МОСКОВСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ
имени М.В.ЛОМОНОСОВА

На правах рукописи

Лунев Кирилл Владимирович

Теоретико-графовые алгоритмы выявления семантической близости между понятиями на основе анализа наборов ключевых слов взаимосвязанных объектов

Специальность 05.13.17 —
«Теоретические основы информатики»

Автореферат
диссертации на соискание учёной степени
кандидата физико-математических наук

Москва — 2021

Работа выполнена на кафедре вычислительной математики механико-математического факультета Московского Государственного Университета им. М.В.Ломоносова.

Научный руководитель: **Васенин Валерий Александрович**
доктор физико-математических наук, профессор

Официальные оппоненты: **Серебряков Владимир Алексеевич,**
доктор физико-математических наук, профессор, главный научный сотрудник Федерального Исследовательского Центра «Информатика и Управление» Российской академии наук.

Кузнецов Сергей Олегович,
доктор физико-математических наук, профессор, заведующий лабораторией интеллектуальных систем и структурного анализа факультета компьютерных наук Национального исследовательского университета «ВШЭ».

Махортов Сергей Дмитриевич,
доктор физико-математических наук, доцент, заведующий кафедрой программирования и информационных технологий факультета компьютерных наук Воронежского государственного университета.

Защита состоится «15» сентября 2021 г. в 16 часов 45 минут на заседании диссертационного совета МГУ.05.01 ФГБОУ ВО «Московский государственный университет имени М. В. Ломоносова» по адресу: Российская Федерация, 119234, Москва, ГСП-1, Ленинские горы, д. 1, ФГБОУ ВО «Московский государственный университет имени М. В. Ломоносова», Механико-математический факультет, ауд. 14-08.
E-mail: vasenin@msu.ru
Диссертация находится на хранении в отделе диссертаций научной библиотеки МГУ имени М.В. Ломоносова (Ломоносовский просп., д. 27). С диссертацией в электронном виде и с информацией о регистрации участия в защите в удаленном интерактивном режиме можно ознакомиться на сайте ИАС «ИСТИНА»: https://istina.msu.ru/dissertations/382381731/.

Автореферат разослан «__» _____ 2021 г..

Ученый секретарь
диссертационного совета
МГУ.05.01,
к.ф.-м.н. М. Кривчиков Кривчиков Максим Александрович

Общая характеристика работы

Актуальность темы. В качестве предмета исследования и анализа в диссертации выступают объекты наукометрической информационно-аналитической системы, которые описываются наборами ключевых слов. Кроме того, объекты такой системы связаны между собой различными отношениями, например, для научной публикации это может быть список соавторов, для научного работника - список проектов, в выполнении которых он принимал участие, или список конференций, в которых он принимал участие. Публикации, персоналии, научные проекты и конференции в данном примере являются объектами и, следовательно, могут иметь собственные наборы ключевых слов.

Побудительным мотивом и конечной целью исследований, результаты которых представлены в настоящей диссертации, является создание интеллектуального программного модуля, встраиваемого в наукометрическую информационно-аналитическую систему, способного по имеющимся в системе ключевым словам и определенным связям между ними выявлять семантическую информацию и с ее помощью решать задачи информационного поиска и классификации. Следует также отметить то обстоятельство, что зачастую информационные системы не обладают большим объемом данных для анализа, что делает затруднительным качественное семантическое сравнение объектов. Однако и в таких системах необходимо уметь точно определять релевантную пользователю информацию. Как следствие, важным требованием к разрабатываемому модулю является его способность эффективно работать в условиях ограниченного объема входной информации.

Кроме того, отсутствие достаточного объема данных в реальных системах показывает актуальность и востребованность на практике исследования, результаты которого представлены в настоящей диссертации.

Целью диссертационной работы является исследование и разработка математических моделей, алгоритмов и программных средств интеллектуального анализа наборов ключевых слов, характеризующих объекты в наукометрических интеллектуальных системах, с использованием методов из теории графов и дополнительной информации онтологического характера об объектах в системе. Такая деятельность соответствует областям исследования, отмеченным в пп. 1, 2, 5, 9 Паспорта специальности 05.13.17 – теоретические основы информатики.

Требования к разрабатываемой системе. Согласно стандарту ГОСТ Р ИСО/МЭК 9126-93 к качеству разрабатываемой системы интеллектуального анализа объектов информационной системы предъявляются следующие требования:

- широкие функциональные возможности;
- надежность;
- эргономичность;
- эффективность;
- сопровождаемость;

– мобильность.

Более подробно изложенные выше пункты определены в приложении А. Описанные в данном приложении характеристики определяют отличительные стороны решаемой в настоящей диссертации задачи от известных работ, связанных с выделением семантической информации между объектами информационных систем. Существующие системы в большинстве своем опираются на обилие входных данных, к числу которых относятся:

– текстовая информация, а именно - аннотации, заголовки, полные тексты документов;
– общие объемы данных системы, которые характеризуются значительным количеством сущностей внутри системы и числом связей между ними.

