Извлечение знаний и фактов из текстов диссертаций и авторефератов для изучения связей научных сообществ*

© Ю.В. Леонова © А.М. Федотов
Институт вычислительных технологий СО РАН,
Новосибирск
juli@ict.nsc.ru fedotov@sbras.ru

Аннотация

В данной работы выполнено исследование диссертаций и авторефератов с целью изучения структуры научных связей ученого (научное окружение ученого), структуры и динамики развития научных коллективов (научные школы), статистического исследование текста диссертаций. Такие исследования дают возможности изучения и оценивания тенденций развития различных научных направлений, идентифицировать персоны, научные центры и организации, научные школы, изучать взаимосвязи между отдельными сообществами.

1 Введение

Целью данной работы является изучение связей научных сообществ, в рамках которых осуществляется научная деятельность, основанное на анализе 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 государственные институты, общественные организации, неформальные группы и т.д. Реализация этой цели включает в себя решение 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 изучения и оценивания тенденций развития различных научных направлений, идентифицировать персоны, научные центры и организации, научные школы, изучать взаимосвязи между отдельными сообществами.

В настоящее время существует много работ [1-7], направленных на анализ диссертаций. Однако в литературе не было найдено примеров использования методов в приложении к техническим наукам. Большинство работ посвящены статистическому анализу диссертаций.

2 Информационная модель фактов

Согласно «Логико-философскому трактату» Л.Витгенштейна [8]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. То есть 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).

Труды 15-й Всероссийской научной конференции «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» — RCDL-2013, Ярославль, Россия, 14-17 октября 2013 г.

*Работа выполнена при частичной поддержке РФФИ, грант №13-07-00258

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.

3 Модель документа в системе

Информационная система представляет собой множество связанных различными отношениями документов, описывающих некие сущности (объекты, факты или понятия) [9]. 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) [10] структура и содержание документа должны описываться в соответствии с международными схемами данных. Для описания соответствующих схем данных используются метаданные, которые определяют структуру и смысловое содержание документа. В нашей системе документом называется информационный ресурс, 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 данной 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.

Информационная система содержит 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 определённого 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 [11]. Примером терминов являются ключевые слова, описывающие содержание диссертации.

4 Модель отношений между документами в системе

Для решения сформулированных выше задач мы должны определить связи (отношения) между документами.

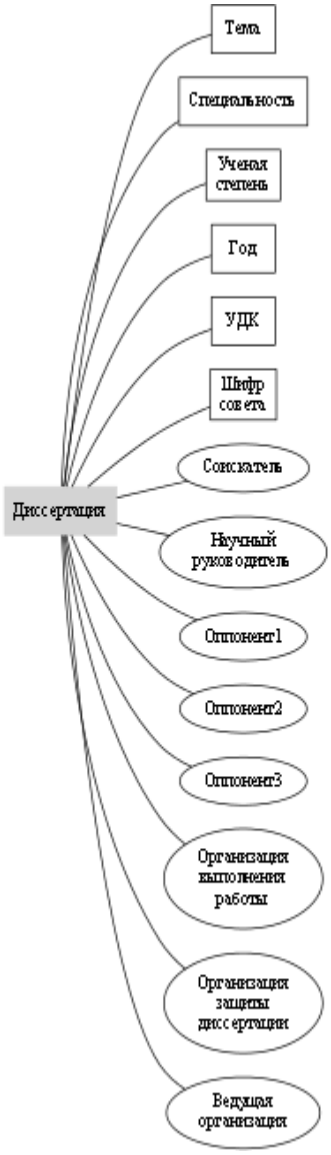

Рис. 1. Информационная модель описания диссертации

В основу нашей модели отношений [9] между документами 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 Барахнин В.Б. занимает некоторую должность (post) в ИВТ СО РАН, записывается как Post(`ИВТ СО РАН', `Барахнин В.Б.'), где Post - то или иное значение из списка (тезауруса) должностей.

