УДК 004.9

Ю. В. Леонова, **А. М. Федотов**

Институт вычислительных технологий СО РАН
пр. Акад. Лаврентьева, 6, Новосибирск, 630090, Россия

Новосибирский государственный университет
ул. Пирогова, 2, Новосибирск, 630090, Россия

E-mail: juli@ict.nsc.ru, fedotov@sbras.ru

ИССЛЕДОВАНИЕ НАУЧНЫХ СВЯЗЕЙ НА ОСНОВЕ АНАЛИЗА ДИССЕРТАЦИОННЫХ РАБОТ *

Приводится описание моделей и методов для анализа диссертаций и авторефератов с целью изучения структуры научных связей ученого (научное окружение ученого), структуры и динамики развития научных коллективов (научные школы), статистического исследования текста диссертаций. Такие исследования дают возможности изучать и оценивать тенденции развития различных научных направлений, идентифицировать персоны, научные центры и организации, научные школы, устанавливать взаимосвязи между отдельными сообществами.

Ключевые слова: извлечение информации, сущность, отношение, именованная сущность, выделение фактов, выделение отношений, информационная модель.

Введение

Целью данной работы является описание подходов, моделей и методов для изучения связей, возникающих в научных сообществах, в рамках которых осуществляется научная деятельность, на основе анализа текстов диссертаций и авторефератов. Научное сообщество понимается как совокупность исследователей-профессионалов, объединенных вокруг единой цели, научной школы или направления, и представляет собой сложную систему, в которой действуют как отдельные ученые, так и разнообразные институты, общественные организации, неформальные группы и т. д. Изучение связей в научном сообществе основано на решении следующих задач: статистическое исследование текста диссертаций, исследование структуры научных связей ученого (научное окружение ученого), исследование структуры и динамики развития незримых научных коллективов (научные школы). Такие исследования дают возможности изучать и оценивать тенденции развития различных научных направлений, идентифицировать персоны, научные центры и организации, научные школы, устанавливать взаимосвязи между отдельными сообществами.

* Работа выполнена при частичной поддержке РФФИ (проекты № 12-07-00472, 13-07-00258), президентской программы «Ведущие научные школы РФ» (грант НШ 6293.2012.9).

Леонова Ю. В., Федотов А. М. Исследование научных связей на основе анализа диссертационных работ // Вестн. Новосиб. гос. ун-та. Серия: Информационные технологии. 2014. Т. 12, вып. 1. С. 34–49.

ISSN 1818-7900. Вестник НГУ. Серия: Информационные технологии. 2014. Том 12, выпуск 1

Проблемы выделения фактографической информации из текстов диссертаций и авторефератов

Значительную часть данных о связях в научных сообществах можно получить в результате анализа информации из разобщенных фактов, содержащихся в текстах диссертаций и авторефератов. Объектом анализа является текст диссертации или автореферата, т. е. текстовый документ в электронной форме, который содержит некоторую информацию. Как правило, тексты:

- неструктурированны – из текста не выделены какие-либо значимые характеристики, вся информация о тексте заключена в самом тексте;
- неоднородны – тексты не являются одинаковыми. Тексты могут относиться к разным предметным областям, могут иметь различную внутреннюю структуру и форматирование. Кроме того, значения одинаковых слов, входящих в разные тексты, может быть различно (например, значение слова *символ* в тексте про языки программирования и про театральное искусство).

Извлечение фактографической информации (понятие «факт» будет описано далее) включает обязательный процесс структуризации входной информации из документа. Структуризация при этом осуществляется через выделение экземпляров информационных объектов определенного типа, информация о которых имеется в документе, и заполнение их реквизитов.

Диссертация является формальным квалификационным сочинением и ее оформление должно соответствовать всем правилам, которые устанавливает ГОСТ [1] для диссертации, что обеспечивает фиксированную структуру данного типа документов и позволяет реализовать методы автоматического извлечения фактографической информации из диссертаций. Процесс автоматического структурирования текстовой информации частично [2] заменяет экспертный процесс выделения фактографической информации и объектов, выполняемый вручную.

Рассмотрим структуру диссертации. Титульный лист диссертации рекомендуется оформлять в соответствии с ГОСТом. На титульном листе сверху вниз указываются: место выполнения диссертации; фамилия, имя, отчество соискателя; название работы; специальность, которой соответствует диссертация; сведения о том, на соискание какой ученой степени и в какой отрасли науки представляется работа; ученая степень, ученое звание, фамилия, имя, отчество научного руководителя (консультанта); город и год написания труда. Если диссертация выполнена под руководством двух научных руководителей (консультантов), то на титульном листе приводятся сведения о них обоих. Текст работы включает следующие элементы: введение, практические рекомендации, выводы и список литературы, а также обязательные главы (обзор литературы, материалы и методики, собственные результаты, обсуждения и заключения).

Основные проблемы, возникающие при анализе данных, изложенных в текстах на естественном языке, заключаются в невозможности использовать исключительно формальные признаки, формализуемые текстовыми последовательностями и подстроками [1]. Это заставляет оперировать объектами (сущностями), отсутствующими в тексте в явном формализованном виде, но описанными автором и несущими собственно реальное значение для решаемых задач. В конкретной формулировке перед нами стоит задача восстановления отдельных объектов и их взаимосвязей, которые были описаны либо упомянуты, либо подразумевались автором неявно.

Анализ диссертационных исследований

В настоящее время существует много работ, направленных как на фактографический поиск, так и на решение задач, связанных с автоматической обработкой текста. Рассмотрим работы, посвященные анализу информации, содержащейся в текстах авторефератов и диссертаций.

В работе [2] для выделения тематически близких к криминалистике дисциплин из ряда иных наук используются методы информационного анализа текста применительно к названиям авторефератов диссертаций (АРД). Так, установлена близость к криминалистике ряда

[1] ГОСТ Р 7.0.4–2006. Система стандартов по информации, библиотечному и издательскому делу. ИЗДАНИЯ. ВЫХОДНЫЕ СВЕДЕНИЯ. Общие требования и правила оформления.

[2] К сожалению, в последнее время далеко не все авторы следуют требованиям ГОСТ.

юридических наук «криминального блока» по некоторым аспектам рядов динамики количества тем АРД, дескрипторов, информационной плотности названий АРД.

Целью анализа являлось отграничение криминалистки от иных дисциплин методами информационно-количественного анализа текста, что достигается решением задач нахождения индикаторов такого отграничения.

Прежде всего, проводилось исследование соотношения семантической близости названий АРД по криминалистике с другими дисциплинами, юридические науки оказались заметно более близкими к криминалистике по изучаемому параметру. Было отмечено, что отсутствие при исследовании лемматизации несколько снижает показатели сходства дисциплин. Это объясняется специфическим употреблением морфологических форм слов в каждой науке. Далее изучались соотношения идентичности дескрипторов названий АРД по криминалистике с другими дисциплинами. Было выявлено, что тенденции при изучении соотношения идентичности названий АРД также сохранились. Авторы сделали вывод: анализ идентичности названий АРД и их дескрипторов по различным наукам может служить индикатором для выделения тематически близких дисциплин. При этом отмечается, что уверенное разделение с помощью данного индикатора близких наук остается проблематичным.