В то же время, разрабатываемый программный комплекс является более гибким решением для систем, не обладающих большим объемом данных. Такие системы с одной стороны не содержат в себе огромного количества различных объектов. С другой стороны, о каждом из объектов известно минимальное количество информации - сущности таких систем должны лишь обладать описывающим их набором ключевых слов, либо быть соединены внутренними связями с сущностями, которым набор ключевых слов ассоциирован. Кроме того, разработанные подходы позволяют получать узконаправленные семантические модели для конкретной области знаний. Ручной труд при внедрении таких систем сводится к минимуму.

В разделе 1.1.4 содержится краткое изложение мотивации предъявленных требований к системе, разрабатываемой в данной работе. Кроме того, представлены недостатки существующих методов решения подобных задач. Описываются проблемные места, которые не позволяют применять эти подходы к некоторому классу систем. В конечном итоге выделяется **специфика** разрабатываемого комплекса, отличающего его от известных аналогов. На основе анализа перечисленных требований была разработана методология решения поставленной задачи.

В работе применяются методы анализа текстов на естественном языке, методы машинного обучения и программной инженерии. При изложении результатов диссертационной работы широко используется аппарат теории графов, а также математической логики и математической статистики.

Положения, выносимые на защиту. На защиту выносятся: обоснование актуальности, научная новизна, теоретическая и практическая значимость работы, а также следующие положения, которые подтверждаются результатами исследования, представленными далее в заключении диссертации.

1. Создание моделей и их программных реализаций вычисления уровня семантической близости между ключевыми словами интеллектуальной системы с учетом специфики этих систем.
2. Создание методов автоматической генерации обучающей выборки для определения семантически близких ключевых слов.

3. Создание модели и ее программной реализации для вычисления семантической близости между парой объектов информационно-аналитической системы по ассоциированным с ними наборами ключевых слов.
4. Решение востребованных практикой задач в рамках рассматриваемой информационно-аналитической системы. Среди таких задач выделяются следующие: поиск экспертов в различных областях научных знаний, кластеризация ключевых слов, определение тематической направленности объекта информационно-аналитической системы.

Научная новизна работы определяется тем, что автором разработаны новые алгоритмы определения семантической близости для пары ключевых слов, а также для пары наборов ключевых слов, описывающих объекты интеллектуальной наукометрической системы. Созданы уникальные методы автоматической генерации обучающей выборки, а также методы автоматической проверки качества работы программных реализаций алгоритмов определения семантически похожих ключевых слов и алгоритмов выявления кластеров близких понятий. Последнее обстоятельство важно, поскольку тестирование программ в данной предметной области требовательно к наличию специалистов, способных точно определить степень близости для пары объектов или понятий. Разработаны алгоритмы построения иерархических классификаторов научных направлений в автоматическом режиме, использующие исключительно наборы ключевых слов.

Важными особенностями указанных алгоритмов являются:

1. отсутствие необходимости больших объемов данных для обучения моделей с приемлемым уровнем качества;
2. возможность использования разработанных моделей для произвольных интеллектуальных систем, использующих ключевые слова для описания сущностей;
3. возможность применения к любым задачам, в которых объекты системы представляются в виде некоторого графа и имеется необходимость в классификации отношений между парой объектов;
4. небольшие человеческие трудозатраты для выставления экспертных оценок.

Проведена работа по уменьшению числа параметров системы, что делает разработанные модели и программные средства эргономичными и легкими для настройки.

Теоретическая значимость работы. Разработаны алгоритмы вычисления уровня семантической близости между ключевыми словами интеллектуальной системы, а также алгоритмы вычисления семантической близости между парой объектов информационно-аналитической системы. Доказана вычислительная сложность разработанных алгоритмов, подтверждающая их адекватность (соответствие) требованиям, предъявляемым к разрабатываемому программному комплексу.

Практическая значимость работы. Рассматриваемый в работе программный комплекс для анализа, обработки и поиска объектов интеллектуальных информационных систем по ключевым словам представляет собой самостоятельный инновационный продукт. Он может использоваться не только в системе, рассматриваемой в данной диссертации, но и в любой информационно-аналитической системе, объекты которой описываются наборами ключевых слов. Кроме того, разработанные автором методики обработки связей между объектами могут быть перенесены на другие задачи анализа взаимосвязанных объектов. Рассматриваемый программный модуль определения семантической близости между словами порождает словарь синонимов той области, на которой был обучен. Этот словарь может быть использован в самых разнообразных задачах информационного поиска и обработки естественного языка, и потенциально может привнести дополнительный полезный сигнал для моделей классификации, ранжирования и кластеризации текстовых или текстово-аннотированных объектов.

