле приведенной выше интерпретации, предлагается использовать ключевые концепты входной коллекции текстовых документов. Более точно, алгоритм метода следующий:

1. Определить ключевые концепты на основе заданной коллекции документов.
2. Для рассматриваемого кандидата в термины: найти все возможные концепты Википедии, такие что их названия совпадают с кандидатом в термины.
3. Для каждого найденного концепта кандидата в термины: вычислить семантическую близость к найденным ключевым концептам.
4. Выбрать максимальное значение по всем концептам кандидата в термины.

Таким образом, значение этого признака будет близко к нулю для слов и словосочетаний, обозначающих понятия, далекие по смыслу от ключевых понятий предметной области.

Для определения ключевых концептов на основе заданной коллекции документов предлагается следующий эвристический алгоритм: извлечь d ключевых концептов из каждого документа коллекции; посчитать *встречаемость* ключевых концептов: сколько раз в коллекции каждый концепт попал в число (лучших d) ключевых концептов документа; взять N ключевых концептов с наибольшей встречаемостью. Для извлечения ключевых концептов из документа используется метод KPMiner.

Семантическую близость между понятием, обозначаемым термином, и набором ключевых понятий, извлеченных из коллекции текстов, предлагается вычислять с помощью взвешенного варианта метода k ближайших соседей (k Nearest Neighbors, kNN), адаптированного для случая только положительных примеров:

$$sim_k(c, C_N) = \frac{1}{k} \sum_{i=1}^{k} sim(c, c_i)$$

где c — концепт термина; C_N — множество ключевых концептов, отсортированных по убыванию семантической близости к c; $sim(c, c_i)$ — функция семантической близости; k — константа, определяющее число ближайших концептов, которые учитываются при вычислении итоговой семантической близости.

В ходе экспериментального исследования были выбраны следующие значения параметров: $N = 200$, $d = 3$ (при наличии большого количества документов, иначе использовать бóльшие значения d), $k = 10$.

Рассмотрим в качестве примера снова *Gene* и *Last card* применительно к предметной области «Настольные игры». Допустим, $k = 2$ и из коллекции текстов про настольные игры были извлечены следующие ключевые концепты:

1. Board game (собственно, Настольная игра)
2. Card game (Карточная игра)
3. Hasbro Inc. (Компания, производящая игрушки и настольные игры)

Как уже упоминалось выше, в Википедии существует статья *Last card* про одноименную игру; значения семантической близости этого концепта с указанными в списке составляют 0.001, 0.037 и 0, соответственно. Таким образом, значение признака составляет 0.019. В то же время для термина *Gene* (Ген) семантическая близость к указанным концептам будет равна нулю и, таким образом, термин Last card является более вероятным термином предметной области «Настольные игры» с точки зрения метода «Близость к ключевым концептам».

В конце второй главы описывается экспериментальное исследование. Оно проводилось на следующих открытых наборах данных: GENIA, Krapivin, FAO, Patents и Board game.

GENIA представляет собой коллекцию из размеченных документов биомедицинской тематики, она является одним из наиболее популярных наборов данных для исследования эффективности извлечения терминов. FAO состоит из размеченных вручную отчетов Продовольственной и сельскохозяйственной организации ООН (Food and Agriculture Organization): в каждом отчете выделялось по 2 термина. Krapivin представляет собой научные статьи по информатике; в качестве эталонного множества терминов используются ключевые слова, выделенные авторами статей. При тестировании в данной работе к этому множеству был добавлен словарь предметной области «Вычислительная техника» (Computing), использованный в качестве эталона в системе Protodog. Patents — набор из патентов в области электротехники, размеченных вручную. Набор данных Boardgame — коллекция из описаний и рецензий настольных игр — был подготовлен в рамках данной работы. Часть документов были размечены вручную, и для тестирования использовались только термины, имеющие хотя бы одно вхождение в размеченные документы.

В таблице 1 показана статистика выбранных наборов данных.

