Московский Государственный Университет
имени М. В. Ломоносова

На правах рукописи

Нокель Михаил Алексеевич

Методы улучшения вероятностных тематических моделей текстовых коллекций на основе лексико-терминологической информации

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 – 2015

Работа выполнена на кафедре алгоритмических языков факультета вычислительной математики и кибернетики федерального государственного бюджетного образовательного учреждения высшего образования «Московский государственный университет имени М. В. Ломоносова».

Научный руководитель:	**Лукашевич Наталья Валентиновна**, кандидат физико-математических наук, ведущий научный сотрудник научно-исследовательского вычислительного центра федерального государственного бюджетного образовательного учреждения высшего образования «Московский государственный университет имени М. В. Ломоносова»
Официальные оппоненты:	**Стрижов Вадим Викторович**, доктор физико-математических наук, ведущий научный сотрудник федерального государственного учреждения «Федеральный исследовательский центр «Информатика и управление» Российской академии наук»
	Поляков Владимир Николаевич, кандидат технических наук, доцент кафедры автоматизированных систем управления федерального государственного автономного образовательного учреждения высшего профессионального образования «Национальный исследовательский технологический университет «МИСиС»
Ведущая организация:	Федеральное государственное бюджетное учреждение науки «Институт системного программирования Российской академии наук»

Защита состоится 11 марта 2016 г. в 11 час. 00 мин. на заседании диссертационного совета Д 501.001.44 при Московском государственном университете имени М. В. Ломоносова по адресу: 119991, г. Москва, ГСП-1, Ленинские горы, д. 1, стр. 52, факультет ВМК, аудитория 685.

С диссертацией можно ознакомиться в Научной библиотеке Московского государственного университета имени М. В. Ломоносова по адресу: 119192, г. Москва, Ломоносовский проспект, д. 27 – а также на официальном сайте факультета ВМК МГУ `https://cs.msu.ru` в разделе «Диссертации».

Автореферат разослан «____» _____ 2015 года.

Учёный секретарь
диссертационного совета Д 501.001.44, О. В. Шестаков
доктор физико-математических наук, доцент

Общая характеристика работы

Актуальность темы. В настоящее время из-за бурного развития сети Интернет наблюдается обилие электронной неструктурированной информации, представленной текстами на естественных языках. Всё более востребованной становится задача автоматической обработки таких текстов для извлечения структурированных данных, которые затем используются при решении различных проблем: извлечения фактических данных, поиска информации и т.п.

Одной из важных задач автоматической обработки текстов является кластеризация текстов и, в частности, выделение тематик, обсуждаемых в коллекциях. Для выявления скрытых тем в текстовых коллекциях в последнее время всё чаще используются **тематические модели**. Это современный инструмент анализа текстов, определяющий, какие темы присутствуют в каждом документе коллекции, и какие слова задают каждую такую тему. При этом темы представляются в виде дискретных распределений на множестве слов, а документы – в виде дискретных распределений на множестве тем. Примером такой темы может служить следующая (для наглядности вероятности слов опущены): *денежный, деньги, обращение, масса, факторинг, средство, функция, оборот...*

Тематические модели исследовались в трудах российских и зарубежных учёных, таких как Воронцов К. В., Машечкин И. В., Петровский М. И., Стрижов В. В., Николенко С. И., Хофманн Т., Блэй Д., Ньюман Д., Мимно Д., Валлах Х., Лау Д. Х. и других авторов.

На данный момент тематические модели успешно применяются в различных сферах информационного поиска, разрешении морфологической неоднозначности, многодокументном аннотировании, машинном переводе, классификации и кластеризации документов и многих других задачах.

В то же время одним из основных недостатков существующих тематических моделей является использование модели «мешка слов», в которой каждый документ представляется в виде множества несвязанных между собой слов. Данная модель не учитывает порядок слов и основывается на гипотезе независимости появлений слов друг от друга в текстах. Это предположение оправдано с точки зрения вычислительной эффективности, но оно далеко от реальности.

Таким образом, актуальной является задача улучшения качества тематических моделей за счёт ухода от модели «мешка слов» путём добавления словосочетаний и многословных выражений (в частности, терминов) и учёта связей между ними и образующими их словами.

Само по себе понятие «термин» не формализовано. Под **термином** чаще всего понимается слово или словосочетание, являющееся точным обозначением определённого понятия какой-либо специальной области науки, техники, искусства, общественной жизни и т.п. Так, в банковской предметной области терминами будут такие слова и словосочетания, как *«банк»*, *«кредит»*, *«инвестиционный фонд»*, *«ипотечный кредит»*, *«лизинг»*. В отличие от общеупотребительных слов термины в рамках конкретной предметной области всегда

однозначны и лишены эмоциональной окраски.

Исторически словари терминов составлялись вручную экспертами. Затем такие словари могли применяться при разработке различных баз знаний, для улучшения качества информационного поиска, машинного перевода, автоматического аннотирования и в других задачах. Однако привлечение экспертов – очень трудоёмкий и дорогостоящий процесс, а получаемые словари, как правило, обладают низкой полнотой покрытия, поскольку не содержат постоянно появляющихся новых терминов. Поэтому актуальной является задача автоматического извлечения терминов для различных предметных областей.

Методы автоматического извлечения терминов исследовались в трудах российских и зарубежных учёных, таких как Большакова Е. И., Браславский П. И., Баева Н. В., Соловьёв С. Ю., Мальковский М. Г., Поляков В. Н., Накагаша Х., Мори Т., Ананиадоу С., Даилле Б., Ахмад К. и других авторов.