Методология исследования включает следующие характеризующие ее аспекты.

- **Концептуальные положения.**
  - Опора на наборы ключевых слов, ассоциированых с объектами информационной системы.
  - Возможность использования различных информационных объектов (источников) - НИР, публикации, патенты и т.п..
  - Наличие механизмов, позволяющих в автоматизированном режиме получать оценки адекватности полученных решений.
- **Модели, методы и средства достижения цели.**
  - Модели, реализующие концептуальные положения создания и развития на основе графовых представлений данных и эвристических алгоритмов над ними, работающие в условиях отсутствия строгого математического описания.
  - Методы машинного обучения, необходимые для улучшения качества определения семантических связей между объектами информационной системы;
- **Инструментальные средства.** Для разработки программных комплексов, решающих поставленные задачи, использованы открытые математические, графовые библиотеки, программные пакеты для обработки естественного языка, программные реализации моделей машинного обучения с отрытым исходным кодом.
- **Перечень и постановка задач, решение которых обеспечивает достижение цели.**
  1. Разработка графовой модели представления данных. Необходимо представить данные системы в виде множества графов, вершинами которых являются некоторые понятия (ключевые

слова/наборы слов/сущности системы), а ребрами - отношения между ними. Такие графы необходимы для вычисления различных характеристик для пар понятий. Решение задачи приводится в главе 2.

2. Разработка моделей определения семантической близости пары ключевых слов. Для этого используются построенные графы, разработанные подходы и технологии машинного обучения. Формируется набор количественных признаков и решающее правило, определяющее степень семантической близости по этому набору. Описание разработанных моделей приводится в главе 2.
3. Разработка моделей определения семантической близости пары наборов ключевых слов. Разработанные модели используют различные графовые представления, подходы и модели, рассмотренные в предыдущих пунктах. Решению этой задачи посвящена глава 3.
4. Апробация разработанных моделей. Используя функцию близости наборов ключевых слов и отношения между сущностями системы, решаются прикладные задачи определения семантической близости пары сущностей. Этой задаче посвящается глава 4.

Для достижения поставленных целей и удовлетворения описанных выше требований были рассмотрены различные методы решения, их преимущества и недостатки. По окончании поиска была составлена методология исследования, наиболее подходящая поставленным в настоящей диссертации задачам в рамках имеющихся особенностей и ограничений в наборов исходных данных. Подробная мотивация выбранной методологии описывается в разделе 1.2.

Соответствие диссертации паспорту научной специальности. Полученные в диссертации результаты соответствуют паспорту специальности 05.13.17 — теоретические основы информатики (физико-математические науки). Теоретические основы информатики – специальность, включающая исследования процессов создания, накопления и обработки информации; исследования методов преобразования информации в данные и знания; создание и исследование информационных моделей, моделей данных и знаний, методов работы со знаниями, методов машинного обучения и обнаружения новых знаний; исследования принципов создания и функционирования аппаратных и программных средств автоматизации указанных процессов. Области исследования:

1. Исследование, в том числе с помощью средств вычислительной техники, информационных процессов, информационных потребностей коллективных и индивидуальных пользователей.
2. Исследование информационных структур, разработка и анализ моделей информационных процессов и структур.

5. Разработка и исследование моделей и алгоритмов анализа данных, обнаружения закономерностей в данных и их извлечениях разработка и исследование методов и алгоритмов анализа текста, устной речи и изображений.
9. Разработка новых интернет-технологий, включая средства поиска, анализа и фильтрации информации, средства приобретения знаний и создания онтологии, средства интеллектуализации бизнес-процессов.

Апробация работы. Результаты, изложенные в диссертации, докладывались на следующих семинарах и конференциях:

1. на всероссийской конференции с международным участием «Знания–Онтологии–Теории (ЗОНТ-2015)»,
2. на международной конференции «Ломоносовские чтения» (2014, 2016, 2018, 2019),
3. на механико-математическом факультете МГУ имени М.В. Ломоносова на семинаре «Проблемы современных информационно-вычислительных систем» под руководством д.ф.-м.н., проф. В.А. Васенина (2013, 2015, 2017, 2018),
4. на семинаре в ВЦ РАН под руководством д.ф.-м.н., проф. В.А.Серебрякова (2018).

Публикации. По теме диссертации опубликовано 7 работ, в том числе одна в зарубежном издании. Из них 5 научных статей опубликованы в журналах Scopus и RSCI, что соответствует требованиям, предусмотренным пунктом 2.3 Положения о присуждении учёных степеней в МГУ.

По результатам апробации модели и реализующих её программных средств получен акт о внедрении.

Объем и структура работы. Диссертация состоит из введения, четырёх глав, заключения и двух приложений. Полный объём диссертации составляет 195 страницы, включая 16 рисунков и 16 таблиц. Список литературы содержит 130 наименований.