Кандидаты для всех оцениваемых методов представляли собой N-граммы (от 1 до 4) с фильтрацией по шаблонам частей речи, по частоте (минимальный порог встречаемости для Board game, Patents и GENIA — 2; для Krapivin и FAO — 3) и по списку стоп-слов.

Результатом работы каждого оцениваемого метода является список кандидатов, отсортированный по вероятности быть термином.

Таблица 1: Статистика наборов данных

	Board game	Patents	GENIA	Krapivin	FAO
Предметная область	Настольные игры	Электротехника	Биомедицина	Информатика	Сельское хозяйство
Документов (размеченных)	1300 (35)	16	2000	2304	778
Слов, тыс.	612	120	484	20566	26364
Слов (на документ)	470.9	7493.0	242.2	8926.3	33887.3
Терминов-эталонов	527	1556	30423	8692	1556

Эффективность оценивалась с помощью средней точности:

$$AvP(N) = \sum_{i=1}^{N} P(i)(R(i) - R(i-1)),$$

где N – общее количество оцениваемых кандидатов, или длина верхней части списка отсортированных кандидатов, $P(i)$ – точность для i лучших кандидатов, $R(i)$ – полнота для i лучших кандидатов. Данная метрика широко используется при оценке эффективности извлечения терминов, поскольку позволяет учитывать точность для всех срезов финального списка терминов. Для всех наборов данных, кроме GENIA, число N равнялось числу терминов в списке эталонов, для GENIA N было выбрано равным 10000, так как итоговый список кандидатов после всех фильтров насчитывает всего 15 тысяч терминов, что в 2 раза меньше размера списка эталонов.

В таблице 2 представлены краткие результаты экспериментального исследования отдельных методов.

	Board game	Patents	GENIA	Krapivin	FAO
Domain Model	0.3767	0.4595	0.6729	**0.4526**	0.3397
C-Value	0.3350	**0.6271**	0.7294	0.3706	**0.3731**
TermExtractor	0.3526	0.4974	**0.7331**	0.3209	0.1543
Novel Topic Model	0.3419	0.5432	0.7151	0.1143	0.0723
Wikipedia Categories (NC)	**0.4062**	0.2934	0.7157	-	-
LinkProbability	0.4483	0.4514	0.7287	0.1816	0.0169
KeyConceptRelatedness	**0.5550**	0.5147	0.7247	0.2283	0.1081

Таблица 2: Сравнение средней точности существующих методов и разработанных методов

Как видно из таблицы, на трех наборах данных разработанные методы показывают результаты, сравнимые с показателями существующих методов, однако на остальных наборах данных эффективность разработанных методов оказывается неудовлетворительной.

В **третьей главе** описывается новый метод извлечения терминов, основанный на алгоритме частичного обучения и не требующий размеченных данных, а также приводится экспериментальное исследование этого метода.

В первом разделе описывается общая идея и формализация нового метода. Как известно, для эффективного извлечения терминов необходим учет множества факторов, однако существующие методы комбинации признаков либо слишком простые и не учитывают скрытые взаимосвязи признаков (линейная комбинация или алгоритм голосования), либо требуют большого числа размеченных данных (алгоритмы машинного обучения с учителем). Одновременно с этим, многие существующие методы извлечения терминов обладают высокой точностью для небольшого числа терминов (50-150 лучших терминов), которая к тому же может быть дополнительно повышена за счет фильтрации терминов, отсутствующих в Википедии.

Разработанный метод основан на данных наблюдениях и состоит в следующем: с помощью специального метода извлечения терминов определяются S лучших кандидатов, которые затем используются как положительные примеры для построения модели алгоритма обучения на основе положительных и неразмеченных примеров (Positive-unlabeled learning, PU-learning) — частного случая алгоритма частичного обучения (Semi-supervised learning). В данном случае неразмеченными примерами служат все остальные кандидаты в термины. Построенная модель алгоритма обучения далее используется для вероятностной классификации каждого кандидата в термины.

Для формализации метода вводятся следующие понятия:

$W = \{w_i\}$ — все слова и словосочетания естественного языка.