На настоящий момент большинство существующих методов для извлечения терминов использует множество критериев, основывающихся на статистических и лингвистических признаках. Однако ни один из таких признаков не является определяющим. Кроме того, они почти не отражают тот факт, что большинство терминов относятся к той или иной тематике, обсуждаемой в текстах коллекции. Поэтому актуальной является задача исследования возможности применения тематической информации при извлечении терминов.

Целью диссертационной работы является разработка методов и программных средств интеграции лексико-терминологической информации в тематические модели, выбирающих наиболее подходящие единицы из слов и словосочетаний для добавления и улучшения качества. Разрабатываемые программные средства и модели должны удовлетворять следующим требованиям:

- Более высокое по сравнению с существующими методами качество тематических моделей;
- Независимость от языков и предметных областей текстовых коллекций;
- Более высокое по сравнению с существующими методами качество извлечённых из текстовых коллекций терминов.

Для достижения этой цели были решены следующие **задачи**:

1. Исследование и разработка методов построения тематических моделей, учитывающих словосочетания и связи между ними и образующими их словами;
2. Разработка и реализация методов извлечения терминов на основе информации, получаемой из тематических моделей.

Основные положения, выносимые на защиту:

1. Предложен и реализован новый метод построения тематических моделей, учитывающий словосочетания и улучшающий характеристики качества

тематических моделей, включая интерпретацию тем экспертами, что полезно для организации человеко-машинных интерфейсов в информационных системах. Для предложенного метода приведено теоретическое обоснование;
2. Предложен и реализован новый итеративный метод добавления словосочетаний в тематические модели, улучшающий меру соответствия тематических моделей словам и словосочетаниям текстовых коллекций (перплексию). Для предложенных методов приводятся теоретические оценки вычислительной сложности;
3. Предложены новые признаки для извлечения терминов, основанные на тематических моделях. Показано, что использование тематической информации улучшает качество извлечения терминов для включения их в базы знаний и терминологические ресурсы;
4. Разработан и выложен в открытый доступ программный комплекс по построению тематических моделей с использованием лексико-терминологической информации.

Научная новизна настоящей диссертационной работы заключается в том, что предложены новые алгоритмы построения тематических моделей, позволяющие добавлять словосочетания и учитывать сходство между ними и образующими их словами. Кроме того, предложены новые признаки для извлечения терминов, основанные на тематической информации. Применимость данных алгоритмов обоснована теоретически, а также численно, на основе тестирования интеграции словосочетаний в тематические модели. Разработанные модели и алгоритмы не зависят от языка и предметной области текстовых коллекций и могут применяться в различных задачах информационного поиска.

Методы исследования. В данной диссертационной работе применялись аналитические методы теории вероятностей, математической статистики, теории обучения машин и теории алгоритмов.

Практическая значимость. На основе предложенных методов спроектирована и разработана многомодульная программная система со следующими функциональными возможностями:

- построения тематических моделей с добавлением словосочетаний, учитывающих сходство между ними и образующими их словами;
- извлечения однословных и двусловных терминов из текстов предметной области;
- автоматических оценок качества построенных тематических моделей и извлечённых терминов.

Таким образом, разработанная система может быть использована как для автоматического пополнения существующих терминологических ресурсов (словарей, тезаурусов), так и для подготовки дополнительной входной информации для других систем обработки текстов на естественном языке.

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

1. Международной конференции по компьютерной лингвистике «Диалог» (Московская область, 30 мая – 3 июня 2012 г.);
2. Летней школе по информационному поиску RUSSIR (Ярославль, 6–10 августа 2012 г.; Казань, 16–20 сентября 2013 г.; Нижний Новгород, 18–22 августа 2014 г.; Санкт-Петербург, 24–28 августа 2015 г.);
3. Международной конференции по информационному поиску ECIR (Москва, 24–27 марта 2013 г.);
4. Всероссийской научной конференции «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» (Ярославль, 14–17 октября 2013 г.; Дубна, 13–16 октября 2014 г.);
5. Международной конференции «Terminology and Artificial Intelligence» (Париж, 28–30 октября 2013 г.);
6. Международной научной конференции студентов, аспирантов и молодых учёных «Ломоносов» (Москва, 7–11 апреля 2014 г.);
7. Международной конференции «Nordic Conference on Computational Linguistics» (Вильнюс, 11–13 мая 2015 г.);
8. Семинаре по многословным выражениям MWE на конференции «North American Chapter of the Association for Computational Linguistics - Human Language Technologies» (Денвер, 31 мая – 5 июня 2015 г.).

Кроме того, результаты обсуждались на семинаре лаборатории анализа информационных ресурсов НИВЦ МГУ и на регулярном семинаре ACM SIGMOD в Москве.

Публикации. Основные результаты по теме диссертации изложены в 12 печатных работах, в том числе в 4 статьях в журналах из списка ВАК [1–4], 3 статьях, входящих в базу SCOPUS [5–7], 1 – в тезисах докладов [8], 4 – в других печатных изданиях [9–12].

Личный вклад автора заключается в выполнении основного объёма теоретических и экспериментальных исследований, изложенных в диссертационной работе, включая разработку теоретических моделей, методик и проведение исследований, анализ и оформление результатов в виде публикаций и научных докладов. В работах [2, 6, 10] Н. В. Лукашевич принадлежит постановка задачи и обсуждение результатов её решения. В работах [1, 5, 9] вклад Е. И. Большаковой ограничен обсуждением результатов, Н. В. Лукашевич – постановкой задачи. В работах [4, 11, 12] Н. В. Лукашевич принадлежит постановка задачи, привлечение лингвистов для получения экспертных оценок тем, выявленных в текстовых коллекциях, и обсуждение результатов.

Результаты научных исследований, представленных в диссертации, были получены в рамках гранта РФФИ №14-07-00383.