Содержание работы

В **первой главе** рассматриваются известные подходы к решению задач в направлении выявления семантической информации, проводится их анализ, выявляются недостатки. После чего автором описывается использованная в настоящей диссертации методология. Эта методология призвана решить недостатки в существующих моделях определения смысловой близости объектов информационных систем по их ключевым словам.

В первом разделе проводится библиографический обзор существующих методов, выделяются сильные и слабые стороны моделей. Анализу подвергаются работы, в которых исследуются семантическая близость для пар слов и пар коротких предложений естественного языка. В частности, обозреваются работы, связанные с анализом ключевых слов.

Второй раздел содержит в себе выводы из обзора и указывает на недостатки рассмотренных методов и на причины неприменимости этих методов к решаемым в данной работе задачам.

В заключающем разделе главы вводится методология, используемая автором в последующих главах для разработки моделей семантической близости. Показывается целесообразность и необходимость рассмотренной методологии, ее преимущества в сравнении с уже существующими. Стоит также отметить, что отдельные пункты методологии представляют собой самостоятельный интерес.

Во **второй главе** подробно описываются разработанные автором эвристические модели, алгоритмы и реализующее их программное обеспечение для определения семантической близости пары ключевых слов. Разработанные автором на основе проведенных исследований с использованием полученных в их ходе эвристических соображений модели определения семантической близости, являются важным результатом деятельности в рамках подготовки настоящей диссертации.

Задача, решение которое представлено в данной главе, имеет следующую математическую постановку. Дано множество объектов системы D и множество ключевых слов W. Каждый элемент $d_i \in D$ представлен *набором ключевых слов* $t_i \in T$, состоящим из k_i ключевых слов из множества W:

$$d_i \to t_i = (w_{i_1}, w_{i_2}, ..., w_{i_{k_i}}).$$

Здесь T - множество всех наборов ключевых слов или *коллекция ключевых слов*. Множество T состоит из *уникальных наборов ключевых слов*. Наборы ключевых слов совпадают, если существует перестановка, отождествляющая эту пару наборов. Кроме того, между объектами системы заданы отношения различных типов, определенные направленным графом $G_{rel} = (D, R)$, в котором D - множество вершин (объекты системы), а R - ребра, определяющие отношения. Ребра этого графа помечены типом отношения: $l : R \to T$, где T - множество возможных типов. Множества (D, T, W, G_{rel}) задают модель информационной аналитической системы.

В рамках данной главы необходимо разработать модель определения семантической близости и в ее рамках такую функцию близости $f : W \times W \to [0, 1]$, бóльшие значения которой означали бы сильный уровень смысловой близости между ключевыми словами.

Для определения семантической близости вводятся вспомогательные графы, вершинами которых являются ключевые слова, а ребра указывают на некоторые свойства пары ключевых слов. В самом простом виде таким свойством может являться факт причастности пары ключевых слов одному набору. Кроме этого существуют более сложные и неявные связи между парами ключевых слов. Примером такой связи может служить *контекстная близость*, определенная в главе 2. Согласно введенной модели, такая близость устанавливается между парой ключевых слов w_a и w_b в том случае, если существует большое число ключевых слов, которые входят одновременно и в наборы, содержащие слово w_a, и

в наборы, содержащие слово w_b. Этот факт означает, что если для пары слов существует достаточное количество общих «соседей по набору», то между ними проявляется такая контекстная связь.

В разделе 2.1 описывается модель близости, названная *WordContSim*. Начало раздела посвящается введению вспомогательного графа ключевых слов, после чего приводится описание и аргументация эвристических идей, заложенных в данную модель. Следующим этапом изложения является представление реализующих данную модель алгоритмов, асимптотических оценок вычислительной сложности и затрат по памяти для их программных реализаций.

В разделе 2.2 рассматриваются методы улучшения качества определения семантической близости для пары ключевых слов с помощью методов машинного обучения с учителем. В тексте подробно описывается сведение рассматриваемой задачи к задаче классификации, то есть к задаче определения целевой метки заданного объекта из заранее сформированного множества меток. Обучение с учителем подразумевают использование обучающей выборки - множества объектов, для которой известны истинные целевые метки. Эта выборка необходима для тренировки модели машинного обучения. Обученная модель машинного обучения применяется к произвольному объекту системы и выдает предсказанную целевую метку.