$y_d : W \to [0, 1]$ — функция, оценивающая вероятность того, что слово или словосочетание w_i является термином заданной предметной области d. Имеется в виду «истинная» оценка без учета осуществимости ее получения, например такой оценкой для определенного w_i может быть число, относительно которого согласны все возможные пользователи приложения, для которого извлекаются термины.

e_d — экспертная оценка значения $y_d(w)$ для каждого w_i. В данной работе, как и в большинстве существующих работ, $e_d(w)$ представляет собой булеву функцию ($e_d : W \to \{0, 1\}$), возвращающую 1 для тех и только тех слов и словосочетаний, которые принадлежат заранее определенному экспертами списку правильных терминов-эталонов, однако возможна и более точная

оценка ($e_d : W \to [0,1]$), например, с помощью усреднения решений нескольких экспертов по данному слову или словосочетанию.

В любом случае, удобно представить $e_d(w)$ в виде $e_d(w) = y_d(w) + \varepsilon_d$, где ε_d — разного рода ошибки экспертов.

$X_{T,K} = \{x_i\}, X_{T,K} \subset W$ — кандидаты в термины, отобранные из заданной коллекции текстовых документов T с помощью заданного метода извлечения кандидатов K.

$f : X_{T,K} \to [0,1]$ — искомая функция, оценивающая вероятность того, что кандидат x_i является термином заданной предметной области. Поскольку функцию y_d невозможно получить на практике, оценка эффективности функции f производится с помощью экспертной оценки e_d.

В разработанном методе предлагается ввести функцию $s : X_{T,K} \to \{u, 1\}$ — метод извлечения терминов для их использования в качестве положительных примеров, где значение 1 соответствует положительному примеру, а u — неразмеченному. Функция $f(x)$ строится на результатах $s(x)$ с помощью алгоритма PU-learning.

Второй раздел третьей главы посвящен специальному методу извлечения терминов, которые будут использоваться в качестве положительных примеров. Формулируются следующие требования к такому методу:

1. Высокая точность: ошибочное отнесение кандидата к положительным примерам приведет к снижению точности функции $f(x)$.

2. Настраиваемая специфичность: как отмечалось выше, для разных приложений может требоваться разный уровень специфичности извлекаемых терминов; поскольку положительные примеры служат для обучения модели извлечения терминов, на практике полезно иметь возможность корректировать специфичность извлекаемых терминов.

Для проверки второго требования в данной диссертационной работе вводится следующее определение: термин t_1 называется более специфичным, чем термин t_2, если термин t_1 является гипонимом или меронимом[3] термина t_2.

На основе работы А. Хипсли[4] делается вывод, что если термин t_1 образован от термина t_2 путем добавления слов-модификаторов, то термин t_1 является более специфичным в смысле принятого определения, чем термин t_2.

На практике важен и обратный вопрос: если термин t_1 является более специфичным в смысле принятого определения, чем термин t_2, какова вероятность, что t_1 образован от термина t_2 путем добавления модификаторов? Для

[3]Отношение «часть-целое» (part-of, substance of, member-of).

[4]Hippisley A., Cheng D., Ahmad K. The head-modifier principle and multilingual term extraction // Natural Language Engineering. 2005. Vol. 11, no. 02. P. 129–157.

удобства назовем *коэффициентом терминообразования* эту вероятность и предположим, что коэффициент терминообразования является характеристикой предметной области, причем для технических и развивающихся предметных областей — именно тех, для которых обычно требуется автоматическое извлечение терминов — он представляет собой достаточно большую величину.

Далее описывается метод извлечения терминов-положительных примеров — собственный метод ComboBasic.

ComboBasic представляет собой модификацию метода Basic, являющегося частью метода Domain Model, который, в свою очередь, является модификацией широко распространенного метода C-Value. Метод Basic вычисляется по формуле:

$$Basic(x) = |x| \log f(x) + \alpha e_x,$$

где x — кандидат в термины, $|x|$ — длина кандидата x (в словах), $f(x)$ — частота вхождений x в коллекции текстов, e_x — количество кандидатов, содержащих кандидата x.