Далее проводился анализ количества тем АРД, а также дескрипторов (динамика по годам) (дисциплина, дисперсия выборки, среднеквадратичное отклонение, медиана, мода, среднее значение, эксцесс, асимметрия относительно среднего, корреляция исследуемых данных). Анализ данных позволил выделить из общего ряда юридические дисциплины на основании показателей эксцесса и корреляции динамики количества работ. Сделан вывод, что указанные индикаторы достаточно отчетливо выделяют юридические науки и могут служить соответствующими индикаторами. Отмечается, что отличие данных по юридическим дисциплинам выражено и в показателях дисперсии (что, видимо, указывает на относительно больший тематический разброс названий в юридических науках вообще), и также в асимметрии относительно среднего, в то же время разница внутри юридических дисциплин по данным параметрам несущественна.

Кроме того, изучалось количество дескрипторов на одну работу по каждому году (информационная плотность). Анализ этих данных показал близость криминалистики и уголовного процесса по параметрам «эксцесс» и «асимметрия относительно среднего» и позволил сделать вывод, что данные индикаторы могут служить для выделения весьма близких дисциплин (обе эти науки относятся к одной научной специальности 12.00.09). Иных отчетливо выраженных индикаторов не было выявлено.

Исследовалась также корреляции количества слов в названии АРД по криминалистике с другими науками. Анализ данных позволил сделать следующие выводы: а) различия в предметных областях наблюдаются во 2-м знаке после запятой; б) четкого выделения юридических наук по данному параметру не наблюдается ввиду близости криминалистики не только с ними, но и с техникой (видимо, здесь, в том числе, отражается близость криминалистики с техническими дисциплинами).

Таким образом, результаты данного исследования позволили прийти к следующим выводам.

- Выделение тематически близких наук (в данном случае дисциплин «криминального блока», близких к криминалистике) из ряда иных осуществимо методами информационного анализа названий авторефератов диссертаций.

При этом возможно применять следующие индикаторы:

– семантическая близость названий авторефератов диссертаций;

– эксцесс и корреляция динамики количества работ, а также дисперсия выборки и асимметрия относительно среднего – при анализе динамики количества тем авторефератов диссертаций и их дескрипторов.

- Индикаторы «эксцесс» и «асимметрия относительно среднего» могут служить для выделения весьма близких дисциплин (в нашем случае криминалистики и уголовного процесса) из ряда иных наук – при изучении динамики количества дескрипторов на одну работу (информационной плотности текста названий АРД).
- Индикатор корреляции количества слов в названии авторефератов диссертаций показывает близкие к криминалистике дисциплины с более широким их охватом (в частности не только юридические науки криминального блока, но и технические).

• Кроме того, проведенный информационный анализ текста названий авторефератов диссертаций подтверждает тезис о том, что криминалистика по исследуемым параметрам относится к юридическим дисциплинам, но имеет и некоторую техническую компоненту.

В работе сделано предположение о том, что только применение выявленных индикаторов в их совокупности позволит выделять различные науки из ряда иных дисциплин. Однако поиск индикаторов для отграничения криминалистики от существенно близких к ней наук (в рамках научной специальности 12.00.09) методами анализа текстов требует более глубоких исследований. При этом информационный анализ текстов позволяет оценивать положение дел в изучаемой дисциплине достаточно точными методами математической статистики.

В [3] выполнен анализ базы диссертаций Центральной библиотеки Пущинского научного центра (ПНЦ) РАН. В работе был проведен количественный и фактографический анализ диссертаций, защищенных в институтах ПНЦ РАН и в профильных организациях.

При проведении содержательного анализа были найдены организации, наиболее часто становившиеся ведущими в области физико-химической биологии, что позволяет судить о связях между учеными различных институтов, разрабатывающими аналогичные или близкие темы. Наиболее часто ведущими организациями становились МГУ им. М. В. Ломоносова, Институт биоорганической химии, Институт молекулярной биологии им. В. А. Энгельгардта, Институт биофизики клетки, Институт биохимии им. А. Н. Баха, Институт физико-химической биологии им. Белозерского (МГУ), Институт теоретической и экспериментальной биофизики, Институт общей генетики, Институт биологии развития им. Н. К. Кольцова АН.

В [4] дан анализ 888 диссертационных исследований российских (советских) социологов на соискание ученой степени доктора социологических наук с момента открытия научной специальности с 1990 по июнь 2010 г. Анализ тем диссертаций докторов социологических наук показал, что авторами диссертаций использовано 1 327 слов (понятий) или их производных (прилагательное, глагол и т. д.). После исключения предлогов и союзов осталось 1 271 слово (понятие). Выделены наиболее часто употребляемые слова (понятия), которые использованы в массиве докторских диссертаций.

Самое распространенное слово (понятие), которым оперируют социологи высшей квалификации, – *социальное* (без учета его производных, к примеру *социально-экономическое* и т. п.), оно использовано 325 диссертантами; *Россия* (*российское*) – 274 раза; *анализ* – 164; *социологическое* – 162; *общество* (*общественное*) – 147; *современность* (*современное*) – 140; *управление* (*управленческое*, *управлять*) – 125; *развитие* – 102; *система* (*системное*) – 94; *теория* (*теоретическое*) – 93; *процесс* (процессуальное) – 81; *условия* (*условное*) – 80; *проблема* (*проблемное*) – 74; *исследование* (*исследовательское*) – 71; *образование* (*обучение*) – 64 раза. В данном перечне представлены первые пятнадцать наиболее часто используемых слов (понятий).

В работе [5] был осуществлен проблемно-тематический анализ диссертаций по социальной педагогике и определен проблемно-понятийный комплекс, отраженный в совокупности понятийных рядов, составленных с учетом терминов и категорий, которые использовались при формулировании тематики диссертационных работ. Понятийные ряды социальной педагогики являются многоуровневыми и в своей совокупности составляют проблемно-понятийный комплекс социальной педагогики, отражающий терминологическую систему социальной педагогики.

В результате группирования тем диссертаций на основе выделения включенных в них терминов и категорий выделены понятийные ряды социально-педагогического знания, которые в той или иной мере отражают структуру социальной педагогики как научной дисциплины: история социальной педагогики; общие вопросы социальной педагогики; теория социального развития человека; теория социально-педагогического сопровождения; теория социально-педагогической инфраструктуры. Сформированная терминологическая система социальной педагогики как совокупность соподчиненных понятийных рядов позволяет осуществить переход к выделению и анализу ее понятийно-проблемных комплексов, сформированных как в рамках социальной педагогики в целом, так и в рамках ее внутренних разделов (теорий).

В [6] проведен анализ списков цитирования 49 диссертаций Индийского института менеджмента г. Ахмедабаб за период 2004–2009 гг. Из 4 319 ссылок цитирования были извлечены названия журналов и произведено их ранжирование. Цели исследования:

1) определить типы информационных ресурсов, наиболее часто используемых аспирантами института;

2) определить журналы, наиболее часто цитируемые аспирантами института;

3) определить группу основных журналов, используемых аспирантами института.

В [7] аналогично из ссылок цитирования диссертаций за 1999–2003 гг., полученных из ProQuest's Dissertations and Theses databases [8] по специальностям в сфере финансов и бухгалтерского учета, извлечены названия журналов для оценки научных интересов новых ученых в сфере бухгалтерского учета и литературных источников, на которые ссылаются. Журналы были классифицированы по специальностям и методам диссертационного исследования. Было показано, что рейтинг журналов варьируется от специальности и методов исследования.

Результаты литературного обзора показали, что проблемы выявления связей между фактами (или объектами) недостаточно разработаны и слабо освещаются. Рассмотренные работы главным образом направлены на извлечение информации о сущностях без установления связей между ними. Кроме того, в литературе не было найдено примеров использования методов содержательного анализа диссертаций в приложении к техническим наукам. Большинство работ посвящены статистическому анализу диссертаций.