Сбор обучающего набора данных является обычно трудоемкой и дорогой работой: как правило для процесса требуется как минимум несколько экспертов в направлении, для которого собираются тренировочное множество. Более того, для некоторых задач, в число которых входит и задача определения семантической близости пар ключевых слов, является сложным даже правильно составить инструкцию для экспертов для правильной оценки примеров. Это происходит по причине того, что семантическая близость - субъективная величина и сильно зависит от области применения, контекста, человека и задачи. Например, слово «стол» может быть близким по смыслу для слова «стул» из бытовых соображений. В то же время если бы эти два слова были синонимами (а следовательно взаимозаменяемыми) в информационной системе, представляющей интернет-магазин, продающий мебель, то имела бы место ситуация, когда пользователь ищет один из этих предметов, а в поисковой выдаче получает другой, что является недопустимым. Рассмотренные в следующих далее разделах алгоритмы сбора обучающего множества призваны разрешить обозначенные проблемы.

Автором настоящей диссертации были разработаны два метода получения обучающих примеров без задействия в процесс экспертов. Первый из них включает в себя набор эвристических алгоритмов создания обучающей выборки. Второй является полностью автоматическим методом и строится исключительно при помощи описанных в предыдущих главах графовых моделей представления данных. Важнейшим преимуществом этого метода является его универсальность и применимость не только к задачам определения семантической близости, но и к любым другим задачам, в которых объекты системы представляются в виде некоторого графа и имеется необходимость в классификации отношений между

парой объектов. Также универсальность проявляется в том, что выборка строится непосредственно по данным и не использует никакие внешние источники. Это позволяет определить отношение близости специфичные для конкретной системы. Другое преимущество метода в возможности использовать эффективные модели машинного обучения с учителем, вместо более слабых моделей без учителя, которые, к тому же представляется сложным валидировать без обучающих примеров.

В заключительном разделе главы приводятся результаты и выводы о проведенных автором исследованиях в направлении создания моделей семантической близости пары ключевых слов.

В **третье главе** представлены разработанные автором настоящей работы алгоритмы выявления смысловой близости между двумя наборами ключевых слов. Этот уровень определяет уровень близости пары объектов интеллектуальной системы, с которыми были ассоциированы данные наборы ключевых слов. Семантическая близость между наборами ключевых слов определяется с использованием введенной в предыдущей главе методах определения семантической близости между парой ключевых слов.

В разделе 3.1 описываются модели и алгоритмы определения смысловой близости наборов ключевых слов.

Решаемая задача имеет следующую постановку. Дано множество документов D и множество ключевых слов W. Каждый элемент $d_i \in D$ представлен набором из k_i ключевых слов из множества $W : d_i = (w_{i,1}, w_{i,2}, ..., w_{i,ki})$. Необходимо разработать такую функцию близости $TupleSim : 2^W \times 2^W \to [0, 1]$, высокие значения которой означали бы высокий уровень смысловой близости между наборами и, следовательно, соответствующими документами системы.

К разрабатываемой метрике близости предъявляются следующие требования.

- Программная реализация должна быть способна вычислять значения метрики в режиме реального времени. Значительные задержки между отправленным пользователем запросом и полученным ответом не могут быть слишком долгими, иначе использование такой системы будет трудновыполнимым или неэффективным.
- Необходимо уметь вычислять уровень близости, в том числе, и для наборов, которые не имеют общих слов. Нулевое число общих слов еще не гарантирует отсутствие семантической близости между наборами. Например, слова одного набора могут быть синонимами, транслитерациями или переводами для слов другого набора. Для таких примеров уровень семантической близости должен быть высоким.

Основой для разрабатываемой функции определения близости являются методы и средства, введенные в предыдущей главе. Таким образом, в качестве базовых блоков определения близости наборов естественно положить идеи о близости между словами, из которых эти

наборы состоят. Формально это можно описать как $TupleSim(d_i, d_j) = TupleSim(WordSim(w_{i,1}, w_{j,1}), ..., WordSim(w_{i,ki}, w_{j,kj}))$.

В заключительном разделе главы приводятся преимущества и недостатки разработанных решений в направлении исследования моделей семантической близости пары наборов ключевых слов.

В **главе 4** рассматриваются решения востребованных на практике задач, необходимые для этого модели и алгоритмы, их программные реализации.

Раздел 4.1 посвящен разработанному автором методу семантической кластеризации ключевых слов. Механизмы кластеризация ключевых слов любой большой информационно-аналитической системы являются важным аналитическим инструментарием анализа данных. В первую очередь, они важны для систем, не обладающих необходимыми для получения надежного результата объемами данных. В таких системах, как правило, нет достаточного количества информации о ее объектах и о соответствующих этим объектам ключевых словах. Как следствие, значительная доля ключевых слов будет иметь низкие частоты встречаемости в системе, что делает затруднительным любые аналитические процедуры для таких слов. Преодолеть эти трудности можно путем решения задачи кластеризации. При таком подходе появляется возможность разбить все множество ключевых слов на кластера из семантически близких слов. После подобной предобработки можно отождествить исходные слова с соответствующими им кластерами, что предоставит более богатую статистику встречаемости для каждого ключевого слова.