В методе ComboBasic вводится дополнительное слагаемое:

$$ComboBasic(x) = |x| \log f(x) + \alpha e_x + \beta e'_x,$$

где e'_x обозначает число кандидатов, *содержащихся* в кандидате x.

Дополнительно рассматривается и модификация метода ComboBasic, в которой производится фильтрация с помощью Википедии: среди результатов оставляются только те термины, которые присутствуют в Википедии в виде названий статей или текстов гиперссылок на статьи. Такая фильтрация позволяет значительно повысить точность, особенно в случае наборов данных небольшого объема или с большим покрытием Википедией.

Параметры α и β были выбраны экспериментально и равняются, соответственно, 0.15 и 3 при наличии фильтрации по Википедии, и 1 и 0.1 иначе.

Для таким образом введенного метода доказывается следующая теорема.

Теорема 1. *Пусть d — коллекция документов из предметной области, такой что ее коэффициент терминообразования равен γ;*
t_1 и t_2 — кандидаты в термины, извлеченные из коллекции документов d, причем t_1 более специфичен, чем t_2;
$C_{\alpha,\beta}(t)$ — значение признака ComboBasic с коэффициентами α и β для кандидата t, посчитанное на основе коллекции d;
$\alpha, \beta_1, \beta_2 \in R^+$, причем $\beta_1 < \beta_2$.

Тогда $P(C_{\alpha,\beta_2}(t_1) - C_{\alpha,\beta_1}(t_1) > C_{\alpha,\beta_2}(t_2) - C_{\alpha,\beta_1}(t_2) \geq 0) = \gamma$.

Из доказанной теоремы следует, что с вероятностью γ увеличение параметра β в методе ComboBasic при прочих равных условиях приводит к извлечению им более специфичных терминов.

Таким образом, если предположить, что в заданной предметной области большая часть специфичных терминов образуется путем добавления модификаторов (коэффициент терминообразования больше 0.5), то, увеличивая параметр β в методе ComboBasic, можно извлекать более специфичные термины.

В третьем разделе этой главы приводится краткий обзор существующих алгоритмов обучения на основе положительных и неразмеченных примеров и описываются выбранные алгоритм и признаки.

К настоящему времени разработано множество алгоритмов обучения на основе положительных и неразмеченных примеров; в данной работе рассмотрены наиболее эффективные из них: Traditional PU learning, Gradual Reduction, Spy-EM и Pairwise Ranking SVM. По результатам экспериментального исследования был выбран метод Traditional.

В большинстве алгоритмов обучения на основе положительных и неразмеченных примеров, в том числе в алгоритме Traditional, итеративно вызывается алгоритм обучения с учителем и в качестве итоговой модели классификации используется модель алгоритма обучения с учителем. Таким образом, возникает задача выбора алгоритма обучения с учителем, для чего формулируются следующие требования:

1. Высокая эффективность вероятностной классификации
2. Высокая эффективность при малом числе признаков
3. Высокая эффективность при относительно малом объеме данных для обучения и перекошенной выборке

На основе данных требований были выбраны логистическая регрессия, наивный байесовский классификатор и Random forest. Экспериментальное исследование показало, что наиболее стабильной — и в большинстве случаев лучшей — эффективностью обладает логистическая регрессия.

Как отмечалось выше, для автоматического извлечения терминов необходимо учитывать множество факторов, то есть признаков для алгоритма обучения; с другой стороны, увеличение количества признаков может привести к повышению корреляции между ними, в то время как вероятностная классификация более эффективна при низкой корреляции между признаками.