Информационная модель фактов

Согласно «Логико-философскому трактату» Л. Витгенштейна [9], мир состоит не из предметов (вещей), а из фактов. Факт выступает как нечто отличное от вещи, как некоторое отношение, как взаимодействие двух предметов. Мир рассматривается как нечто, определяемое связями (взаимодействиями). Любой факт при этом – фиксация некоего отношения. Все факты фиксируются фразами, например «молоток забивает гвоздь». Любое предложение структурировано вполне конкретным образом: оно может быть представлено как 2 (или 3, 4…) объекта, которые как-то связаны между собой. Элементарное предложение связывает 2 объекта, а вещь – нечто общее совокупности фактов. Таким образом, отношения и факты объявляются первичными, а вещи представляют собой пересечение, совокупность возможных отношений. Иначе говоря, с вещью можно соотнести общую область «пересечения» множества фактов. Атомарный факт есть соединение (двух) объектов. Анализ фактов дает объекты или предметы. При этом по мере накопления фактов представление о вещи может меняться. Благодаря такой трактовке мира вещь выступает не как нечто данное, застывшее, вполне определенное, а как некоторая сущность с размытыми границами, и эти границы уточняются по мере выявления класса возможных для данной сущности отношений (фактов). Чтобы определить вещь, надо зафиксировать все факты (положительные – где может встречаться эта вещь, и отрицательные – где не может).

Таким образом, мир подразделяется на факты. Факт – существование событий. Событие – связь объектов (предметов, вещей).

Факты в тексте можно представить в виде языковой модели, способной содержать, хранить и передавать информацию. Языковые модели, содержащие целенаправленно отобранную информацию, принято называть информационными моделями.

Нетрудно заметить, что процитированные положения «Логико-философского трактата» (прежде всего, ключевые определения: «Атомарный факт есть соединение объектов (вещей, предметов)... Структура факта состоит из структур атомарных фактов») практически полностью воспроизводятся в модели данных «сущность – связь» [10], являющейся основой для унификации различных представлений данных.

Для единообразия определения понятия «факт» удобно использовать модификацию модели данных «сущность – связь», называемую моделью множества сущностей. Ее отличительные особенности заключаются в том, что, во-первых, в ней все трактуется как объекты (в том числе, например, цвет, в то время как в модели «сущность – связь» цвет обычно трактуется как «значение», а согласно «Логико-философскому трактату» «пространство, время и цвет

(цветность) есть формы объектов»), а во-вторых, все связи в этой модели – бинарные. Связи между объектами в модели множества сущностей также рассматриваются как объекты, связанные, в свою очередь, с объектами – атрибутами связей.

Таким образом, разумно следующее определение факта [11]: содержащаяся в тексте и метаданных документа совокупность связей между сущностями, описываемыми в онтологии информационной системы.

Модель документа

Информационная система представляет собой множество связанных различными отношениями документов, описывающих некие сущности (объекты, факты или понятия) [12]. Информация о той или иной сущности содержится в системе либо непосредственно в виде документа, который ее представляет, описывает или моделирует, либо в виде упоминаний об этой сущности, которые имеются в других документах, т. е. содержат опосредованную информацию об этой сущности.

Согласно стандартам построения открытых систем (OSI) структура и содержание документа должны описываться в соответствии с международными схемами данных [3]. Для описания соответствующих схем данных используются метаданные, которые определяют структуру и смысловое содержание документа. В нашей системе документом называется информационный ресурс, имеющий уникальный идентификатор и снабженный метаописанием (метаданными) в соответствии с рекомендациями OSI.

Дадим два определения.

Документом d_i называется пара $d_i = <S_i, V_i>$, где S_i – структура документа в соответствии с выбранной схемой данных; V_i – содержание документа (информационное наполнение).

Коллекция – множество документов с выделенной фиксированной структурой, содержание которых имеет одинаковую тематическую направленность.

С точки зрения унификации работы с документами будем представлять информационную систему в виде набора коллекций. Метаданные, описывающие структуру и содержание документов в коллекциях, подразделяются на описательные и структурные.

Структурные метаданные определяют структуру и свойства документов, в соответствии с которыми осуществляется их обработка (типы, связи, форматы представления, ограничения на управление доступом и т. п.). Описательные метаданные описывают смысловое содержание документа (его название, краткое содержание и т. п.). Отметим, что описательные метаданные, характеризующие документ, могут являться частью документа и в то же время могут содержать в соответствии с выбранной схемой данных сведения о документе (основные и дополнительные, такие как, например, авторы, название, дата создания и т. д.).

Элемент схемы данных этой коллекции будем называть структурным элементом.

Структурный элемент (далее просто элемент) имеет идентификатор и обладает некоторыми свойствами. Таким образом, элемент E – это совокупность $<ID, P>$, где ID – идентификатор элемента; P – свойства элемента.

Экземпляр элемента имеет значение (или содержание). Свойства элемента определяют характер работы с элементом. Элемент обладает типом, выбираемым из словаря. Тип определяет правила работы с элементом и, следовательно, является свойством элемента.

Примеры элементов: заголовок документа, аннотация документа, фамилия в визитной карточке, авторы документа. Значение элемента – его конкретная содержательная часть, а свойства элемента описывают его структуру. Для элемента визитной карточки «Фамилия» значение – Матвеев, идентификатор – 1, свойства – тип «word».

Структура документа – это набор структурных элементов. Содержание документа – объединение значений экземпляров элементов, составляющих документ.

---

[3] Р 50.1.041-2002. Руководство по проектированию профилей среды открытой системы (СОС) организации-пользователя / Госстандарт России. М., 2004.

ISO/IEC 7498-1:1994. Information technology – Open Systems Interconnection (OSI) – Basic Reference Model: The Basic Model.

ISO/IEC 14252-1996 (ANSI/IEEE Std1003.0-1995). Information technology Guide to the POSIX Open Systems Environment (OSE).

Для информационного анализа диссертаций была создана Информационная система (ИС), содержащая следующие коллекции.

1. Персоны и организации, диссертационные советы.

2. Авторефераты и диссертации. Диссертация обладает документной и лингвистической информативностью. Документная информативность связана с реализацией сигнальной функции, которая дает информацию организационного характера, т. е. извещает о том, что диссертация подготовлена и поступила в библиотеку организации по месту работы диссертационного совета, о месте и времени защиты, об ученых, являющихся оппонентами по диссертации. Она реализуется в таких атрибутах описания, как «соискатель», «тема», «специальность», «дата защиты», «организация, в которой выполнена работа», «шифр совета», «научный руководитель» (ФИО, ученая степень, звание), «оппоненты», «ведущая организация», «название организации, где можно ознакомиться с диссертацией», «дата рассылки автореферата», «ученый секретарь», «УДК». Лингвистическая информативность реализуется в автореферате или диссертации в атрибуте «Текст».

3. Термины. Особым видом объектов ИС является термин. Термин – слово или словосочетание, название определенного понятия какой-нибудь специальной области науки, техники, искусства, общественной жизни и т. п. Термин называет специальное понятие и в совокупности с другими терминами данной системы является компонентом научной теории определенной области знания. Примером терминов являются ключевые слова, описывающие содержание диссертации.

Модель отношений между документами в системе

Для решения сформулированных выше задач мы должны определить связи (отношения) между документами – объектами, составляющими содержание перечисленных выше коллекций.