В разделе описывается разработанный автором подход к решению данной задачи. В первых подразделах описывается новая модель кластеризации ключевых слов информационно-аналитической системы, основанная на мере *контекстной близости*, которая была введена ранее. С помощью такой меры конструируется граф, именуемый далее *контекстным графом*. После этого используются дополнительные эвристические соображения, позволяющие отбросить наименее значимые связи в этом графе. Такой граф получил название *усеченный контекстный граф*. Далее описывается метод кластеризации этого грава.

Задача кластеризации формулируется следующим образом. Задано множество ключевых слов рассматриваемой системы $W = w_1, \cdots, w_N$. Необходимо построить такое отображение $c : W \to \mathbb{N}$, ставящее в соответствие каждому ключевому слову системы номер кластера, которому это слово принадлежит. Условие, которое накладывается на эту функцию c, заключается в том, что все слова, попавшие в один кластер, должны быть семантически близкими понятиями, а слова из разных кластеров должны семантически различаться.

Заключительная часть раздела посвящена апробации программной реализации разработанной модели и выводам.

В разделе 4.2 ставится задача определения тематической направленности набора ключевых слов и дается разработанное автором настоящей диссертации решение. Для решения задачи вводится понятие степени абстрактности (далее

для краткости - абстрактности) ключевого слова. Под абстрактностью ключевого слова понимается степень общности значения этого слова.

Задача определения уровня абстрактности формулируется следующим образом. Рассматривается модель информационно-аналитической системы, введенная в начале настоящей главы. Необходимо по коллекции ключевых слов T и множества W ключевых слов данной коллекции разработать такую меру абстрактности

$$a_T : W \to \mathbb{R} + ,$$

чтобы бо̀льшим значениям меры соответствовали слова более широкого значения, а меньшим - слова более конкретного, обладающего определённой спецификой значения.

Далее в разделе 4.2 описывается решение задачи определения тематической направленности набора ключевых слов и дается разработанное автором настоящей диссертации решение. Задача формулируется следующим образом. Пусть дано множество документов D и множество ключевых слов W. Каждый элемент $d_i \in D$ представлен набором из k_i ключевых слов из множества $W : di = (w_{i,1}, w_{i,2}, ..., w_{i,ki})$. Необходимо разработать методы определения тематики для каждого документа коллекции. Под тематикой следует понимать название некоторой области знаний, дисциплины, специальности или направления. Определив такое слово для документа, можно с высокой вероятностью сказать о чем этот документ. Тематическое ключевое слово - это такое ключевое слово из W, которое может являться тематикой документа. Таким образом, необходимо из множества ключевых слов W выделить подмножество $T \subset W$ тематических ключевых слов и для каждого документа из коллекции определить соответствующее ему множество тематических ключевых слов, т.е. определить отображение $t : D \to 2^T$. Отметим, что принадлежность ключевых слова к тематике зависит от набора документов.

В множестве ключевых слов, которые приписаны документам из некоторой коллекции, можно выделить «тематические ключевые слова» - ключевые слова, семантическое значение которых является названием некоторого «широкого» направления. Примерами таких ключевых слов могут быть: «динамика», «статистика», «право», «численные методы». Выделение таких слов из набора является простой задачей для человека, но очень сложной для машины. В настоящем разделе описывается алгоритм, определяющий тематику набора ключевых слов, а также приводятся результаты его работы на реальных данных. Предложенные алгоритмы опираются на понятие абстрактности ключевого слова, а также на алгоритм её определения.

Заключительная часть раздела посвящена тестовым испытаниям программных реализаций модели абстрактности ключевого слова и модели определения тематической направленности набора ключевых слов.

В разделе 4.3 представлено описание моделей, алгоритмов и программных комплексов, решающих востребованную практикой задачу поиска эксперта. В рамках этой задачи необходимо по запросу, состоящему из набора ключевых

слов, находить наиболее релевантные объекты информационно-аналитической системы, наилучшим образом удовлетворяющие исходному запросу. Задача поиска эксперта является важной для наукометрических систем, поскольку позволяют существенно улучшить качество поиска необходимой пользователю информации. Модели и их программные реализации, описанные в настоящем разделе, прошли апробацию в системе ИАС «ИСТИНА».