Объём и структура диссертации. Диссертация состоит из введения, четырёх глав, заключения и двух приложений. Полный объём диссертации со-

ставляет 137 страниц с 14 рисунками и 30 таблицами, объём приложений – 22 страницы. Список литературы содержит 101 наименование.

Содержание работы

Во **введении** обоснована актуальность диссертационной работы, сформулирована цель и аргументирована научная новизна исследований, показана их практическая значимость, представлены выносимые на защиту научные положения.

Первая глава посвящена описанию основных подходов, применяемых в задачах построения тематических моделей и автоматического извлечения терминов. Особое внимание уделяется рассмотрению существующих способов добавления словосочетаний и многословных выражений в тематические модели. Кроме того, в данной главе описываются существующие признаки ранжирования слов и словосочетаний-кандидатов в термины. Целью данной главы является анализ достоинств и недостатков существующих методов интеграции словосочетаний в тематические модели и автоматического извлечения терминов.

Тематические модели предназначены для выявления скрытых тем в текстовых коллекциях. Они успешно применяются в ряде задач информационного поиска. Однако большинство из них, включая широко известные методы Латентного Размещения Дирихле (LDA) и Вероятностного Латентного Семантического Анализа (PLSA), использует модель «мешка слов», которая далека от реальности и не учитывает порядок слов и их зависимость друг от друга. В последнее время появились подходы, позволяющие добавлять словосочетания в тематические модели. На данный момент существуют два таких подхода: создание единой унифицированной тематической модели, одновременно моделирующей появления слов и словосочетаний[1], и предварительное извлечение словосочетаний для последующего добавления в тематические модели[2].

К достоинствам первых методов относится теоретически элегантное обоснование. Существенным же их недостатком является большое число параметров, которые сложно настраивать на реальных данных. К достоинствам вторых методов относится то же самое число параметров, что и в исходных моделях, а также возможность использования с различными моделями. Однако они склонны к ухудшению основной меры качества – перплексии[3].

Таким образом, для качественного решения задачи построения тематических моделей необходимо разработать метод добавления словосочетаний, не усложняющий исходные модели и улучшающий все меры качества. Кроме то-

[1]Wallach H. Topic Modeling: Beyond Bag-of-Words // Proceedings of the 23rd International Conference on Machine Learning. – 2006. – P. 977-984.

[2]Lau J. H., Baldwin T., Newman D. On Collocations and Topic Models // ACM Transactions on Speech and Language Processing. – ACM Press, 2013. – Vol. 10, № 3. – P. 1-14.

[3]Daud A., Li J., Zhou L., Muhammad F. Knowledge discovery through directed probabilistic topic models: a survey // Frontiers of Computer Science in China. – Higher Education Press, 2010. – Vol. 2, № 2. – P. 280-301.

го, такие модели смогут использоваться и в других прикладных задачах для улучшения качества (в частности, в задаче извлечения терминов).

Несмотря на обилие предложенных для извлечения терминов признаков ранжирования слов и словосочетаний, ни один из них так и не стал определяющим. При этом в последнее время для повышения качества решаемой задачи всё чаще применяются методы машинного обучения. Поэтому актуальной является разработка и комбинирование с помощью машинного обучения новых признаков, позволяющих полнее описывать характеристики извлекаемых терминов, в частности, признаков, основанных на тематиках, обсуждаемых в текстах предметной области, для чего полезно применение тематических моделей.

Во **второй главе** предлагается новый алгоритм построения тематических моделей PLSA-SIM, являющийся модификацией алгоритма PLSA. Предлагаемый алгоритм позволяет добавлять словосочетания и учитывать сходство между ними и образующими их словами. Для выбора и последующего включения словосочетаний в тематические модели изучается возможность применения ассоциативных мер. Также предлагается новый итеративный алгоритм без учителя PLSA-ITER, позволяющий добавлять наиболее подходящие словосочетания. Помимо автоматического выбора словосочетаний рассматривается возможность интеграции в предложенные методы ручных терминологических ресурсов, разработанных экспертами.

Модель «мешка слов», на которой основываются алгоритмы *PLSA* и *LDA*, далека от реальности, поскольку в документах есть много слов и словосочетаний, связанных между собой по смыслу: например, словосочетания, содержащие общие слова (например, *бюджетный – бюджетные расходы – бюджетные доходы – бюджетные средства*). Стоит отметить, что в рамках данной диссертационной работы под **словосочетанием** понимается сочетание двух слов. При этом основная идея предлагаемой модели заключается в том, что если словосочетания, содержащие общие слова, встречаются часто в одних и тех же документах, то их целесообразно отнести к одним и тем же темам, поскольку многие из таких словосочетаний обладают тематической близостью:

$$\begin{aligned} \exists d_i \in D : |D \setminus \{d_i\}| << |D|,\ \forall w \in W : w_1, \ldots, w_k \in S_w, \\ n_{d_i w_1} > 0, \ldots, n_{d_i w_k} > 0 \Longrightarrow r(w_1, \ldots, w_k, t), \end{aligned} \tag{1}$$

где $d_i \in D$ – документы, $w \in W$ – слова, $w_1, \ldots, w_k$ – слова и словосочетания, S_w – множество похожих слов и словосочетаний, $n_{d_i w_k}$ – частотность w_k в документе d_i, $r(w_1, \ldots, w_k, t)$ – отношение принадлежности $w_1, \ldots, w_k$ к темам t.

В результате реализации данной модели был предложен новый алгоритм **PLSA-SIM**, являющийся модификацией исходного алгоритма PLSA. При его описании будут использоваться следующие обозначения[4]:

[4]Воронцов К. В., Потапенко А. А. Модификации EM-алгоритма для вероятностного тематического моделирования // Машинное обучение и анализ данных. – 2013. – Т. 1, № 6. – С. 657-686.