В основу нашей модели отношений между документами [13] в информационной системе легла модель RDF. В нашей системе связи между документами устанавливаются путем задания на множестве документов бинарных отношений, которые в соответствии с правилами RDF могут быть записаны в виде $A(R,V)$: объект R имеет атрибут A со значением V. Например, тот факт, что В. Б. Барахнин занимает некоторую должность (post) в ИВТ СО РАН, записывается как Post(`ИВТ СО РАН', `Барахнин В. Б.'), где Post – то или иное значение термина из списка (тезауруса) должностей.

Связь – это направленное или ассоциативное отношение между объектами системы, например А. А. Петров преподает в НГУ. Факт – событие (как правило, зафиксированное и произошедшее), которое может сопровождаться временной и географической метками и др., например, П. П. Иванов защитил кандидатскую диссертацию в 1994 г. в Новосибирске. События представляют действия, происходящие в реальном мире, и определяются указанием типа действия и ролей, которые играют сущности в этом действии. Факт может быть извлечен из текста документов автоматически либо определен экспертом.

Как говорилось ранее, если событие – связь объектов, то факт может определить как отношение между объектами, которое может иметь временные и географические атрибуты, например, год – 1994, географическую привязку – Новосибирск.

Можно выделить следующие виды связей.

Прямые. В этом случае есть факт о связи двух понятий, например, отношение соискатель-оппонент.

Нечеткие (не представленные фактом):

- по общему месту и времени у пары различных фактов различных объектов, например, дата и место защиты диссертации, позволяет установить соискателей, защитивших диссертацию в один день в одном совете;
- косвенные (транзитивные) – через общий третий объект – отношение у пары фактов различных объектов, например, связь диссертация – ключевые слова. Установление связи подобных диссертаций выполняется через ключевые слова.

Факты можно выразить посредством высказываний с использованием предикатов. Методы математической логики позволяют формализовать эти утверждения и представить их в виде, пригодном для анализа.

Рассмотрим высказывание «Преподаватель Иванов А. А, родился в 1962 году». Оно выражает следующие свойства сущности «Иванов А. А.»:

- в явном виде – год рождения;
- в неявном виде – принадлежность к преподавателям.

Первое свойство устанавливает связь между парами сущностей «Иванов А. А.» и «год рождения», а второе свойство устанавливает связь между парами сущностей «Иванов А. А.» и «множество преподавателей». Формализация этого высказывания представляется как результат присваивания значений переменных, входящих в следующие предикаты:

РОДИЛСЯ (Иванов А. А., 1962);

ЯВЛЯЕТСЯ ПРЕПОДАВАТЕЛЕМ (Иванов А. А.).

Пример информационной модели описания диссертаций представлен на рис. 1. Существенными характеристиками диссертации являются соискатель, тема, специальность, ученая степень, год, организация, в которой выполнена работа, организация, в которой защищалась диссертация, шифр совета, научный руководитель, оппоненты, ведущая организация, УДК.

Связи между документом и его элементами представлены на рисунке, который дает схемное описание рассматриваемой модели. В этом описании используются следующие элементы: соискатель, оппонент 1, оппонент 2, оппонент 3, научный руководитель, организация выполнения работы и организация защиты диссертации, ведущая организация – объекты, тема, специальность, ученая степень, шифр совета, УДК – текстовые значения, год – числовое.

Формализованное описание данной модели является предикатом с именем диссертация: диссертация (соискатель, тема, год, специальность, ученая степень, организация выполнения работы, организация защиты диссертации, ведущая организация, шифр совета, научный руководитель, оппонент 1, оппонент 2, оппонент 3, УДК).

Для конкретных значений аргументов этот предикат превращается в факт. Например, если Барахнин В. Б. защитил диссертацию «Программные системы информационного обеспечения научной деятельности: модели, структуры и алгоритмы» в 201 г., то имеет место факт: Диссертация (Барахнин В. Б., Программные системы информационного обеспечения научной деятельности: модели, структуры и алгоритмы, 2011, 05.13.17, доктор технических наук, Институт вычислительных технологий СО РАН, Московский государственный университет печати, Институт математики СО РАН, Д 212. 147.03, Федотов А. М., Шайдуров В. В., Хорошевский В. Ф., Мальцева С. В., 004). С помощью таких фактов можно выделить различные характеристики диссертаций, например, найти соискателей, защитивших диссертацию по специальности 05.13.17 в 2011 г.

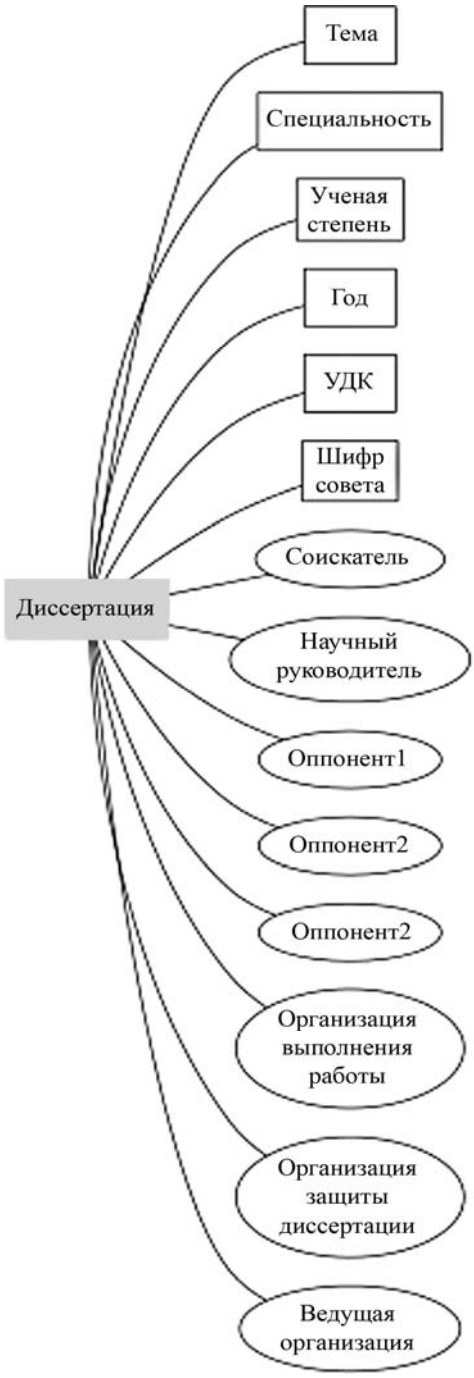

Рис. 1. Информационная модель описания диссертации

Статистическое исследование текста диссертации

При исследовании текста диссертаций используется метод контент-анализа – метод качественно-количественного анализа содержания документов с целью выявления или измерения различных фактов и тенденций, отраженных в этих документах. Сущность метода контент-анализа состоит в выделении в содержании научных документов некоторых ключевых признаков (содержательных единиц анализа, проблем, категорий), которые отражают существенные (фактические и смысловые) стороны содержания с последующим подсчетом частоты употребления этих единиц.

В данной работе используется тезаурусный метод, являющийся разновидностью контент-анализа, суть которого состоит в сведении рассматриваемого текста к ограниченному набору элементов и терминов, которые затем подвергаются анализу.