Рассмотрим математически формализованную постановку задачи. Дано множество экспертов системы E и множество ключевых слов W. Каждый эксперт $e_i \in E$ представлен набором ключевых слов $t_i \in T$, состоящим из k_i ключевых слов из множества W наборов ключевых слов W_X и объектов информационной системы X, а также множество Q запросов к системе. Обозначим за W множество всех уникальных ключевых слов из всех наборов W_X. Каждый элемент $x_i \in X$ множества объектов ассоциирован с набором ключевых слов $W_i = \{w_{i_0}, w_{i_1}, ..., w_{i_{n_i}}\} \in W_X \in 2^W$. Точно также каждый запрос $q_j \in Q$ связан с некоторым набором ключевых слов $W_j = \{w_{j_0}, w_{j_1}, ..., w_{j_{n_j}}\} \in 2^W$. Необходимо определить меру близости пары запрособъект для каждого объекта и каждого запроса, т.е. функцию $f : Q \times X \to R$. Поскольку запросам и объектам единственным образом сопоставляются наборы ключевых слов, то задача сводится к определению меры близости на наборах: $f_w : 2^W \times 2^W \to \mathbb{R}$. Кроме того, необходимо разработать эффективный алгоритм, который, используя меру близости и некоторые дополнительные идеи, мог бы по запросу выдавать множество объектов, наиболее релевантных данному запросу.

В разделе 4.4 описывается построение тезауруса ключевых слов по коллекции наборов ключевых слов. Схема построения построения приводится на рисунке 1.

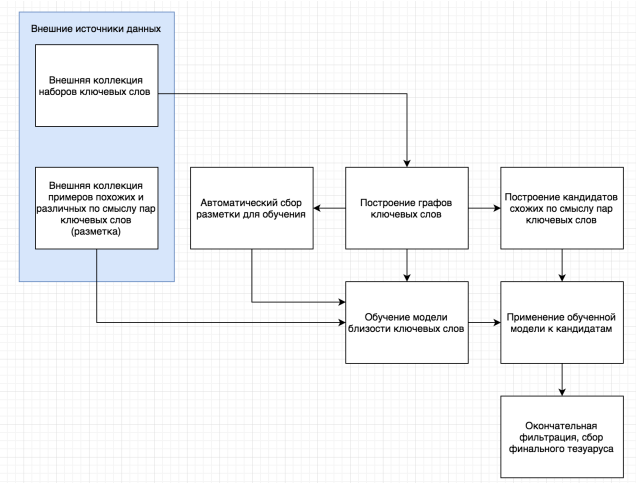

Рис. 1 — Схема построения словаря тезауруса ключевых слов

В качестве внешних данных используется коллекция наборов ключевых слов научных публикаций, собранная из сети Веб. Это коллекция, насчитывающая сотни тысяч наборов ключевых слов для русскоязычных статей и миллиона наборов на английском языке. С помощью этих наборов поочередно строится несколько графов близости. Одним из таких графов является, например, граф, в вершинах которого находятся ключевые слова, а в ребро означает причастность двух ключевых слов одному набору. Вес ребра тем больше, чем чаще пара слов встречается в одном наборе. Другим примером контекстного семантического графа является граф, в вершинах которого как и прежде стоят ключевые слова, а ребра указывают на наличие у этой пары общего ключевого слова, которое входит в некоторые наборы вместе и с первым словом из пары, и со вторым (но не с двумя сразу). Было показано, что ребра такого графа гораздо сильнее отражают семантическую близость, в то время как ребра первого графа служат для улучшения механизмов поисковых подсказок, потому что предлагают связанные, но более разнообразные слова для заданного.

В разделе 4.5 представлены программные реализации поисковых алгоритмов, использующих идеи семантической близости между объектами информационно-аналитической системы.

В разделе 4.6 приводится важное с практической точки зрения приложение реализаций моделей, описанных в главе 2. Разработанные автором подходы позволяют в значительной мере улучшить качество решения задачи поиска эксперта в системах, объекты которых задаются графами знаний. Методы хранения и обработки информации, основанные на использовании графов знаний, часто встречаются в современных информационно-аналитических системах. В этой связи важной является возможность применения разработанных в рамках диссертации подходов к системам такого типа.

В разделе 4.7 показаны результаты анализа разработанного автором программного модуля на соответствие заявленным требованиям.

В заключении главы представлены выводы, а также дальнейшее направление в применении разработанных программных комплексов в реальных приложениях.

В **заключении** настоящей диссертации представлены описание методологии, моделей, алгоритмов и программных средств интеллектуального анализа наборов ключевых слов, характеризующих объекты в интеллектуальной системе. В основе разработанных подходов лежат методы теории графов, анализа текстов на естественном языке, машинного обучения и программной инженерии.

Отмечается, что предложенные автором работы подходы к построению моделей семантического анализа могут применяться не только для наукометрических систем, но также и для других систем, объекты которой описываются набором слов естественного языка. Важным преимуществом разработанных моделей является их возможность эффективного внедрения в системы, не обладающие большим объемом исходных данных. Другое их достоинство заключается

в способности использовать произвольные связи между объектами системы для улучшения качества определения уровня семантической близости.

Заключение

В ходе работ по подготовке настоящей диссертации получены следующие **основные результаты**.