Связь — это направленное или ассоциативное отношение между объектами системы, например Петров А.А. преподает в НГУ. Факт — событие (как правило, зафиксированное и произошедшее), которое может сопровождаться временной и географической метками и др., например, Иванов П.П. защитил кандидатскую диссертацию в 1994 году в г. Новосибирск. 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 либо определен экспертом.

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

Можно выделить следующие виды связей:

- Прямые. В этом случае есть факт о связи двух объектов, 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 (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, оппонент1, оппонент2, оппонент3, 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, оппонент1, оппонент2, оппонент3, 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 2011 году, 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, можно выделить соискателей, защитивших диссертацию по специальности 05.13.17 в 2011 году.

5 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 [12, 13].

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:

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 разработке и практическом применении контент-анализа выделяют несколько стадий. После того, как сформулированы тема, задачи и гипотезы исследования, определяются категории анализа, т.е. наиболее общие, ключевые понятия, соответствующие исследовательским задачам.

В данном исследовании категорией анализа содержания диссертации является ее тема, соответствующая 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.

Единицы анализа, взятые изолированно, могут быть не всегда правильно истолкованы, поэтому они рассматриваются на фоне более широких лингвистических или содержательных структур, указывающих на характер членения текста, в пределах которого идентифицируется присутствие или отсутствие единиц анализа — контекстуальных единиц. Например, простейшим элементом текста является слово, для единицы анализа «слово» контекстуальная единица – «предложение».

Смысловыми единица контент-анализа могут быть:

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. Выделение единиц счета, которые могут совпадать либо не совпадать с единицами анализа. В нашем случае процедура сводится к подсчету частоты упоминания выделенной смысловой единицы (интенсивность).

6 Научные связи

Научное пространство учёного N определим как совокупность учёных {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.[14].

7 Научные коллективы

Коллектив – устойчивая во времени организационная группа взаимодействующих людей со специфическими органами управления, объединенных целями совместной общественно-полезной деятельности и сложной динамикой

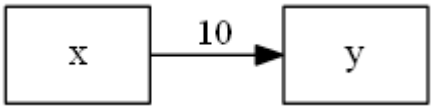

Рис. 2. Элемент графа

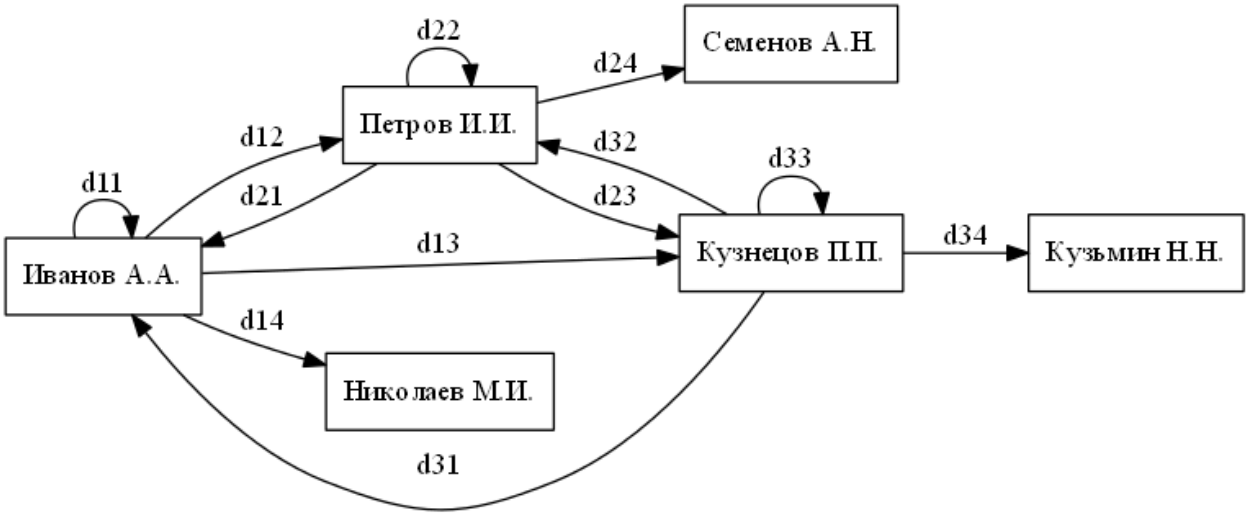

Рис.3. Фрагмент графа

формальных (деловых) и неформальных взаимоотношений между членами группы. Т.о. коллектив имеет сложную структуру, спектр всевозможных отношений, связей и взаимосвязей его членов весьма широк. Аппаратом описания структур коллективов, как и аппаратом описания отношений вообще является теория графов.