Не все документы могут выступить объектом контент-анализа. Необходимо, чтобы исследуемое содержание позволило задать однозначное правило для надежного фиксирования нужных характеристик (принцип формализации), а также чтобы интересующие исследователя элементы содержания встречались с достаточной частотой (принцип статистической значимости). Можно выделить следующие направления применения контент-анализа [14; 15]:

а) выявление того, что существовало до текста и что тем или иным образом получило в нем отражение (текст как индикатор определенных сторон изучаемого объекта – окружающей действительности, автора или адресата);

б) определение того, что существует только в тексте как таковом (различные характеристики формы – язык, структура и жанр сообщения, ритм и тон речи);

в) выявление того, что будет существовать после текста, т. е. после его восприятия адресатом (оценка различных эффектов воздействия).

Основой содержания диссертации является принципиально новый материал, включающий описание новых фактов, явлений и закономерностей, или рассмотрение имеющегося материала в совершенно ином аспекте. Таким образом, автор диссертации сосредоточен на описании новых фактов, их точном представлении научной общественности, и их контент-анализ предполагает выявление фактов, существовавших до написания текста диссертации.

В данном исследовании категорией анализа содержания диссертации является соответствие ее темы специальности ВАК.

После того, как категории сформулированы, необходимо выбрать соответствующую единицу анализа – лингвистическую единицу речи или элемент содержания, служащие в тексте индикатором интересующих исследователя явлений.

Смысловыми единицами контент-анализа в нашем случае являются:

а) понятия, выраженные в отдельных терминах;

б) темы, выраженные в целых смысловых абзацах, частях текстов, статьях;

в) имена, фамилии людей, названия организаций;

г) события, факты и т. п.

Наконец, необходимо установить единицу счета – количественную меру взаимосвязи текстовых и внетекстовых явлений, выделить единицы счета, которые могут совпадать либо не совпадать с единицами анализа. В нашем случае процедура сводится к подсчету частоты упоминания выделенной смысловой единицы (интенсивность).

Научные связи

Определим научное пространство ученого N как совокупность ученых $\{S\}$, связанных с N различными научными отношениями, как, например, связи типа соискатель – научный руководитель, соискатель – оппонент, автор книги – редактор, автор книги – рецензент (не анонимный) и т. д.

В рамках математической теории организации рассматриваются свойства научных коллективов и групп на основе теории графов. Коллектив представляет собой комплексную структуру, систему, множество различных отношений, зависимостей [16].

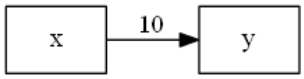

Рис. 2. Элемент графа

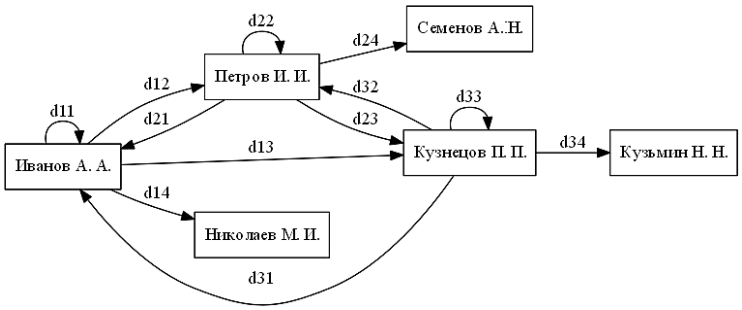

Рис. 3. Фрагмент графа

Средством представления незримых коллективов является сеть (сеть идейного, творческого и прочего влияния) (рис. 2, 3). Звено сети (см. рис. 2) характеризует степень влияния *x* на *y* и может означать, например, что «*y* цитирует *x*» 10 раз. Иначе говоря, *y* использовал концепции, идеи, факты *x*, развивал их и т. д. Тем самым между *x* и *y* имеется устойчивая информационная связь, причем число 10 – характеристика интенсивности этой связи [15; 16].

Если построить сеть взаимных ссылок, то можно выделить подграфы, элементы которых интенсивно связаны друг с другом. Такие подграфы образуют незримые коллективы (на рис. 3 подграф Иванов А. А., Петров И. И., Кузнецов П. П. – научный неформальный коллектив).

Например, неформальный коллектив из N элементов ($N = 3$) может быть представлен следующей матрицей $N \times N$ [4] [17]:

$$\mathrm{D} = \begin{array}{c} \\ a \\ b \\ c \end{array} \begin{array}{c} \begin{array}{ccc} a & b & c \end{array} \\ \begin{pmatrix} d_{11} & d_{12} & d_{13} \\ d_{21} & d_{22} & d_{23} \\ d_{31} & d_{32} & d_{33} \end{pmatrix} \end{array}.$$

Здесь d_{ji} – количество ссылок j на i (иначе говоря, мера неформального воздействия i на j). Здесь d_{13} – количество ссылок c на a и наоборот, d_{31} – количество ссылок a на c. Можно также ввести меру $m(x)$ неформального (идейного, научного и проч.) статуса индивидуума x, например, следующего вида:

$$m(a) = \frac{d_{13}}{d_{31}} + \frac{d_{12}}{d_{21}} \text{ или } m(a) = \frac{d_{13}}{d_{31}} + \frac{d_{12}}{d_{12}}.$$

Эти меры используют различные выражения отношения «влияния *a* на остальных» к «влиянию остальных на *a*».

Лицо *x* с максимумом $m(x)$ может быть названо лидером неформального коллектива. Между формальными и неформальными отношениями существуют определенные причинно-следственные связи. Например, может наблюдаться следующая последовательность их развития:

- *a* и *b* образуют неформальный коллектив (взаимные ссылки);
- *a* и *b* печатаются в соавторстве;
- *a* и *b* начинают работать вместе.

Выявление неформальных лидеров и коллективов способствует лучшей организации выполнения проектов путем привлечения в формальный коллектив единомышленников.

Описанный выше подход является статическим. Можно рассматривать развитие коллектива в динамике, когда с течением времени к графу добавляются новые вершины и ребра и одновременно часть прежних элементов удаляется. Такие графы достаточно наглядно отображают перемены в коллективе, связанные, например, с уходом прежнего формального лидера.

[4] Elements of the mathematical theory of organization // Portal Cadmium. URL: http://cadmium.ru/content/view/832/45/

Другим видом научных коллективов являются научные школы, информацию о которых можно получить на основе анализа таких реквизитов диссертации, как учебное заведение, в котором выполнена работа, научный руководитель, ведущая организация, дата и время защиты, шифр совета и т. д. Понятие научной школы употребляют «применительно к относительно небольшому научному коллективу, объединенному не столько организационными рамками, не только конкретной тематикой, но и общей системой взглядов, идей, интересов, традиций – сохраняющейся, передающейся и развивающейся при смене научных поколений».

Рассмотрим структуру графа диссертаций [17]. Вершины ориентированного графа диссертаций соответствуют диссертантам, руководителям и оппонентам диссертантов. Бинарное отношение на парах вершин задается естественным образом: дуги выходят из вершин-руководителей и вершин-оппонентов и входят в соответствующую вершину-диссертант. Сохраняется информация о годе защиты, совете защиты, ведущей организации и т. п. Типичный фрагмент графа должен содержать 4 или более вершин (рис. 4).

1. Вершины ориентированного графа диссертаций соответствуют диссертантам, руководителям и оппонентам диссертантов. Бинарное отношение на парах вершин задается естественным образом: дуги выходят из вершин-руководителей и вершин-оппонентов и входят в соответствующую вершину-диссертант. Сохраняется информация о годе защиты, совете защиты, ведущей организации и т. п.

