Московский государственный университет имени М. В. Ломоносова

На правах рукописи

Турдаков Денис Юрьевич

# Методы и программные средства разрешения лексической многозначности терминов на основе сетей документов

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

АВТОРЕФЕРАТ

диссертации на соискание ученой степени

кандидата физико-математических наук

Москва – 2010

Работа выполнена на кафедре системного программирования факультета вычислительной математики и кибернетики Московского государственного университета имени М. В. Ломоносова.

| | |
|---|---|
| Научный руководитель: | *доктор технических наук,* *профессор* ***Кузнецов Сергей Дмитриевич***. |
| Официальные оппоненты: | *доктор физико-математических наук,* *профессор* ***Соловьев Валерий Дмитриевич***; *кандидат физико-математических наук,* *заведующий лабораторией* ***Добров Борис Викторович***. |
| Ведущая организация: | *Санкт-Петербургский государственный университет* |

Защита состоится «5» марта 2010 года в 11 часов на заседании диссертационного совета *Д 501.001.44 Московского государственного университета имени М.В. Ломоносова* по адресу: *119991, ГСП-1, Москва, Ленинские горы, МГУ имени М.В.Ломоносова, 2-й учебный корпус, факультет ВМК, ауд. 685.*

С диссертацией можно ознакомиться в библиотеке *факультета ВМК МГУ*. С текстом автореферата можно ознакомиться на официальном сайте ВМК МГУ http://cs.msu.ru в разделе «Наука» — «Работа диссертационных советов» — «Д 501.001.44»

Автореферат разослан «5» февраля 2010 г.

Ученый секретарь
диссертационного совета
*профессор* *Н.П. Трифонов*

## Общая характеристика работы

### Актуальность темы

Разрешение лексической многозначности является одной из центральных задач обработки текстов. Задача заключается в установлении значений слов или составных терминов в соответствии с контекстом, в котором они использовались. Разрешение лексической многозначности используется для повышения точности методов классификации и кластеризации текстов, увеличения качества машинного перевода, информационного поиска и других приложений.

Для решения задачи необходимо определить возможные значения слов и отношения между этими значениями и контекстом, в котором использовались слова. На данный момент основным источником значений являются словари и энциклопедии. Для установления связей между значениями лингвистами создаются тезаурусы, семантические сети и другие специализированные структуры. Однако создание таких ресурсов требует огромных трудозатрат.

В начале 21-го века исследователи в области обработки естественного языка заинтересовались возможностью использования сетей документов, таких как Веб и Википедия, связанных гиперссылками, созданных огромным числом независимых пользователей, и обладающих высокой степенью актуальности.

Открытая энциклопедия Википедия является беспрецедентным ресурсом. Она позволяет автоматически составить словарь терминов, достаточный для описания любых текстовых документов, сопоставить термины со значениями, описанными в статьях Википедии, и на основе ссылочной структуры вывести отношения между этими значениями. Словарь Википедии позволяет автоматически находить в документах как отдельные слова, так и составные термины. На основе разрешения лексической многозначности выделенных

терминов, возможно определить основные тематические линии, нахождение которых необходимо для большого числа практических приложений.

**Цель диссертационной работы**

Целью диссертационной работы является разработка методов и программных средств разрешения лексической многозначности терминов на основе структурной и текстовой информации сетей документов. Разрабатываемые методы должны обладать следующими свойствами: они должны быть полностью автоматическими; соотношение точности и полноты должно быть равно или превышать аналогичный показатель методов, представленных в современной литературе; время работы алгоритмов должно линейно зависеть от количества обрабатываемых терминов; методы не должны быть привязаны к синтаксису конкретных языков.

Для достижения этой цели были поставлены следующие задачи:

1. разработать метод для автоматического определения отношений между значениями терминов Википедии;

2. разработать методы разрешения лексической многозначности терминов, на основе структурной и текстовой информации Википедии.

**Научная новизна**

Научной новизной обладают следующие результаты работы:

1. предложен подход к разрешению лексической многозначности терминов на основе сети документов Википедии.

2. разработан метод разрешения лексической многозначности, основанный на Марковской модели высокого порядка, где параметры модели оценивались на основе структурной и текстовой информации Википедии;

3. предложено обобщение Марковской модели на случай множества независимых Марковских процессов и разработан алгоритм вычисления наи-

более вероятной последовательности состояний, удовлетворяющей ограничениям модели;

4. разработан метод разрешения лексической многозначности и выделения лексических цепей, основанный на обобщенной Марковской модели.