.Средством представления незримых коллективов является сеть (сеть идейного, творческого и пр. влияния) (рис. 2, 3). Звено сети (рис. 2) характеризует степень влияние 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 [14].

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

Неформальный коллектив из N элементов (N = 3) может быть представлен следующей матрицей NxN:

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

Здесь d_{ij} — количество ссылок j на i (иначе говоря, мера неформального воздействия i на j). Например, d_{13} — количество ссылок с на а, и наоборот,— количество ссылок а на с. Здесь также можно ввести меру m(x) неформального (идейного, научного и пр.) статуса индивидуума x, например, следующего вида:

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

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

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 рёбра и одновременно часть прежних элементов удаляется. Такие графы достаточно наглядно отображают перемены в коллективе, связанные, например, с уходом прежнего формального лидера.

Другим видом научных коллективов являются научные школы, информацию о которых можно получить на основе анализа таких реквизитов

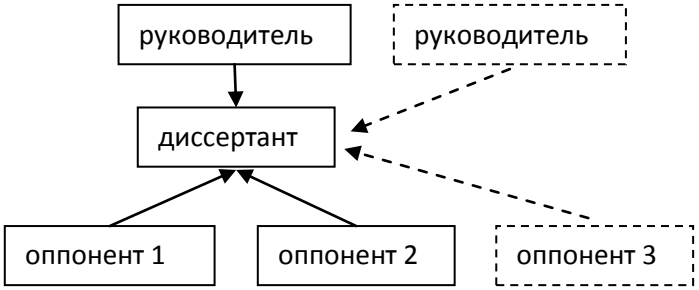

Рис. 4. Фрагмент графа диссертаций

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 [15].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 (см. 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.

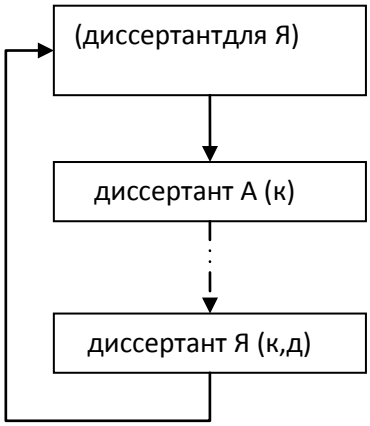

Рис. 5. Контур графа диссертаций

4. **Большие степени в графе** выявляют персон, оказавших большой 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.

Граф может иметь контуры. На рисунке ниже 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.