2. Число входящих дуг в вершину-диссертант лежит в границах от 3 до 8. Максимальная входящая степень будет у вершин-диссертантов, которые защитили кандидатскую и докторскую степени, имеют несколько руководителей и консультантов. Степени вершин-руководителей и вершин-оппонентов могут быть очень большими.

3. Из вершины-диссертанта дуга будет выходить, если он в дальнейшем стал руководителем или оппонентом какой-либо диссертации.

4. Большие степени в графе выявляют персон, оказавших большое влияние на формирование коллектива специалистов в данной области. Длинная цепь в графе показывает протяженный во времени процесс защит диссертаций, где в качестве руководителя выступает бывший диссертант и т. д. Таким образом, наличие больших степеней и длинных цепей позволяет предполагать существование школы по рассматриваемому направлению.

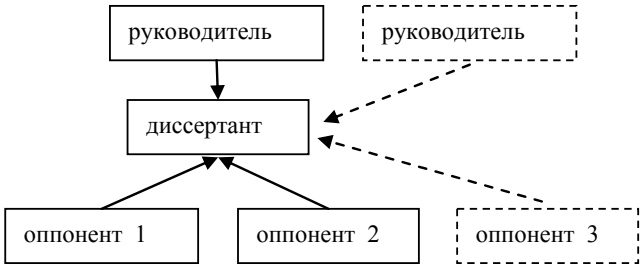

Рис. 4. Фрагмент графа диссертаций

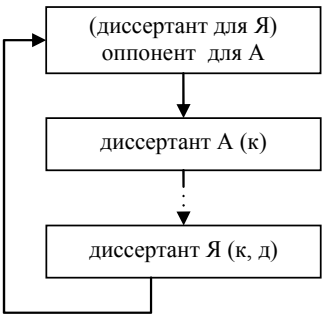

Рис. 5. Контур графа диссертаций

Граф может иметь контуры. На рис. 5 показан пример образования контура: диссертант А защитил кандидатскую (к) диссертацию, далее стал руководителем другого диссертанта и т. д. После последовательности защит диссертант Я защитил кандидатскую и докторскую (д) диссертации и затем стал оппонентом докторской диссертации для кандидата наук, бывшего оппонентом диссертанта А.

Лингвистическая модель обработки фактов

Средством классификации понятий в определенной области является тезаурус, который содержит:

1) структурированную систему терминов (понятий) с определением иерархических и ассоциативных отношений между ними;

2) список терминов, определяющих каждое из понятий; все термины, определяющие одно и то же понятие, называются синонимами. Сведение синонимов вместе реализует термино-

логический контроль, обеспечивающий возможность выражения одного и того же понятия разными способами.

Тезаурус содержит набор дескрипторов для индексирования и поиска документов. Современный этап развития систем информационного поиска ведет начало от работ Кельвина Муэрса (Calvin Mooers), предложившего описывать содержание документов простым перечислением дескрипторов – терминов, употребляющихся в самом документе и тем самым определяющих его содержание в пределах терминологии данной предметной области. Предполагалось, что перечень дескрипторов, существенных для описания документов определенной предметной области, не слишком велик, что его можно свести в словарь, в котором будут заданы отношения между дескрипторами наподобие соотношения иерархии классов понятий, и использовать этот ограниченный словарь, называемый информационно-поисковым тезаурусом, как систему фасетного индексирования документов. Специфика некоторых тематических областей, например, сферы науки и образования, заключается, в частности, в том, что ее терминология не имеет узкоотраслевого характера, а включает, наряду со специальной научной и педагогической лексикой, также все термины тех отраслей знания, по которым ведется исследование и обучение. Поэтому с узкоотраслевыми тезаурусами должны использоваться универсальные тезаурусы научной и технической лексики. Очевидно, что применять неопределенный набор тезаурусов в качестве единой классификационной системы невозможно.

Современный подход использования тезаурусов состоит не в индексировании документов, а в определении релевантности документа поисковому запросу. При этом классификационными признаками документов служат сами слова, которые употреблены в документе (ключевые слова), а критерий соответствия документа запросу определяется на основе семантической информации по возможности обо всех ключевых словах данной специальности. Каждый пользователь, заинтересованный в работе по своему направлению (специальности), должен дополнять встроенный в систему общетехнический словарь только своим специфическим словарем.

Рассмотрим задачу классификации диссертаций и авторефератов по специальностям ВАК с использованием тезаурусного метода анализа, суть которого состоит в сведении рассматриваемого текста диссертаций и авторефератов к ограниченному набору элементов и терминов, которые затем подвергаются анализу.

Поскольку в ВАК нет формальных критериев, по которым можно определить соответствие диссертации определенной специальности, то построить тезаурус ключевых слов для классификации диссертаций по всем специальностям затруднительно. В ВАК существуют только лексические критерии соответствия формулировок документов и результатов исследования. При этом формулировки одних и тех же результатов с небольшими отличиями могут соответствовать разным специальностям. Кроме того, в разных ученых советах предъявляются разные требования к диссертациям по одной и той же специальности. Одним из способов классификации диссертаций может быть классификация фактов, содержащихся в диссертациях, в соответствии со специальностями ВАК с использованием методов обучения.

Методы извлечения понятий из текста диссертации

Рассмотрим подробнее методику извлечения фактов из текста диссертации. Извлечение понятий из текста представляет собой технологию, обеспечивающую получение информации в структурированном виде. В качестве структур могут запрашиваться как относительно простые понятия (ключевые слова, персоны, организации, географические названия), так и более сложные, например, имя персоны, ее должность в конкретной организации и т. п.

Данная технология включает три основных метода:

а) извлечение слов или словосочетаний, важных для описания содержания текста. Это могут быть списки терминов предметной области, персон, организаций, географических названий и др.;

б) прослеживание связей между извлеченными понятиями;

в) извлечение сущностей, распознавание фактов и событий.

Подходы к извлечению различных типов понятий из текстов существенно различаются. Например, для выявления принадлежности документа к тематической рубрике могут использоваться методы классификации. Для выявления названий организаций и персон применяются как система шаблонов, так и результаты структурного исследования текста, например, используется таблица префиксов названий организаций. Выявление географических названий предполагает использование таблиц, в которых кроме шаблонов написания этих названий применяются коды и названия стран, регионов и отдельных населенных пунктов. Таким образом, методы извлечения из текста сущностей и терминов имеют свою специфику для каждого типа.

Методы автоматического извлечения понятий можно разделить на 2 типа.

- Методы машинного обучения. Основываются на статистических (вероятностных) методах извлечения знаний. Для обучения системы необходим размеченный корпус текстов.
- Методы, основанные на знаниях. Основываются на языках описания правил-шаблонов, которые составляются экспертами. Основой недостаток метода – написание правил может занимать много времени.

Методы, основанные на знаниях, используются при необходимости обеспечить максимально возможное качество извлечения, однако для их работы необходимо иметь словари, списки слов и экспертов – инженеров по знаниям, при этом отсутствует необходимость иметь много размеченных данных.

Методы машинного обучения используются при необходимости обеспечить хорошее качество извлечения, при этом отпадает необходимость в экспертах и словарях, необходимо иметь большой объем размеченных данных.

Наиболее эффективными являются комбинированные методы.

Извлечение именованных сущностей. Выделение сущностей является ключевым этапом предобработки текста для решения более сложных задач извлечения информации.

Под термином *именованная сущность* будем понимать объект определенного типа, имеющий имя, название или идентификатор.