**Практическая значимость** Разработанные методы разрешения лексической многозначности, основанные на Википедии, могут применяться для повышения точности реальных практических приложений, предназначенных для обработки и анализа текстовых данных.

На основе предложенных методов разработан прототип системы разрешения лексической многозначности. Этот прототип был использован в качестве основы для создания в Институте системного программирования РАН системы анализа текстов «Texterra».

**Апробация работы и Публикации.**

По материалам диссертации опубликовано восемь работ [1–8]. Основные положения докладывались на следующих конференциях и семинарах:

- на четвертом и пятом весеннем коллоквиуме молодых исследователей в области баз данных и информационных систем (SYRCoDIS) (2007 и 2008 гг.);
- на сто двадцать пятом и сто тридцать шестом заседаниях Московской Секции ACM SIGMOD (2008 и 2009 гг.);
- на тридцать четвертой международной конференции по очень большим базам данных (VLDB) (2008 г.);
- на международном симпозиуме по извлечению знаний из социального Веба (KASW) (2008 г.);

- на одиннадцатой Всероссийской научной конференции «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» (2009 г.);

- на двадцать третей международной конференции по проблемам языка, информации и вычислений (PACLIC) (2009 г.).

**Структура и объем диссертации**

Работа состоит из введения, трех глав, заключения и списка литературы. Общий объем диссертации составляет 138 страниц. Список литературы содержит 119 наименований.

## Содержание работы

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

**Первая глава** является обзорной и содержит изложение методов и средств, которые лежат в основе разработок, описываемых в последующих главах. В данной главе определяются основные понятия, необходимые для дальнейшего описания методов и алгоритмов. Описывается процесс лингвистической обработки текстов, и очерчиваются границы решаемой задачи.

В первом разделе главы обсуждаются вопросы терминологии, используемой в дисертационной работе. Исследователи выделяют несколько типов многозначности естественного языка, и для работы с каждым из этих типов существуют собственные методы. Данная работа посвящена лексической многозначности и методам ее разрешения. Под разрешением лексической многозначности понимается установление значений слов или составных терминов

в соответствии с контекстом, в котором они использовались.

Во втором разделе первой главы описываются основные проблемы, возникающие при решении задачи разрешения лексической многозначности. В этом разделе рассмотрены существующие подходы к определению значений терминов и контекста, а также методы сравнения различных алгоритмов разрешения лексической многозначности.

В третьем разделе приводится обзор существующих методов разрешения лексической многозначности. Основная часть современных алгоритмов была разработана в 90-е годы прошлого века и основана на тезаурусе WordNet. Из множества всех существующих алгоритмов можно выделить два доминирующих класса: методы, основанные на основе внешних источниках знаний, и методы, основанные на обучении по размеченным корпусам.

Алгоритмы разрешения лексической многозначности на основе внешних источников знаний могут быть легко адаптированы к документам, полученным из любых источников и не привязаны к конкретному языку. Методы, основанные на машинном обучении показывают лучшие результаты из всех алгоритмов, представленных в современной литературе, однако требуют обучения, на документах, похожих на обрабатываемые в дальнейшем. Это связано с проблемой разреженности языка.

Для установления отношений между значениями используются статистические оценки и оценки полученные на основе лексико-семантических ресурсов, называемые в литературе семантической близостью значений.

**Определение 1.** *семантической близостью называется отображение* $f : X \times X \to \mathcal{R}$*, ставящее в соответствие паре слов, терминов или их значений действительное число и обладающее следующими свойствами:*

1. $0 \leq f(x, y) \leq 1$,
2. $f(x, y) = 1 \Leftrightarrow x = y$.

Семантическая близость часто используется методами из обоих описанных классов класса, для преодоления проблемы разреженности языка. На практике для определения таких отображений используются специализированные лексико-семантические ресурсы, создающиеся экспертами в области лингвистики, что существенно усложняет процесс разработки практических приложений.