- D – текстовая коллекция;
- T – множество полученных тем;
- W – словарь коллекции (множество уникальных слов и словосочетаний в коллекции документов D);
- $\Phi = \{\phi_{wt} = P(w|t)\}$ – распределение слов (словосочетаний) по темам t;
- $\Theta = \{\theta_{td} = P(t|d)\}$ – распределение тем t по документам d;
- $S = \{S_w\}$ – множества похожих слов и словосочетаний, где S_w – множество слов и словосочетаний, похожих на w:

$$S_w = \{w, \bigcup_v wv, \bigcup_v vw\}, \tag{2}$$

  где w – лемматизированное слово, wv и vw – лемматизированные словосочетания, содержащие w;
- n_{dw} и n_{ds} – частотности слов (словосочетаний) w и s в документе d;
- n_d – длина документа d (общее число слов (словосочетаний) в документе);
- $\hat{n}_{wt}$ – оценка частотности слова (словосочетания) w в теме t;
- $\hat{n}_{td}$ – оценка частотности темы t в документе d;
- $\hat{n}_t$ – оценка частотности темы t в коллекции документов D;
- $P(t|d, w)$ – условная вероятность отнесения вхождения слова (словосочетания) w в документе d к теме t.

Единственная модификация в алгоритме PLSA-SIM (см. Алгоритм 1) касается строчки 7, где в рассмотрение добавляются предварительно вычисленные множества похожих слов и словосочетаний. Тем самым вес подобных слов увеличивается в каждом документе коллекции.

Стоит отметить, что алгоритм PLSA-SIM не увеличивает число параметров оригинального алгоритма – оно остаётся равным $WT + DT$.

Для предложенного алгоритма сформулирована и доказана следующая теорема, отражающая основную идею: похожие слова и словосочетания, часто встречающиеся в одних и тех же документах, имеют максимальную вероятность среди выявленных тем.

Теорема 1. *Пусть имеется коллекция текстов D со словарём W. Пусть u – самое частотное слово в коллекции D:*

$$\forall w \in W\{u\} : \sum_{d \in D} n_{du} > \sum_{d \in D} n_{dw}$$

Тогда при добавлении любых словосочетаний вида uv_j $(j = 1, \ldots, l)$ в словарь W так, что $n_{du} > \sum_{j=1}^{l} n_{duv_j}$, и построении тематической модели алгоритмом PLSA-SIM с одной темой t будут выполнены следующие неравенства:

$$\forall j = 1, \ldots, l \ \forall w \in W\{u, v_1, \ldots, v_l\} :$$

Алгоритм 1: Алгоритм PLSA-SIM

Вход: коллекция документов D, число тем T, множества похожих слов и словосочетаний S, начальные приближения Φ и Θ

Выход: распределения Φ и Θ

while *не выполнится критерий остановки* **do**
  for $d \in D$, $w \in W$, $t \in T$ **do**
    $\hat{n}_{wt} = 0$, $\hat{n}_{td} = 0$, $\hat{n}_t = 0$
  for $d \in D$, $w \in W$ **do**
    for $t \in T$ **do**
      $P(t|d,w) = \frac{\phi_{wt}\theta_{td}}{\sum\limits_{s \in T} \phi_{ws}\theta_{sd}}$
      $\hat{n}_{wt}, \hat{n}_{td}, \hat{n}_t += \left(n_{dw} + \sum\limits_{s \in S_w} n_{ds}\right) P(t|d,w)$
  for $d \in D$, $w \in W$ **do**
    $\phi_{wt} = \frac{\hat{n}_{wt}}{\hat{n}_t}$
  for $d \in D$, $t \in T$ **do**
    $\theta_{td} = \frac{\hat{n}_{td}}{n_d}$

$$P(uv_j|t) \geq P(u|t)$$

$$P(uv_j|t) > P(w|t) \ и \ P(u|t) > P(w|t)$$

Также была предложена итеративная модель учёта словосочетаний в определении тематической структуры текстов, основывающаяся на идее, что для добавления в тематические модели наиболее подходящие словосочетания выбираются исходя из вида верхушек списков слов, образующих темы. Для этой цели в каждой теме из первых слов составляются все возможные словосочетания, которые затем добавить в тематическую модель при построении:

$$W = \{u_1, \ldots, u_n, u_{1t_1}u_{2t_1}, \ldots, u_{it_k}u_{jt_k}\}, \tag{3}$$

где W – словарь коллекции, u_i – слова, $u_{it_k}u_{jt_k}$ – добавляемые в модель словосочетания, составленные из верхушек темы t_k.

Для реализации данной модели был предложен новый итеративный алгоритм **PLSA-ITER** (см. Алгоритм 2). При описании предлагаемого алгоритма используются следующие дополнительные обозначения:

- B – множество всех словосочетаний в коллекции документов D;
- B_A – множество словосочетаний, добавленных в тематическую модель;
- B_i – множество словосочетаний, добавленных в тематическую модель на i-й итерации;

- S_i – множество потенциальных кандидатов на похожие слова и словосочетания на i-й итерации;
- $(u_1, u_2, \ldots, u_{10})$ – первые 10 слов в теме t;
- $TF(u_i u_j)$ – частотность словосочетания $u_i u_j$.