8 Методы извлечение 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

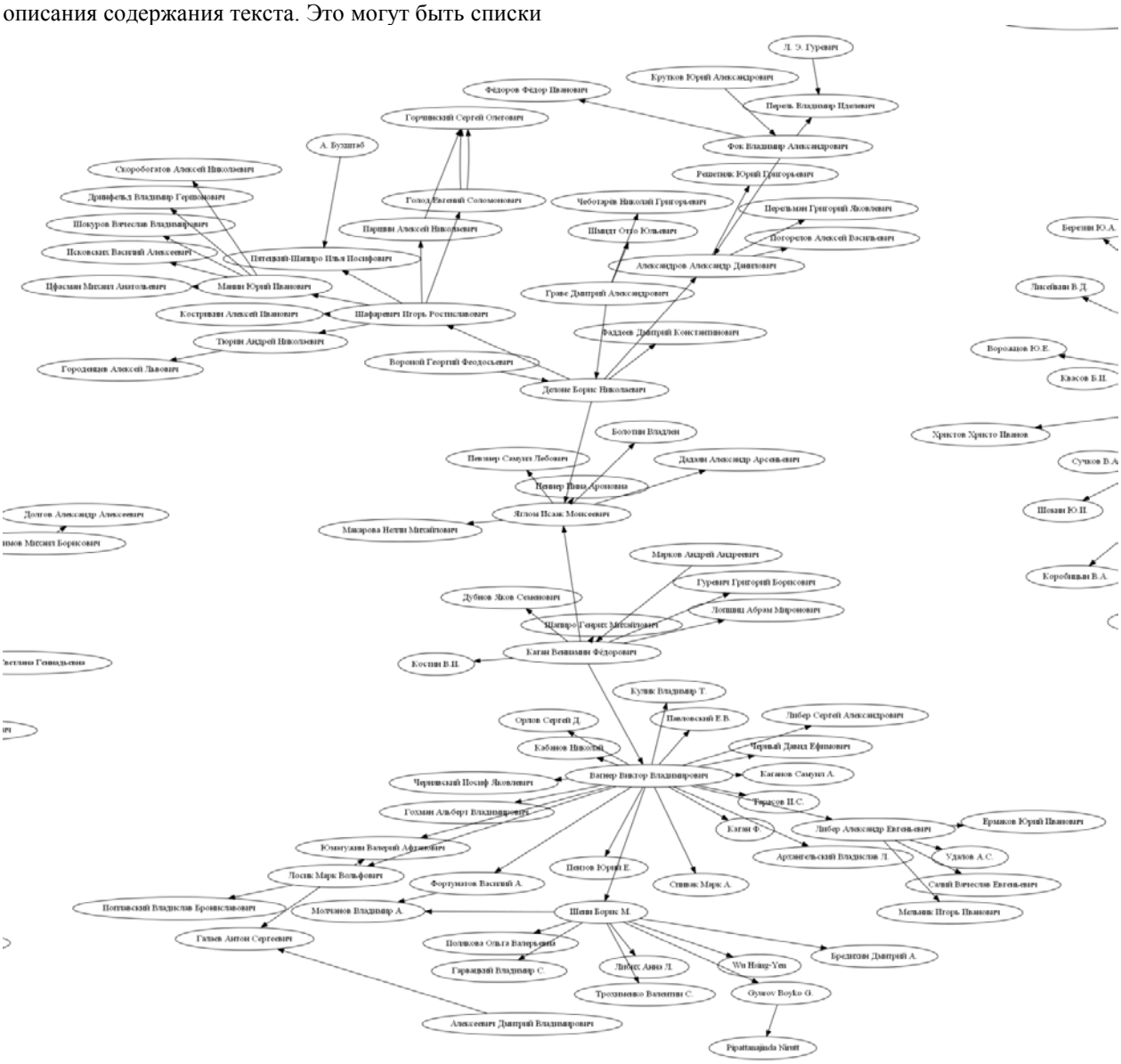

Рис. 6. Фрагмент графа диссертаций

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 используются 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, но 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.