В начале 21-го века исследователей в области обработки естественного языка заинтересовала возможность использования сетей документов, таких как Веб и Википедия, связанных гиперссылками и созданных огромным числом независимых пользователей, для автоматического построения лексико-семантических ресурсов, что позволит исключить наиболее дорогостоящий этап в создании систем обработки естественного языка.

**Во второй главе** определяется понятие сети документов, рассматриваются отличия сетей документов, от лексико-семантических ресурсов, созданных экспертами и описываются методы вычисления семантической близости значений терминов Википедии, на основе ее структурной и текстовой информации.

Под *сетью документов* понимается случайный граф, вершинами которого являются текстовые документы, а ребрами — гипертекстовые ссылки между ними. При этом, распределение степеней узлов в таких сетях, отвечает степенному закону, то есть доля $P(k)$ вершин в сети, имеющих $k$ связей с другими узлами описывается законом $P(k) \sim k^{-\gamma}$, где $\gamma$ — константа, для большинства реальных сетей находящаяся в интервале $2 < \gamma < 3$. Таким образом, сети документов являются частным случаем безмасштабных графов (scale-free graph).

В первом разделе второй главы описываются свойства сетей документов и безмасштабных графов и показывается, что сети документов структурно отличаются от семантических сетей и тезаурусов, созданных экспертами. Это

накладывает отпечаток на способы их использования, в частности, на методы вычисления семантической близости.

Во втором разделе данной главы рассматриваются подходы к вычислению семантической близости в сетях документов и определяется класс методов, пригодных для решения поставленной задачи.

В работах, представленных в современной литературе, было показано, что процесс образования графов, где узлами служат текстовые документу хорошо описывается гипотезой, что «документы с похожим содержимым имеют тенденцию ссылаться друг на друга», что является основанием для использования ссылочной структуры сетей документов для определения коэффициента семантической близости между этими документами.

Методы, основанные на ссылочной структуре графов, представленные в современной литературе, можно разделить на два класса: локальные и глобальные. Наиболее эффективными считаются локальные методы, основанные на пересечении множеств ближайших соседей узлов сети, где соседними считаются узлы, непосредственно соединенные ссылкой, так как они позволяют получить хорошие оценки при малой вычислительной сложности.

Глобальные методы основаны на рекурсивном обходе графа. В диссертационной работе приводится обзор глобальных методов, и выделяются проблемы, связанные с практическим применением описанных методов для вычисления семантической близости узлов в сети документов. Так как реально существующие сети документов содержат миллионы узлов и сотни миллионов ребер, их размер накладывает жесткие требования на алгоритмы вычисления семантической близости. Так на примере алгоритма SimRank показано, что большая вычислительная сложность и объем получаемых данных не позволяют эффективно использовать этот класс алгоритмов для вычисления семантической близости между узлами Википедии.

В третьем разделе предлагается собственный метод вычисления семанти-

ческой близости между документами Википедии. Википедия — это современная электронная энциклопедия, обладающая огромным словарем и высокой степенью актуальности, при этом сравнительно легко поддающаяся автоматической обработке, что позволяет создать высококачественную базу знаний, необходимую для построения современных систем интеллектуального анализа текстов.

Каждая статья Википедии имеет заголовок, состоящий из одного или нескольких слов, и тела, описывающего значение термина-заголовка. Все статьи Википедии связаны гипертекстовыми ссылками и образуют граф, обладающий свойствами сети документов. Кроме того, существуют специальные страницы для определения синонимов и списков значений многозначных слов. Таким образом, можно автоматически создать словарь терминов и определить их возможные значения. При этом, вычислив семантическую близость между статьями Википедии, можно оценить отношения между значениями терминов.

Так как граф Википедии обладает свойствами сети документов, то на основании произведенного анализа, выбор меры близости производился из класса локальных мер. В ходе диссертационной работы было замечено, что ссылки Википедии обладают различной степенью релевантности по отношению к семантической близости в зависимости от места в статье, где они использовались. На основании этого замечания был предложен способ взвешивания графа Википедии и вычисления семантической близости концепции во взвешенном безмасштабном графе с использованием теории нечетких множеств.