Алгоритм 2: Итеративный алгоритм PLSA-ITER

Вход: коллекция документов D, число тем $|T|$, множество словосочетаний B

Выход: полученные темы T

```
Запуск алгоритма PLSA на коллекции D для получения тем T
B_A = ∅
while не выполнится критерий остановки do
    S_i = ∅, B_i = ∅
    for t ∈ T do
        S_i = S_i ⋃ {u_1, u_2, ..., u_10}
        for u_i, u_j ∈ (u_1, u_2, ..., u_10) do
            if u_i u_j ∈ B и u_j u_i ∈ B и TF(u_i u_j) > TF(u_j u_i) then
                B_i = B_i ⋃ {u_i u_j}
    Создание множества похожих слов и словосочетаний S из S_i ⋃ B_i
    B_A = B_A ⋃ B_i
    Запуск PLSA-SIM с множествами S и B_A
```

Стоит отметить, что алгоритм PLSA-ITER не увеличивает число параметров оригинального алгоритма – оно остаётся равным $WT + DT$.

Для оценки качества тематических моделей использовались следующие меры:

- *Перплексия*, рассчитываемая по контрольной выборке:

$$Perplexity(D) = \exp\left(-\frac{1}{n}\sum_{d \in D}\sum_{w \in d} n_{dw} \log P(w|d)\right), \tag{4}$$

  где n – число всех рассматриваемых слов в текстовой коллекции, D – множество всех документов в коллекции, n_{dw} – частотность слова w в документе d, $P(w|d)$ – вероятность появления слова w в документе d;
- *TC-PMI*, оценивающая согласованность тем:

$$TC\text{-}PMI = \frac{1}{|T|}\sum_{t=1}^{|T|}\sum_{j=2}^{10}\sum_{i=1}^{j-1} \log \frac{P(w_i w_j)}{P(w_i)P(w_j)}, \tag{5}$$

  где T – множество полученных тем, $(w_1, w_2, \ldots, w_{10})$ – первые 10 слов в рассматриваемой теме t, $P(w_i)$ и $P(w_j)$ – вероятности слов w_i и w_j соответственно, $P(w_i w_j)$ – вероятность словосочетания $w_i w_j$;

- *TC-PMI-nSIM* – предложенная модификация меры *TC-PMI*, учитывающая первые 10 непохожих слов в каждой теме;
- *Экспертные оценки.* Перед экспертами ставилась задача классификации тем на два класса в зависимости от того, можно ли дать каждой теме некоторое название или нет.

Тестирование алгоритма PLSA-SIM показало существование группы из 6 мер, ранжирующих словосочетания таким образом, что при добавлении 1000 лучших словосочетаний по любой из них в алгоритм PLSA-SIM значительно улучшается качество тематических моделей по всем целевым мерам независимо от языка и предметной области. В Таблице 1 приведены результаты добавления 1000 лучших словосочетаний, упорядоченных по *частотности (TF)*. Результаты остальных мер из данной группы похожи на результаты, приведённые в этой таблице. Экспертные оценки полученных тем также подтверждают улучшение качества независимо от текстовой коллекции.

Таблица 1: Сравнение алгоритмов PLSA и PLSA-SIM

Корпус	**Модель**	**Перплексия**	**TC-PMI**	**TC-PMI-nSIM**
Банковский	*PLSA*	1724.2	86.1	86.1
	PLSA-SIM	**1450.6**	**156.5**	**102.6**
Europarl	*PLSA*	1594.3	53.2	53.2
	PLSA-SIM	**1431.6**	**127.7**	**84.7**
JRC-Acquiz	*PLSA*	812.1	67	67
	PLSA-SIM	**743.7**	**108.4**	**76.9**
ACL Anthology	*PLSA*	2134.7	74.8	74.8
	PLSA-SIM	**1806.4**	**152.7**	**87.8**

При тестировании же алгоритма PLSA-ITER выяснилось, что в процессе его работы на каждой итерации образуется очень мало множеств похожих слов и словосочетаний. Для нахождения большего числа похожих слов и словосочетаний из образующихся кандидатов использовались стеммеры, т.е. алгоритмы, пытающиеся найти основы для заданных слов.

Для того чтобы учесть стеммеры в итеративном алгоритме, было модифицировано определение множеств похожих слов и словосочетаний $S = \{S_w\}$, где множество S_w теперь задаётся следующим образом:

$$S_w = \{w,\ \bigcup_u u,\ \bigcup_{u,v} uv : stem(u) = stem(w) \text{ или } stem(v) = stem(w)\}, \qquad (6)$$

где w и u – лемматизированные слова, uv – лемматизированное словосочетание, $stem(u)$ – основа слова u, получающаяся в результате работы стеммера.

В Таблице 2 представлены результаты работы алгоритма PLSA-ITER после первой итерации вместе с результатами работы алгоритма PLSA-SIM с добавленными 1000 самых частотных словосочетаний. Экспертные оценки полученных тем тоже подтверждают улучшение качества независимо от коллекции.

Таблица 2: Сравнение алгоритма PLSA-ITER со стеммерами и алгоритма PLSA-SIM с добавленными 1000 самых частотных словосочетаний

Коллекция	**Модель**	**Перплексия**	**TC-PMI**	**TC-PMI-nSIM**
Банковская	*PLSA-SIM*	1450.6	156.5	102.6
	PLSA-ITER + Snowball	**1265.1**	137.6	96.7
Europarl	*PLSA-SIM*	1431.6	127.7	84.7
	PLSA-ITER + Ланкастер	**1077.7**	105	55.2
JRC-Acquiz	*PLSA-SIM*	743.7	108.4	76.9
	PLSA-ITER + Ланкастер	**736.5**	94.5	68.6
ACL	*PLSA-SIM*	1806.4	152.7	87.8
	PLSA-ITER + Ланкастер	**1772.1**	121.3	76.5

Данные результаты показывают, что использование стеммера Snowball[5] для русского языка и жадного стеммера Ланкастерского университета[6] для английского языка в алгоритме PLSA-ITER приводит к дальнейшему улучшению качества по перплексии с незначительным падением согласованности тем.