1. Разработан ряд моделей вычисления уровня семантической близости между ключевыми словами интеллектуальной системы. Программные реализации моделей позволяют вычислять такую близость, автоматически учитывая специфику системы, в которую модели внедряются. Проведены многочисленные тестовые испытания, подтверждающие высокий уровень полученных результатов. Получены аналитические оценки, характеризующие вычислительную сложность программных реализаций этих моделей.
2. Созданы уникальные методы автоматической генерации обучающей выборки для определения семантически похожих ключевых слов. Разработанные методы избавляют от больших человеческих трудозатрат для выставления экспертных оценок. Кроме того, алгоритмы генерации позволяют использовать методы машинного обучения с учителем для решения задачи семантической близости пары слов. Это обстоятельство в значительной мере упрощает подбор параметров модели и улучшает качество определения семантической близости.
3. Разработана модель и ее программная реализация для вычисления семантической близости между парой объектов. Для решения задачи были использованы дополнительные связи между сущностями системы. Модель протестирована на данных из ИАС «ИСТИНА». Показано, что модель удовлетворяет предъявляемым к ней требованиям.
4. На основе разработанных моделей решены важные, востребованные практикой задачи поиска экспертов в различных областях научных знаний, кластеризации ключевых слов, определения тематической направленности объекта информационно-аналитической системы. Программные реализации решений опробированы на данных из ИАС «ИСТИНА», включены в состав ее программного обеспечения, а также получили высокие оценки качества квалифицированных экспертов. Получен акт о внедрении.

Благодарности. Автор выражает огромную благодарность своему научному руководителю доктору физико-математических наук, профессору Валерию Александровичу Васенину за внимание к работе на всех ее этапах, редакторскую правку диссертационной работы и опубликованных статей, а также за ценные наставления, терпение и понимание.

Автор благодарит кандидата физико-математических наук, доцента С.А.Афонина за активное участие в постановке задач, за плодотворное обсуждение результатов работы, а также за техническую помощь в проведении исследований и экспериментов.

Работы автора по теме диссертации.

Научные статьи, опубликованные в журналах RSCI:

1. Лунев К. В. *К вычислению смысловой близости предложений* // Программная инженерия. — Москва, 2014. — No 8. — С. 30—39 (ИФ РИНЦ 0.267).
2. Лунев К. В., Афонин С. А. *Выявление тематических направлений в коллекции наборов ключевых слов* // Программная инженерия. — Москва, 2015. — No 2. — С. 29—39. (Луневым К.В. разработаны и протестированы программные реализации алгоритма определения уровня абстрактности ключевого слова, алгоритма определения тематической направленности ключевого слова, а также алгоритма выявления тематических направлений в наборе ключевых слов) (ИФ РИНЦ 0.267).
3. Васенин В. А., Зензинов А. А., Лунев К. В. *Использование наукометрических информационно-аналитических систем для автоматизации проведения конкурсных процедур на примере информационно-аналитической системы 'ИСТИНА'* / // Программная инженерия. — Москва, 2016. — Т. 7, No 10. — С. 472—480. (Луневым К.В. предложена модель поиска эксперта по запросу из набора ключевых слов) (ИФ РИНЦ 0.267).
4. Лунев К. В. *Графовые методы определения семантической близости пары ключевых слов и их применения к задаче кластеризации ключевых слов* // Программная инженерия. — Москва, 2018. — Т. 9, No 6. — С. 262—271 (ИФ РИНЦ 0.267).
5. Лунев К. В. *Алгоритм автоматизированной генерации обучающей выборки для решения задачи выявления семантической близости между парой ключевых слов методами машинного обучения* // Программная инженерия. — Москва, 2021 —Т. 12. No 6. — С. 283—293 (ИФ РИНЦ 0.267).

Другие публикации:

6. Васенин В. А., Афонин С. А., Зензинов А. А., Лунев К. В., Шачнев Д. А. *Механизмы системы «ИСТИНА» для интеллектуального анализа состояния и стимулирования хода выполнения проектов в сфере науки и высшего образования* // Научный сервис в сети Интернет: труды XXI Всероссийской научной конференции (23-28 сентября 2019 г., г. Новороссийск). — М.: ИПМ им. М.В.Келдыша, 2019. — С. 210-221. (Луневым К.В. создан алгоритм и его программная реализация для определения семантической близости отдельных ключевых слов или их наборов).

7. V. Vasenin, K. Lunev, S. Afonin, D. Shachnev. *Methods for intelligent data analysis based on keywords and implicit relations: The case of "ISTINA" data analysis system* // Proc. of the International Conference Actual Problems of Systems and Software Engineering (APSSE 2019). — IEEE Conference Publications. — США, 2019. — с. 151–155. (Луневым К.В. предложены модель определения уровня семантической близости для пары ключевых слов WordContSim, а также алгоритм определения степени абстрактности ключевого слова).