Оценка качества мер близости может производиться как с помощью корреляции с экспертными оценками, так и на основе влияния на высокоуровневые алгоритмы, использующие семантическую близость. Так как в рамках данной работы вычисление семантической близости является лишь вспомогательной задачей, то был выбран второй подход и найдена мера дающая наилучшие

результаты в контексте применения к задаче устранения лексической многозначности:

$$sim(A,B) = \frac{\sum\limits_{N \in n(A) \cap n(B)} [w(A,N) + w(B,N)]}{\sum\limits_{N \in n(A)} w(A,N) + \sum\limits_{N \in n(B)} w(B,N)} \ , \qquad (1)$$

где $n(X)$ — множество узлов, непосредственно связанных с узлом $X$. Эксперименты проводились на алгоритмах и тестовых множествах, описанных в третьей главе данной работы.

Также в третьем разделе второй главы описан процесс извлечения информации из Википедии, приводится обзор существующих работ по этой теме и обосновывается необходимость создания собственного анализатора Википедии. Так как использование Википедии является новым направлением, существующие системы извлечения информации из этой энциклопедии являются недостаточно гибкими для использования с методами, разработанными в диссертационной работе. Описание разработанного анализатора, представленное в данном разделе, показывает процесс построения словарей и основные трудности, возникающие при обработке Википедии. После обработки Википедии был получен словарь из 5.5 млн. терминов, соответствующих 2.8 млн. значений. При этом словарь содержит 290000 многозначных слов, а среднее количество значений равно 5,4.

В четвертом разделе приводится обзор работ, использующих Википедию для устранения лексической многозначности. Рассмотренные алгоритмы устранения лексической многозначности продолжают использоваться парадигмы, заданные в 90-х годах 20 века.

Результаты второй главы опубликованы в работах [2, 3, 7].

**В третьей главе** описываются три метода снятия лексической многозначности, основанные на семантической близости значений терминов Википедии. Каждый следующий метод устраняет недостатки предыдущего.

В первом разделе третьей главы приводится общий для всех методов процесс предварительной обработки текстов. На этом этапе текст разбивается на слова, определяются части речи слов, и производится поиск терминов из словаря Википедии, с учетом возможных словоформ. Лингвистическая обработка текста производится с помощью открытого пакета OpenNLP, алгоритмы которого основаны на методе максимальной энтропии.

Во втором разделе описан метод устранения лексической многозначности терминов, основанный на выборе значения, наиболее близкого к однозначному контексту. Аналогом этого метода может считаться широко известный алгоритм Леска, отличающийся способом определения отношений между значениями терминов. Метод, предложенный в данном разделе, является самым простым с вычислительной точки зрения, однако не позволяет достичь точности последующих методов.

Для определения наиболее близкого к контексту значения используется семантическая близость между значениями терминов. Так как контекст может состоять из нескольких терминов, в данном разделе предлагается способ вычисления семантической близости между множествами значений. Результаты предложенного метода зависят от однозначного контекста, которого может быть недостаточно для принятия правильного решения. Этот существенный недостаток, исправляется в последующих методах, где задача устранения лексической многозначности сводится к задаче максимизации.

Также в данном разделе описываются три тестовые коллекции, которые используются для оценки качества всех методов, представленных в диссертационной работе. Первая коллекция разработана в Институте системного программирования РАН и основана на коллекции новостных документов и

научных статей. Во второй коллекции используются статьи самой Википедии. Оценка производится только для терминов, представленных в ссылках, а значения этих терминов описываются статьями на которые указывает соответствующая ссылка. Еще одна используемая коллекция была разработана исследователями новозеландского университета Уэйкато и размечена с помощью сервиса Amazon Mechanical Turk.

Следующие две секции диссертационной работы посвящены одному из перспективных, но мало изученных направлений в области устранения лексической многозначности — использованию методов оптимизации.

В третьей секции показано как решать задачу разрешения лексической многозначности с помощью скрытой модели Маркова. Имея последовательность наблюдаемых переменных $\tau = (t_1, \ldots, t_n)$ (терминов текста), необходимо найти наиболее вероятную последовательность состояний модели $\mu = (m_1, \ldots, m_n)$ (соответствующих значений терминов). Марковская модель задает ограничения на зависимость между переменными, а задача разрешения лексической многозначности сводится к поиску наиболее вероятной последовательности значений:

$$\hat{\mu} = \arg\max_{\mu} \left( \prod_{i=1}^{n} P(m_i | m_{i-k:i-1}) \cdot P(t_i | m_i) \right),$$

где $(m_{i-k:i-1})$ — сокращенная запись для последовательности переменных $(m_{i-k}, \ldots, m_{i-1})$.