Помимо автоматического выбора словосочетаний также рассматривалась возможность интеграции в предложенные методы словосочетаний из ручных терминологических ресурсов, разработанных экспертами (так называемых *«золотых стандартов»*). Под **термином** в рамках данной диссертационной работы понимается эталонное слово или сочетание двух слов, выделенное экспертом, содержащееся в «золотом стандарте».

В Таблице 3 представлены результаты интеграции 1000 самых частотных терминов в алгоритм PLSA-SIM вместе с результатами первой итерации алгоритма PLSA-ITER со стеммерами, добавляющего словосочетания-термины.

Таким образом, предложенные алгоритмы позволяют интегрировать в тематические модели словосочетания из ручных терминологических ресурсов, улучшая качество по сравнению с оригинальным алгоритмом.

Третья глава посвящена исследованию возможности применения тематической информации в задаче автоматического извлечения терминов. Для это-

[5] `http://snowball.tartarus.org/algorithms/russian/stemmer.html`

[6] Paice C. D. Another Stemmer // SIGIR Forum. – 1990. – Vol. 24, № 3. – P. 56-61.

Таблица 3: Результаты интеграции терминов в предложенные алгоритмы

Коллекция	**Модель**	**Перплексия**	**TC-PMI**	**TC-PMI-nSIM**
Банковская	*PLSA*	1724.2	86.1	86.1
	PLSA-SIM + термины	1475.2	**151.8**	**102.9**
	PLSA-ITER + термины	**1267.6**	134.9	96
Europarl	*PLSA*	1594.3	53.2	53.2
	PLSA-SIM + термины	1522.4	**133.9**	**84.8**
	PLSA-ITER + термины	**1193.5**	97.4	66.6

го предлагаются новые признаки терминологичности слов и словосочетаний-кандидатов, основанные на тематических моделях, и разрабатываются комбинированные модели извлечения терминов. Главными целями являются нахождение наилучших индивидуальных признаков и определение вклада предложенных тематических признаков, а также комбинаций признаков, осуществляющих ранжирование так, чтобы в начале итогового списка стояли кандидаты, с наибольшей вероятностью являющиеся терминами.

Автоматическое извлечение терминов представляет собой процедуру упорядочивания множества слов и словосочетаний S из текстовой коллекции D так, чтобы эталонные термины T_e оказались в начале списка. Итак, пусть имеется X – множество описаний слов и словосочетаний, использующихся в текстах некоторой предметной области, и 2 целевых класса $\{0, 1\}$, соответствующие тому, что рассматриваемое слово (словосочетание) является термином или нет. При этом на объектах обучающей выборки $X^m = \{x_1, \ldots, x_l\}$ известен правильный порядок $i \prec j$ на парах $(i, j) \in \{1, \ldots, l\}$. Требуется построить ранжирующую функцию $a : X \to \{0, 1\}$, сохраняющую правильный порядок на парах (i, j):

$$i \prec j \Longrightarrow a(x_i) \prec a(x_j) \tag{7}$$

и максимизирующую число эталонных терминов в начале списка:

$$a^* = \arg\max_a |\{t_e \in T_e | \forall s \in S \setminus T_e : a(s) \prec a(t_e)\}| \tag{8}$$

В качестве описаний слов и словосочетаний используются различные признаки. Признаком называется отображение $f : X \to D_f$, где D_f – множество допустимых значений признака. Если заданы признаки $f_1, \ldots, f_n$, то вектор $x = (f_1(x), \ldots, f_n(x))$ называется признаковым описанием объекта $x \in X$, отождествляемое, как правило, с самим объектом.

На текущий момент все предложенные признаки извлечения терминов слабо отражают тот факт, что большинство терминов относятся к той или иной **тематике**, обсуждаемой в рамках текстов предметной области. Иными словами, множество всех терминов $Term$ представимо в виде:

$$Term = Term_{general} \bigcup \left(\bigcup_{t \in T} Term_t \right), \tag{9}$$

где $Term_{general}$ – множество общих терминов, не относящихся к какой-либо тематике, T – множество всех тематик в текстовой коллекции, $Term_t$ – множество терминов, относящихся к тематике t.

При этом, как правило, мощность множества общих терминов намного меньше мощности множеств терминов, относящихся к тематикам, а сами эти множества не пересекаются:

$$\forall t \in T : |Term_{general}| << |Term_t| \text{ и } Term_{general} \bigcap Term_t = \emptyset \tag{10}$$

При этом множества $Term_t$ могут пересекаться, поскольку некоторые термины могут относиться сразу к нескольким тематикам в рамках текстов предметной области. Для выявления тематик, как правило, применяются подходы, основанные на тематическом моделировании.

Для проверки предположения о принадлежности терминов тематикам были предложены новые признаки, вычисляемые по построенным тематическим моделям. При их описании используются следующие обозначения:

- $P(w|t)$ – условная вероятность слова или словосочетания w в теме t;
- $DF(w)$ – число тем, содержащих слово или словосочетание w с $P(w|t) > \epsilon$;
- T – множество тем, полученных из текстовой коллекции.