Особенностями этого вида объектов являются:

- большое множество разных сущностей;
- отсутствие строгих правил именования сущностей;
- постоянное появление новых сущностей.

Какие типы выделяет система, определяется в рамках конкретной задачи. Для диссертаций и авторефератов это *люди* (PER), *места* (LOC), *организации* (ORG), *время* (TIME). В общем случае системе на вход поступает текст, на выходе система сообщает информацию о положении имен в тексте и информацию о классах, которые им соответствуют. Набор классов фиксируется заранее. Приведем пример размеченного текста:

[PER Барахнин Владимир Борисович]. Программные системы информационного обеспечения научной деятельности: модели, структуры и алгоритмы: диссертация доктора технических наук: 05.13.17 / Место защиты: [ORG Моск. гос. ун-т печати].- [LOC Новосибирск], [TIME 2010].- 315 с.

Для извлечения именованных сущностей применяются несколько типов признаков [18]:

1) признаки уровня слов (*N*-граммы, суффиксы, префиксы, части речи и т. д.);

2) признаки уровня документа (наличие акронимов в корпусе, позиция термина в предложении, наличие термина в заголовке или тексте и т. д.);

3) дополнительная информация (слова-указатели, например, Inc., Corp., списки стоп-слов, слов с капитализацией, которые не являются именованными сущностями, и т. д.).

В пределах одного документа может быть несколько вхождений одного и того же имени, которое может относиться к одной сущности или же к различным объектам. В простейшем случае обычно исходят из предположения, что в одном документе одно и то же имя относится к одной и той же сущности. Базовый набор признаков составлен из признаков первой группы для слов, находящихся в скользящем по тексту окне размера до 5 токенов. Под токеном подразумеваются не только слова, но и символы пунктуации.

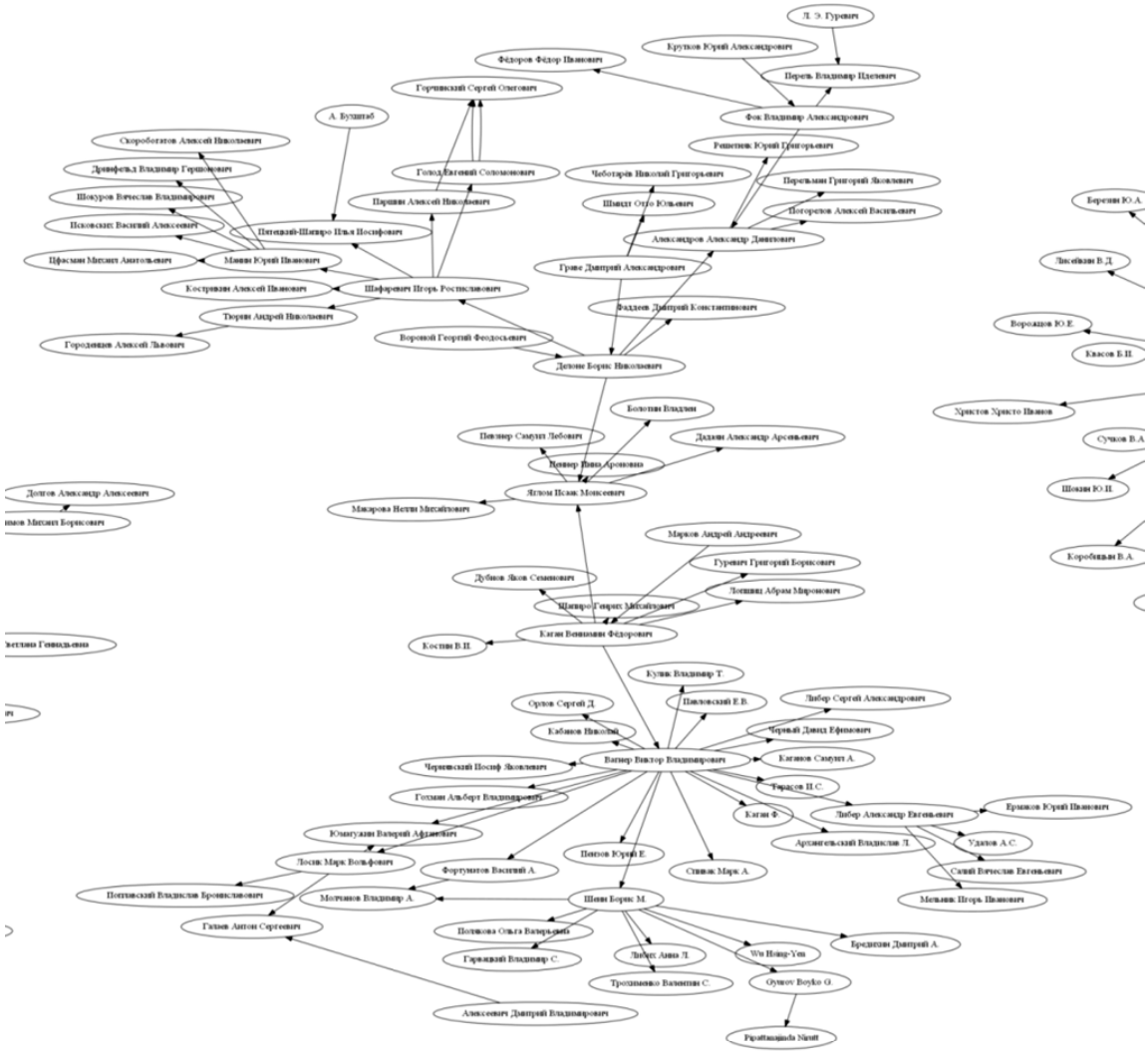

Рис. 6. Фрагмент графа диссертаций

Были проанализированы 4 587 диссертаций и авторефератов и получен граф связей между персонами в диссертации на основании вышеприведенной модели (см. рис. 4). Граф распадается на множество несвязанных компонент, в которых можно отыскать подграфы (рис. 6) с длинными цепями с длиной 2, что позволяет говорить о наличии научной школы.

Извлечение ключевых терминов из текста. Ключевыми терминами (ключевыми словами или ключевыми фразами) являются важные термины в документе, которые могут дать высокоуровневое описание содержания документа для читателя. Извлечение ключевых терминов является базисным этапом для многих задач обработки естественного языка, таких как классификация документов, кластеризация документов, суммаризация текста и вывод общей темы документа [19; 20].

В данной работе используется метод выделения терминов на основе морфологических шаблонов, а ключевые термины выражаются именными словосочетаниями. В именных словосочетаниях главным словом (основным носителем смысла) является, как правило, первое слева существительное, а остальные слова служат для уточнения значения главного слова.

Для выделения ключевых терминов используются следующие виды шаблонов:

- П + С – согласованные прилагательное + существительное;
- С + Срод.п. – существительное + существительное в родительном падеже;
- С + Ств.п. – существительное + существительное в творительном падеже;
- П + П + С – согласованные прилагательное + прилагательное + существительное;
- С + П + Срод.п. – существительное + согласованное прилагательное + существительное в родительном падеже;

- С + П + Ств.п. – существительное + согласованное прилагательное + существительное в творительном падеже.