Основной трудностью применения марковских моделей к обработке естественного языка является оценка параметров модели — вероятностей, входящих в произведение. В работе предложен способ оценки параметров модели с использованием Википедии как размеченного корпуса. Для преодоления проблемы разреженности языка предложено комбинировать статистические оценки и значения семантической близости концепций Википедии. Для этого ведена эвристика.

**Эвристика 1.** *Вероятность значения $m_i$, при условии предыдущего контекста $m_{i-h}, \ldots, m_{i-1}$ пропорциональна линейной комбинации (а) близости значения к контексту и (б) априорной вероятности этого значения.*

$$P(m_i|m_{i-h}, \ldots m_{i-1}) = \hat{P}(m_i|m_{i-h}, \ldots m_{i-1}) = \\ = \alpha \cdot (sim(m_i; m_{i-h}, \ldots m_{i-1}) + \beta \cdot P(m_i)) \quad (2)$$

Эвристика позволяет дать экспертные оценки параметров марковских моделей любого порядка. Для задачи разрешения лексической многозначности такой результат представлен впервые.

После определения всех параметров модели задача максимизации решается с помощью алгоритма Витерби. Чтобы сократить время работы алгоритма для моделей высокого порядка, было выдвинуто предположение:

**Эвристика 2.** *для задачи устранения лексической многозначности, наиболее вероятный путь до состояния $m_i$ зависит только от n последних значений наиболее вероятного пути до состояния $m_{i-1}$ .*

Эта эвристика позволила существенно увеличить скорость обработки текстов, при этом потери в точности оказались несущественны.

Метод основанный на марковской модели показывает результаты лучше, чем первый метод, так как не зависит от количества и местоположения однозначных терминов. Однако марковская модель не позволяет полностью описать структуру дискурса, состоящего из нескольких тем, поэтому не дает существенного прироста точности разрешения лексической многозначности.

В разделе 3.4 для моделирования структуры текста предложено обобщение марковской модели на случай множества независимых марковских процессов, и показан способ применения этого обобщения к задаче устранения лексической многозначности.

Также как и в классической марковской модели, исследуются стационарные процессы, моделируемые марковскими цепями некоторого порядка $m$. Основное различие между классической марковской моделью и моделью представленной в данном разделе заключается в том, что текущее состояние может стать частью одной из существующих цепей, а может сформировать новую цепь, независимую от других цепей. Для этого в определение модели вводится дополнительная вероятность, что некоторое состояние принадлежит одной или нескольким из существующих цепей.

$$\hat{\mu} = \arg\max_{\mu} \left( \prod_{i=1}^{n} P(m_i \in \lambda) P(m_i|\lambda) P(t_i|m_i) \right),$$

где $\lambda$ — некоторое подмножество существующих цепей. Таким образом, задачей становится поиск наиболее вероятной последовательности состояний и их разделения на отдельные цепи. При этом модель предполагает ограничения на цепи, к которым может присоединиться текущее состояние.

**Предположение 1.** *Вероятность того, что текущее состояние $m_k$ принадлежит цепи $\mathcal{L}$, зависит только от конечного числа предыдущих состояний $m_{t_1}$, $m_{t_2}$, ..., $m_{t_h}$ цепи $\mathcal{L}$, где $t_i < k, \forall i = \overline{1,h}$.*

В дальнейшем будем называть состояния $m_{t_1}$, $m_{t_2}$, ..., $m_{t_h}$, описанные в сделанном предположении активными; цепи, содержащие активные состояния, будем называть активными цепями. Количество активных состояний определяет порядок обобщенной модели.

**Предположение 2.** *Для различных состояний $m_i$, $m_j$ и $m_k$, событие «$m_i$ и $m_k$ принадлежат одной цепи» (обозначаемое $\widehat{m_i m_k}$) является независимым от события «$m_j$ и $m_k$ принадлежат одной цепи». То есть для $i \neq j$, $i \neq k$ и $j \neq k$:*

$$P(\widehat{m_i m_k} \text{ and } \widehat{m_j m_k}) = P(\widehat{m_i m_k}) \cdot P(\widehat{m_j m_k}) \ .$$

Эти ограничения позволяют выразить вероятность $P(m_i \in \lambda)$ через вероятности $P(\widehat{m_i m_k})$ и существенно сократить сложность вычисления наиболее вероятной последовательности состояний.