Решение этого противоречия основано на предположении, что признаки из разных категорий (например, на основе тематических моделях и на основе контекстов вхождений) обладают меньшей корреляцией. С учетом этого, были выбраны лучшие признаки из каждой категории, поддерживающие работу с терминами любой длины:

1. C-Value в модификации для поддержки однословных терминов — признак на основе частот вхождений.
2. Novel Topic Model — признак на основе тематических моделей.
3. Relevance — признак на основе анализа внешней коллекции документов.
4. Domain Model — признак на основе контекстов вхождений.
5. Близость к ключевым концептам — признак на основе Википедии, учитывающий близость к предметной области.
6. Вероятность быть гиперссылкой — признак на основе Википедии, учитывающий специфичность термина.

Четвертый раздел третьей главы посвящен экспериментальному исследованию разработанного метода. Оно проводилось при тех же условиях, что и в предыдущей главе. В методах на основе положительных и неразмеченных примеров число лучших кандидатов S, используемых как положительные примеры для обучения, равнялось 100.

C-Value (mod) обозначает модификацию метода C-Value для поддержки однословных терминов. Traditional LR — предложенный метод без фильтрации положительных примеров по Википедии; Traditional LR Wiki — тот же метод, но с фильтрацией.

Результаты экспериментального исследования представлены в таблице 3.

Таблица 3: Сравнение разработанного подхода с существующими методами

Метод	Board games	Patents	GENIA	Krapivin	FAO
TF-IDF	0.3882	0.4922	0.6936	0.3619	**0.4270**
C-Value	0.3350	**0.6271**	0.7294	0.3706	0.3731
C-Value (mod)	0.3967	0.5874	0.7376	0.3911	0.4080
Domain Relevance	0.3253	0.4943	**0.7425**	0.3218	0.1534
Domain Model	0.3821	0.4594	0.6729	**0.4609**	0.3447
Novel Topic Model	0.3684	0.5426	0.7129	0.1271	0.0593
Wiki Categories (NC)	**0.4110**	0.2934	0.7157	0.1671	0.0765
LinkProbability	0.4482	0.4522	0.7276	0.1815	0.0169
KeyConceptRelatedness	**0.5470**	0.5237	0.7253	0.2282	0.1089
Traditional LR Wiki	**0.5925**	0.6161	**0.7745**	**0.5128**	**0.5117**
Traditional LR	**0.4756**	**0.6321**	**0.7645**	**0.4957**	**0.4742**

Как видно из таблицы, разработанный метод без фильтрации по Википедии показывает лучшие результаты на всех наборах данных, с фильтрацией —

на четырех из пяти (кроме Patents), причем для этих четырех результаты значительно лучше по сравнению с методом без фильтрации.

Превосходство разработанного метода над лучшим из существующих методов, позволяющих извлечение терминов любой длины, на всех наборах данных подтверждено статистическим критерием знаковых рангов Уилкоксона (p-value = 0.05). Для получения набора выборок, на которых происходила проверка статистической гипотезы, использовался метод расщепления выборки (Jackknife resampling): набор разбивался случайным образом на N частей, на $N-1$ частях проводилось N тестирований.

Четвертая глава посвящена разработанной программной системе: описывается архитектура и приводятся оценки вычислительной сложности.

Разработанная программная система состоит из следующих модулей:

1. Модуль обработки документов, производящий чтение документов, а также — с помощью системы Текстерра — разбиение их на предложения и токены, определение частей речи и ключевых концептов.

2. Модуль извлечения кандидатов в термины, собирающий N-граммы из токенов и фильтрующий их по шаблонам частей речи, частоте вхождений и наличию стоп-слов.

3. Модуль обработки терминов, вычисляющий признаки и производящий вероятностную классификацию.

4. Модуль фильтрации терминов, производящий отбор кандидатов на основе значений, полученных на этапе обработки терминов.

5. Модуль оценки эффективности извлечения терминов, подсчитывающий метрики точности, полноты и средней точности.

Важной особенностью разработанной архитектуры является кэширование значений признаков для кандидатов. Это позволяет, в частности, производить отбор признаков путем полного перебора, поскольку подсчет каждого признака будет произведен только один раз, а именно этап вычисления признаков обладает наибольшей вычислительной сложностью.