Всего было предложено **7** тематических признаков:

1. *Частотность (TF)*, поощряющая слова и словосочетаний с большой суммарной вероятностью во всех темах:

$$TF(w) = \sum_{t \in T} P(w|t) \tag{11}$$

2. *Максимальная частотность (Maximum TF)*, ранжирующая вверх слова и словосочетания, имеющие наибольшую вероятность в какой-либо из тем:

$$Maximum\ TF(w) = \max_{t \in T} P(w|t) \tag{12}$$

3. *TF-IDF*, поощряющий слова и словосочетания, встречающиеся часто в малом числе тем:

$$TF\text{-}IDF(w) = TF(w) \times \log \frac{|T|}{DF(w)} \tag{13}$$

4. *Domain Consensus*, основанный на энтропии и поощряющий слова и словосочетания, встречающиеся в различных темах:

$$\boldsymbol{Domain\ Consensus}(w) = -\sum_{t \in T} (P(w|t) \times \log P(w|t)) \tag{14}$$

5. *Term Score (TS)*, принимающий низкие значения для слов и словосочетаний, имеющих высокую вероятность принадлежности всем темам:

$$TS(w) = \sum_{t \in T} TS_t(w), \text{ где } TS_t(w) = P(w|t) \times \log \frac{P(w|t)}{\left(\prod\limits_{t \in T} P(w|t)\right)^{\frac{1}{|T|}}} \tag{15}$$

6. *Maximum Term Score (Maximum TS)*, представляющий собой максимальное $\boldsymbol{Term\ Score}$ среди всех тем:

$$\boldsymbol{Maximum\ TS}(w) = \max_{t \in T} TS_t(w) \tag{16}$$

7. *TS-IDF*, комбинирующий идеи признаков $\boldsymbol{TF\text{-}IDF}$ и $\boldsymbol{Term\ Score}$:

$$\boldsymbol{TS\text{-}IDF}(w) = TS(w) \times \log \frac{|T|}{DF(w)} \tag{17}$$

Для оценки качества извлечённых терминов использовалась мера *Средней точности* на уровне первых n кандидатов:

$$AvP@n = \frac{1}{m} \sum_{k=1}^{n} \left(r_k \times \left(\frac{1}{k} \sum_{1 \le i \le k} r_i \right) \right), \tag{18}$$

где $r_i = 1$, если i-й кандидат является эталонным термином, и $r_i = 0$ в противном случае, а m – число эталонных терминов среди извлечённых кандидатов.

Для определения лучшей тематической модели для извлечения однословных и двусловных терминов сравнивались различные традиционные и вероятностные тематические модели: метод K-средних, сферический K-средних, иерархические алгоритмы, метод NMF (неотрицательной матричной факторизации), PLSA и LDA. Экспериментально было выявлено, что лучшее качество независимо от языка и предметной области даёт тематическая модель **PLSA**.

Для изучения вклада тематической информации в задачу извлечения терминов результаты предложенных тематических признаков, посчитанных для лучшей модели PLSA, сравнивались с остальными признаками. Так, лучшими признаками для извлечения однословных терминов оказались предложенные тематические (**TS-IDF** для банковского корпуса и **Maximum TS** – для корпуса Europarl), а для извлечения двусловных терминов – предложенный контекстный признак **Modified Gravity Count**:

$$\boldsymbol{Modified\ Gravity\ Count}(x, y) = \log \left(\frac{TF(x,y) \times l(x)}{TF(x)} + \frac{TF(x,y) \times r(y)}{TF(y)} \right), \tag{19}$$

где $l(x)$ – число различных слов, встретившихся слева от x, $r(y)$ – число различных слов, встретившихся справа от y.

Для оценки вклада тематических признаков в модели извлечения однословных и двусловных терминов модель, учитывающая тематические признаки, сравнивалась с моделью, не учитывающих их. Результаты сравнения на уровне 5000 самых частотных слов и словосочетаний-кандидатов в термины приведены в Таблице 4.

Таблица 4: Средняя точность AvP@5000 моделей извлечения терминов с тематическими признаками и без них

Модель	**Однословные термины**		**Двусловные термины**	
	Банковский корпус	**Корпус Europarl**	**Банковский корпус**	**Корпус Europarl**
Без тематических признаков	57.3	58.5	70.8	60.0
С тематическими признаками	**59.0**	**58.7**	**71.6**	**60.3**

С помощью одностороннего непараметрического критерия знаковых рангов Вилкоксона была показана статистическая значимость результатов, из которой следует, что тематические модели вносят дополнительную информацию в процесс извлечения терминов.

Также в задаче извлечения однословных терминов были апробированы модели, получаемые предложенным алгоритмом PLSA-SIM. Для данной цели были выбраны тематические модели, построенные алгоритмом PLSA-SIM с 1000 самых частотных словосочетаний. По построенным моделям вычислялись те же самые предложенные тематические признаки. Однако для учёта влияния добавленных в тематическую модель словосочетаний использующаяся в данных признаках вероятность принадлежности слова w теме t $P(w|t)$ была заменена на $\hat{P}(w|t)$, вычисляемую согласно признаку $C\text{-}Value$:

$$\hat{P}(w|t) = P(w|t) - \frac{\sum\limits_{p \in P_w} P(p|t)}{|P_w|}, \tag{20}$$

где w – слово, p – словосочетание, $P(w|t)$ и $P(p|t)$ – вероятности слова w и словосочетания p принадлежности теме t, P_w – множество всех добавленных в тематическую модель словосочетаний, содержащих слово w.

Для оценки вклада предложенных признаков в общую модель извлечения однословных терминов модель, учитывающая данные признаки, сравнивалась с моделью, не использующей их. Результаты сравнения на уровне 5000 самых частотных слов приведены в Таблице 5.

Таблица 5: Средняя точность AvP@5000 моделей извлечения однословных терминов с признаками, посчитанными по модели PLSA-SIM, и без них

Модель	**Русский корпус**	**Английский корпус**
Без признаков по PLSA-SIM	59.0	58.7
С признаками по PLSA-SIM	**59.9**	**58.9**

С помощью одностороннего непараметрического критерия знаковых рангов Вилкоксона была показана статистическая значимость результатов, из которой следует, что признаки, посчитанные по модели PLSA-SIM, вносят дополнительную информацию в процесс извлечения однословных терминов.