Далее в тексте представлен алгоритм для нахождения наиболее вероятной последовательности состояний обобщенной модели, являющийся модификацией алгоритма Витерби. Алгоритм линейно зависит от количества обрабатываемых терминов, однако на каждом шаге производит сравнение всех возможных цепей. Число таких цепей описывается последовательностью Белла $B_n$, где n — число активных состояний. Таким образом, сложность алгоритма растет в экспоненциальной зависимости от порядка обобщенной модели.

Далее рассмотрен частный случай модели, не учитывающий зависимость переменных, относящихся к одной цепи. Для этого случая предложена оптимизация алгоритма, основанная на теореме, показывающей что:

1. если на некотором шаге алгоритма для фиксированных значений переменных состояний вероятность объединения этих состояний в одну цепь больше вероятности создания нескольких цепей, и если переменные состояния в итоговом наиболее вероятном пути будут иметь те же значения, то они также будут принадлежать одной цепи;

2. на любом шаге алгоритма, для любого количества различных активных цепей, существуют такие значения текущего состояния $m_i$ и состояний, входящих в эти цепи, что вероятность случая, где все $n$ цепей различны, а часть из них, возможно, объединяется состоянием $m_i$, меньше чем вероятность случая, где все эти состояния образуют одну цепь.

На основании рекурсивной структуры алгоритма, теорема сформулирована для промежуточных состояний $m_i$. Чертой сверху обозначены компоненты, составляющие одну цепь.

**Теорема 1.**

*1.* $\forall m,\ \forall \mathcal{L} \in \Lambda,\ \mathcal{L} = \overline{\mathcal{L}_1, \ldots, \mathcal{L}_n},\ \forall n \in \mathbb{N},\ \forall k = 1..n$*:*

$$P(\overline{\mathcal{L}m}) \geq P(\mathcal{L}, m) \Rightarrow P(\overline{\mathcal{L}m}) \geq \prod_{i=1}^{k-1} P(\mathcal{L}_i) \times P(\overline{\mathcal{L}_k m}) \times \prod_{i=k+1}^{n} P(\mathcal{L}_i)\ ,$$

$$P(\mathcal{L}, m) \geq P(\overline{\mathcal{L}m}) \Rightarrow P(\mathcal{L}, m) \geq \prod_{i=1}^{k-1} P(\mathcal{L}_i) \times P(\overline{\mathcal{L}_k m}) \times \prod_{i=k+1}^{n} P(\mathcal{L}_i)\ .$$

*2.* $\forall \lambda \in \Lambda,\ \lambda = \{\mathcal{L}_1, \ldots, \mathcal{L}_n\},\ \exists m$*:*

$$P(\overline{\mathcal{L}_1 \ldots \mathcal{L}_n m}) \geq P(\mathcal{L}_{i_1}, \ldots\ \mathcal{L}_{i_k}, \overline{\eta m}),\ \textit{где}\ \eta = \{\mathcal{L}_{i_{k+1}}, \ldots, \mathcal{L}_{i_n}\}.$$

Эта оптимизация позволяет уменьшить время вычисления алгоритма для практических задач, таких как задача разрешения лексической многозначности, так как для нахождения наиболее вероятного пути в данном случае, на каждом шаге алгоритма необходимо сравнивать только вероятности существующих активных цепей.

В следующей части раздела приведен способ оценки параметров обобщенной модели, для применения к задаче разрешения лексической многозначности. В дополнение к вероятностям перехода и наблюдения, которые можно оценить по аналогии с классической марковской моделью, необходимо оценить вероятность принадлежности состояния к существующим цепям. Последняя вероятность, на основании введенных предположений, выражается через вероятность события, что два состояния принадлежат одной цепи. В терминах задачи устранения лексической многозначности это означает, что два значения относятся к одной лексической цепи. Для оценки этой вероятности вводится дополнительная эвристика.