Кроме того, в случае использования системы извлечения терминов в рамках практического приложения, поддерживающего собственную быструю оценку эффективности, становится возможным выбрать наилучшие комбинации признаков и значений параметров. Например, если извлечение терминов применяется для последующего определения ключевых фраз и существует набор документов с эталонным множеством ключевых фраз, то можно перебрать несколько

тысяч комбинаций признаков и значений параметров метода извлечения терминов и выбрать ту комбинацию, на которой достигается максимальная эффективность.

В конце четвертой главы оценивается вычислительная сложность разработанного метода и доказывается следующая теорема.

Теорема 2. ***Временная сложность*** *разработанного метода на основе алгоритма PU-learning составляет:*

$$O(n(log(n) + n_{lt} + D_{avg} + n_c + m) + D(V_{avg} + log(D)) + w_f(m + log(w_f)) + v_{all}),$$

где n — число терминов (кандидатов в термины); n_{lt} — среднее число терминов, содержащих другие термины; D_{avg} — среднее число документов, в которых встречается один и тот же термин; n_c — среднее количество концептов термина; m — среднее число вхождений терминов; V_{avg} — среднее число различных слов в документе после фильтров по частоте и стоп-словам (для построения тематической модели); D — число документов входной коллекции; w_f —число слов после фильтров по частям речи и распространенности (для построения модели домена); v_{all} — общее число слов (токенов) во всех документах.

Пространственная сложность *разработанного метода на основе алгоритма PU-learning составляет:*

$$O(n(n_{lt} + D_{avg}) + V + D + w_f),$$

где V — размер словаря, или общее число разных слов.

В **заключении** приведены основные результаты работы:

1. Предложен подход к использованию информации Википедии для задачи извлечения терминов, основанный на структуре гиперссылок Википедии.
2. Предложен подход к извлечению терминов на основе алгоритма частичного обучения, не требующий размеченных данных.
3. В рамках предложенных подходов разработан метод автоматического извлечения терминов.
4. Разработана программная система и проведено экспериментальное исследование, доказывающее повышение эффективности разработанного метода по сравнению с существующими методами.

Публикации автора по теме диссертации

1. Астраханцев Н., Турдаков Д.. Методы автоматического построение и обогащения неформальных онтологий // Программирование. 2013. Т. 39, № 1. С. 23-34.

   Автору принадлежит основополагающий вклад: проведен анализ существующих работ и написаны введение и основной текст статьи, заключение написано совместно с Д. Турдаковым.

2. Федоренко Д., Астраханцев Н.. Автоматическое извлечение новых концептов предметно-специфичных терминов // Труды Института системного программирования РАН. 2013. Т. 25. С. 167-178.

   Автором сформулирована общая концепция работы и, совместно с Д. Федоренко, проведены экспериментальные исследования.

3. Texterra: инфраструктура для анализа текстов / Д. Турдаков, Н. Астраханцев, Я. Недумов [и др.] // Труды Института системного программирования РАН. 2014. Т. 26. № 1. С. 421-438.

   Автором написана глава 4, посвященная базе знаний системы Текстерра, в том числе обогащению базы знаний.

4. Астраханцев Н. Автоматическое извлечение терминов из коллекции текстов предметной области с помощью Википедии // Труды Института системного программирования РАН. 2014. Т. 26. № 4. С. 7-20.

5. Fedorenko D., Astrakhantsev N., Turdakov D. Automatic recognition of domain-specific terms: an experimental evaluation // Proceedings of SYRCoDIS 2013. 2013. P. 15-23.

   Автором сформулированы общая концепция и план работы и, совместно с Д. Федоренко, проведены экспериментальные исследования.

6. Astrakhantsev N., Fedorenko D., Turdakov D. Automatic Enrichment of Informal Ontology by Analyzing a Domain-Specific Text Collection // Computational Linguistics and Intellectual Technologies: Papers from the Annual International Conference “Dialogue”. 2014. Vol. 13. P. 29-42.

   Автору принадлежит основополагающий вклад: проведен анализ существующих работ, написан текст статьи и, совместно с Д. Федоренко, проведены экспериментальные исследования.