В рамках данной диссертационной работы был разработан программный комплекс по построению тематических моделей с использованием лексико-терминологической информации и применению тематических моделей для извлечения терминов, описание которого приведено в **четвёртой главе**. Данный комплекс выложен в открытый доступ[7] и включает в себя следующие условно-независимые пакеты программ:

- Пакет программ построения тематических моделей с возможностью добавления словосочетаний и похожих слов, объединяющий в себе следующие модули: *модуль преобразования входных данных*, *модуль добавления словосочетаний в тематические модели*, *модуль построения инвертированного индекса* и *модуль построения тематических моделей*.
- Пакет программ извлечения однословных и двусловных терминов, объединяющий в себе следующие модули: *модуль извлечения кандидатов в термины*, *модуль вычисления признаков* и *модуль машинного обучения*;

Данные пакеты программ объединены в единое целое и могут взаимодействовать друг с другом по принципу конвейера в любой последовательности.

Также в рамках диссертационной работы сформулировано и доказано утверждение о вычислительных сложностях предложенных алгоритмов построения тематических моделей PLSA-SIM и PLSA-ITER.

Утверждение 1. *Вычислительные сложности алгоритма PLSA-SIM и одной итерации PLSA-ITER совпадают и равны* $O(NTi + NK + WK\log K + DK\log K)$, *где* N *– число ненулевых элементов терм-документной матрицы*[8], T *– число тем,* i *– число итераций алгоритма PLSA-SIM,* K *– число добавленных словосочетаний,* W *– размер словаря,* D *– число документов в коллекции.*

[7] `https://bitbucket.org/Meister17/dissertation`

[8] *Терм-документная матрица* – матрица, строки которой соответствуют документам в коллекции, столбцы – словам и словосочетаниям, а в ячейках записана частотность слов и словосочетаний в каждом документе.

Стоит отметить, что на реальных данных выполняются следующие условия: $W << N$ и $D << N$. Тогда оказывается справедливой следующая теорема.

Теорема 2. *При условиях* $W = O\left(\frac{N}{\ln T}\right)$ *и* $D = O\left(\frac{N}{\ln T}\right)$ *вычислительные сложности алгоритмов PLSA-SIM и одной итерации PLSA-ITER совпадают с вычислительной сложностью алгоритмов PLSA и LDA и равны* $O(NTi)$.

Данная теорема подтверждает, что предложенные алгоритмы не увеличивают вычислительную сложность исходных алгоритмов.

В **заключении** приведены основные результаты данной работы, а также обозначены направления для дальнейших исследований.

В **приложении А** приведён список первых 10 слов из тем, полученных оригинальным алгоритмом PLSA на банковском корпусе.

В **приложении Б** приведён список первых 10 слов и словосочетаний из тем, полученных предложенным алгоритмом PLSA-SIM на банковском корпусе с добавлением 1000 самых частотных словосочетаний.

Публикации автора по теме диссертации

Публикации из списка ВАК

1. Большакова Е. И., Лукашевич Н. В., Нокель М. А. Извлечение однословных терминов из текстовых коллекций на основе методов машинного обучения // Информационные технологии. – 2013. – № 7. – С. 31-37.

2. Нокель М. А., Лукашевич Н. В. Тематические модели в задаче извлечения однословных терминов // Программная инженерия. – 2014. – № 3. – С. 34-40.

3. Нокель М. А. Метод учёта структуры биграмм в тематических моделях // Вестник ВГУ, Серия: Системный анализ и информационные технологии. – 2014. – № 4. – С. 89-97.

4. Нокель М. А., Лукашевич Н. В. Тематические модели: добавление биграмм и учет сходства между униграммами и биграммами // Вычислительные методы и программирование. – 2015. – Т. 16, № 2. – С. 215-234.

Публикации из международной базы цитирования Scopus

5. Bolshakova E., Loukachevitch N., Nokel M. Topic Models Can Improve Domain Term Extraction // ECIR Proceedings. Серия LNCS. – Издательство SPRINGER HEIDELBERG, 2013. – Т. 7814. – С. 684-687.

6. Nokel M., Loukachevitch N. Application of topic models to the task of single-word term extraction // CEUR Workshop Proceedings. – 2013. – Vol. 1108. – P. 52-60.

7. Nokel M. Topic models: Taking into account similarity between unigrams and bigrams // CEUR Workshop Proceedings. – 2014. – Vol. 1297. – P. 243-252.

Прочие публикации

8. Нокель М. А. Использование лингвистической информации в тематической модели PLSA // Сборник тезисов XXI Международной научной конференции студентов, аспирантов и молодых учёных «Ломоносов-2014». Секция «Вычислительная Математика и Кибернетика». – 2014. – С. 120-121.

9. Nokel M., Bolshakova E., Loukachevitch N. Combining Multiple Features For Single-Word Term Extraction // Компьютерная лингвистика и интеллектуальные технологии. По материалам конференции Диалог-2012. – 2012. – С. 490-501.

10. Nokel M., Loukachevitch N. An Experimental Study of Term Extraction for Real Information-Retrieval Thesauri // Proceedings of the 10th International Conference on Terminology and Artificial Intelligence (TIA-2013). – 2013. – P. 69-76.

11. Nokel M., Loukachevtich N. Topic Models: Accounting Component Structure of Bigrams // Proceedings of the 20th Nordic Conference of Computational Linguistics (NODALIDA-2015). – 2015. – P. 145-152.

12. Nokel M., Loukachevitch N. A Method of Accounting Bigrams in Topic Models // Proceedings of North American Chapter of the Association for Computational Linguistics – Human Language Technologies (NAACL-HLT). – 2015. – P. 1-9.