**Эвристика 3.** *Вероятность события, что два значения принадлежат одной цепи, является функцией от семантической близости:*

$$P(\widehat{m_1 m_2}) = \phi(sim(m_1, m_2))\ . \tag{3}$$

Далее предлагается метод, для автоматической оценки параметра $P(\widehat{m_1 m_2})$, основанный на кластеризации графа значений. Оценка производилась на неразмеченной коллекции новостных статей. Каждый документ коллекции был представлен в виде взвешенного графа, где вершинами являлись значения терминов документа, а ребра между вершинами имели вес равный семантической близости этих значений. После этого к каждому полученному графу был применен алгоритм кластеризации. Два значения считались принадлежащими одной лексической цепи, если соответствующие вершины графа принадлежали одному кластеру. Таким образом был получен тренировочный корпус для оценки функции $\phi$, где пары концепций, принадлежащих одному кластеру, являлись положительными примерами, а отрицательные примеры формировались парами концепций, принадлежащих одному документу, но разным кластерам. Было получено множество из 137,324 положительных и 859,076 отрицательных примеров, на их основе из пространства ступенчатых функций с шагом 0.01 выбиралась функция $\phi$. Заметим, что никакая ручная обработка текста не требовалась.

В конце третьей главы описываются проведенные эксперименты, приводится сравнение точности и полноты алгоритмов и показывается, что алгоритм, основанный на обобщенной марковской модели, дает результаты, превосходящие все, представленные в современной литературе. Частный случай модели, для которого предложена оптимизация, в рамках данной задачи показывает схожие результаты.

Результаты третьей главы опубликованы в работах [2, 5, 6, 8].

**В Заключении** диссертационной работы перечисляются ее основные результаты.

## Основные результаты работы

1. Предложен подход к разрешению лексической многозначности терминов на основе сети документов Википедии.

2. Предложен метод измерения семантической близости узлов взвешенной сети документов.

3. В рамках предложенного подхода разработаны и формально обоснованы методы разрешения лексической многозначности терминов на основе структурной и текстовой информации сетей документов с использованием: контекста из однозначных терминов; марковской модели высокого порядка; обобщения марковской модели.

4. Для экспериментального подтверждения эффективности предложенных методов разработан прототип системы разрешения лексической многозначности терминов Википедии и проведены эксперименты, доказывающие эффективность предложенных методов.

5. Разработанный прототип был использован в качестве основы для создания в Институте системного программирования РАН системы анализа текстов Texterra.

## Список публикаций

[1] *Denis Turdakov.* Recommender System Based on User-generated Content // Proceedings of the SYRCODIS 2007 Colloquium on Databases and Information Systems. — 2007.

[2] *Denis Turdakov, Pavel Velikhov.* Semantic Relatedness Metric for Wikipedia Concepts Based on Link Analysis and its Application to Word Sense Disam-

biguation // Proceedings of the SYRCODIS 2008 Colloquium on Databases and Information Systems. — 2008.

[3] *Dmitry Lizorkin, Pavel Velikhov, Maxim Grinev, Denis Turdakov.* Accuracy estimate and optimization techniques for SimRank computation // *Proceedings of the 34rd International Conference on Very Large Data Bases.* — 2008. — Vol. 1, no. 1. — Pp. 422–433.

[4] *Maria Grineva, Maxim Grinev, Denis Turdakov et al.* Harnessing Wikipedia for Smart Tags Clustering // KASW: International Workshop on «Knowledge Acquisition from the Social Web». — 2008.

[5] *Д. Ю. Турдаков, С. Д. Кузнецов.* Автоматическое разрешение лексической многозначности терминов на основе сетей документов // *Программирование.* — 2010. — Vol. 36, no. 1. — Pp. 11–18.

[6] *Турдаков Денис.* Устранение лексической многозначности терминов Википедии на основе скрытой модели Маркова // XI Всероссийская научная конференция «Электронные библиотеки: перспективные методы и технологии, электронные коллекции». — 2009.

[7] *Dmitry Lizorkin, Pavel Velikhov, Maxim Grinev, Denis Turdakov.* Accuracy estimate and optimization techniques for SimRank computation // *The VLDB Journal.* — 2009. `http://dx.doi.org/10.1145/1453856.1453904`.

[8] *Denis Turdakov, Dmitry Lizorkin.* HMM Expanded to Multiple Interleaved Chains as a Model for Word Sense Disambiguation // Proceedings of the 23rd Pacific Asia Conference on Language, Information and Computation. — Hong Kong: City University of Hong Kong, 2009. — December. — Pp. 549–558.