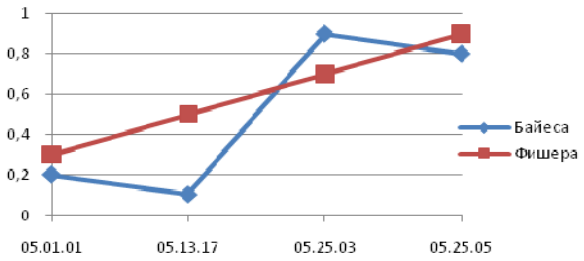

Рис. 7. Точность. Зависимость от категории

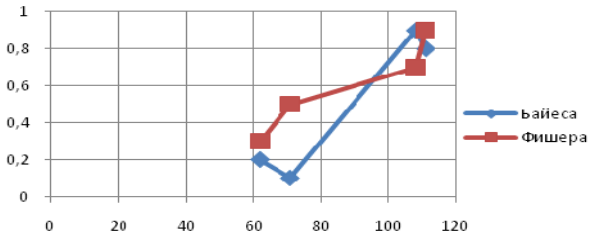

Рис. 8. Точность. Зависимость от количества документов в рубрике

После выделения терминов определяется их тематика с помощью метода классификации – отнесение документа к одной из нескольких категорий на основании семантического содержания документа. Для классификации применяются методы обучения с учителем, которые позволяют провести классификацию или спрогнозировать значение исходя из ранее предъявленных примеров. Из множества существующих методов были выбраны метод наивной классификации Байеса и метод Фишера. Существенное преимущество наивных байесовских классификаторов по сравнению с другими методами заключается в том, что их можно обучать и затем опрашивать на больших наборах данных [21]. Даже если обучающий набор очень велик, обычно для каждого образца есть лишь небольшое количество признаков, а обучение и классификация сводятся к простым математическим операциям над вероятностями признаков. Это особенно важно, когда обучение проводится инкрементно, – каждый новый предъявленный образец можно использовать для обновления вероятностей без использования старых обучающих данных. (Отметим, что код для обучения байесовского классификатора запрашивает по одному образцу за раз, тогда как для других методов, скажем деревьев решений или машин опорных векторов, необходимо предъявлять сразу весь набор.) Поддержка инкрементного обучения очень важна в случаях расширении набора категорий в классификаторе, который постоянно обучается на вновь поступающих документах, должен обновляться быстро и, возможно, даже не имеет доступа к старым документам. Еще одно достоинство наивных байесовских классификаторов – относительная простота интерпретации того, чему классификатор обучился. Метод Фишера – альтернативный метод классификации, обеспечивает большую гибкость при настройке параметров классификации.

Результаты тестирования точности алгоритмов классификации терминов приведены на рис. 7 и 8, что позволяет сделать выводы о точности алгоритмов классификации около 90 % при количестве документов в рубрике более ста.

Список литературы

1. *Шокин Ю. И., Федотов А. М., Барахнин В. Б.* Проблемы поиска информации. Новосибирск: Наука, 2010. 198 с.

2. *Бугаев К. В.* Отграничение криминалистики от иных наук методами информационного анализа текста // Юридический мир. 2011. № 8. С. 40–43.

3. *Бескаравайная Е. В., Митрошин И. А.* Анализ базы данных диссертаций ПНЦ РАН // Информационное обеспечение науки: новые технологии. М.: Научный Мир, 2011. С. 124–133.

4. *Прошанов С. Л.* Докторские диссертации по социологии (1990–2010 гг.) // Социологические исследования. 2011. № 1. С. 30–39.

5. *Липский С. И.* Проблемно-тематический анализ диссертационных исследований по социальной педагогике (1971–2008 гг.): Автореф. дис. … канд. пед. наук. Кострома, 2009.

6. *Anil Kumar H., Mallikarjun Dora.* Citation analysis of doctoral dissertations at IIMA: A review of the local use of journals // Library Collections, Acquisitions, and Technical Services. 2011. Vol. 35, Issue 1. P. 32–39.

7. *Kam C. Chan, Kam C. Chan, Gim S. Seow, Kinsun Tam* Ranking accounting journals using dissertation citation analysis: A research note // Accounting, Organizations and Society. 2009. Vol. 34, Issues 6–7. P. 875–885.

8. *Dilek Altun, Çağla Öneren Şendil, İkbal Tuba Şahin.* Investigating the National Dissertation and Thesis Database in the Field of Early Childhood Education in Turkey // Procedia – Social and Behavioral Sciences. 2011. Vol. 12. P. 1–654. International conference on education and educational psychology, 2–5 December 2010, Cyprus.

9. *Wittgenstein L.* Logisch-Philosophische Abhandlung // Annalen der Naturphilosophie. Leipzig: Verlag Unesma, 1921. Vol. 19. Parts 3/4. P. 185–262.

10. *Chen P. P.* The entity-relational model. Toward a unified view of data // ACM TODS. 1976. № 1. P. 9–36 / Рус. пер. Чен П. П.-Ш. Модель «сущность – связь» – шаг к единому представлению данных // СУБД. 1995. № 3. С.137–158.

11. *Барахнин В. Б., Федотов А. М.* Построение модели фактографического поиска // Вестн. Новосиб. гос. ун-та. Серия: Информационные технологии. 2013. Т. 11, вып. 4.

12. *Барахнин В. Б., Леонова Ю. В.* Информационная модель отношений между документами в информационной системе // Вычислительные технологии. 2005. Т. 10. Спец. выпуск. С. 129–137.

13. *Leonova Yu. V., Barakhnin V. B.* On the Problem of Modeling of the Relations between Documents // Computer Science and Information Technology. 2013. Vol. 1 (2). P. 138–144.

14. *Манаев О. Т.* Контент-анализ как метод исследования // «ПСИ-ФАКТОР». URL: http://psyfactor.org/lib/content-analysis3.htm

15. *Хайтун С. Д.* Наукометрия: Состояние и перспективы. М.: Наука, 1983.

16. *Бурков В. Н., Заложнев А. Ю., Новиков Д. А.* Теория графов в управлении организационными системами. М.: Синтег, 2001.

17. *Леонова Ю. В., Добрынин А. А., Веснин А. Ю.* Построение графа диссертаций // XIV Российская конференция с участием иностранных ученых «Распределенные информационные и вычислительные ресурсы» (DICR-2012): Программа конференции и тезисы докладов (Новосибирск, Россия, 26–30 ноября 2012). Новосибирск: ИВТ СО РАН. 2012. С. 17.

18. *Ермакова Л. М.* Методы извлечения информации из текста // Вестник Пермского университета. Сер.: Математика. Механика. Информатика. 2012. Вып. 1 (9). С. 77–84.

19. *Manning C. D., Schtze H.* Foundations of Statistical Natural Language Processing. The MIT Press, 1999.

20. *Гринева М., Гринев М., Лизоркин Д.* Анализ текстовых документов для извлечения тематически сгруппированных ключевых терминов // Тр. Ин-та системного программирования РАН. URL: http://citforum.ru/database/articles/kw_extraction/

21. *Сегаран Т.* Программируем коллективный разум: Пер. с англ. СПб.: Символ-Плюс, 2008.

Материал поступил в редколлегию 25.03.2014

Yu. V. Leonova, A. M. Fedotov

INVESTIGATON THE SCIENTIFIC RELATIONS BASED ON THE ANALYSIS OF THESES

In this paper describes the models and methods for the analysis \ of theses and abstracts in order to study the structure of scientific relations scientist (scientific environment scientist), the structure and dynamics of the research teams (research schools), the statistical study of the text theses. Such studies provide opportunities to study and evaluate trends in the development of different scientific fields to identify the person, research centers and organizations, academics, school, study the relationship between the individual communities.

Keywords: information extraction, entity relation, named entity, selection of facts, selection relations, information model.