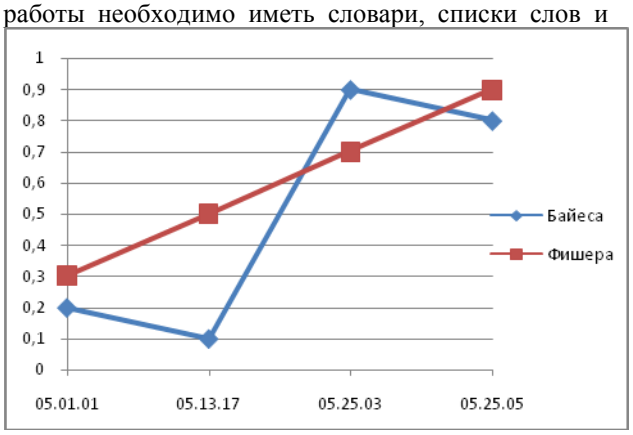

Рис.7. Точность. Зависимость от категории

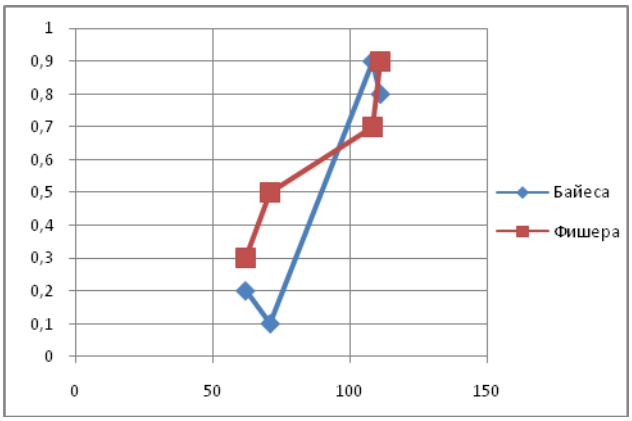

Рис.8. Точность. Зависимость от количества документов в рубрике

8.1 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;
- Отсутствуют строгие правила именования сущностей;
- Постоянно появляются новые сущности.

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 БарахнинВладимирБорисович]. Программныесистемы информационного обеспечения научной деятельности : модели, структуры и алгоритмы : диссертация доктора технических наук: 05.13.17 / Место защиты: [ORG Моск. гос. ун-тпечати].- [LOC Новосибирск], [TIME 2010].- 315 с.

Для извлечения именованных сущностей применяются несколько типов признаков [16]:

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.

Были проанализировано 4587 диссертаций и авторефератов и получен граф связей между персонами в диссертации на основании вышеприведенной модели (Рис. 4). Граф распадается на множество несвязанных компонент, в которых можно отыскать подграфы (Рис.6) с длинными цепями с длинной 2, что позволяет говорить о наличии научной школы.

8.2 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 [17,18].

В данной работе используется метод выделения терминов на основе морфологических шаблонов, аключевые 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.

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 [19]. 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 для таких 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 терминовприведены на Рис.7 и Рис.8., что позволяет сделать выводы о точности алгоритмов классификации около 90%при количестве документов в рубрике более ста.

Литература

[1] К.В. Бугаев Отграничение криминалистики от иных наук методами информационного анализа текста// Юридический мир. -2011. - № 8. - С. 40 – 43.

[2] Бескаравайная Е. В.. Анализ базы данных дисертаций ПНЦ РАН / Е. В. Бескаравайнова, И. А. Митрошин // Информационное обеспечение науки: новые технологии. - М.: Научный Мир, 2011. - С. 124-133.

[3] Прошанов С.Л.Докторские диссертации по социологии (1990-2010 гг.) // Социологические исследования. - 2011.-№1. - С.30-39.

[4] Липский С. И. Проблемно-тематический анализ диссертационных исследований по социальной педагогике (1971-2008 гг.) Автореферат диссертации, Кострома - 2009

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

[6] 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 - Vol. 34, Issues 6–7, 2009, P. 875–885

[7] DilekAltun, ÇağlaÖnerenŞendil, İkbal Tuba Şahin Investigating the National Dissertation and Thesis Database in the Field of Early Childhood Education in Turkey // Procedia - Social and Behavioral Sciences - Vol. 12, P. 1-654 (2011) - International conference on education and educational psychology, 2–5 December 2010, Cyprus

[8] Гайдадымов Евгений - Философия (Конспект лекций) //ЭЛЕКТРОННАЯ БИБЛИОТЕКА ModernLib.Ru

[9] Барахнин В.Б., Леонова Ю.В. Информационная модель отношений между документами в информационной системе. Вычислительные технологии. – 2005. - Том 10. Специальный выпуск. - С. 129-137.

[10] Концепция открытых систем // Материалы к межотраслевой Программе “Развитиеи применение открытых систем”. [http://www.informika.ru/text/inftech/opensys/3/concept/os_1.html]

[11] Большой Энциклопедический словарь. 2000

[12] О.Т. Манаев Контент-анализ как метод исследования // «ПСИ-ФАКТОР»

[13] Хайтун С.Д. Наукометрия: Состояние и перспективы. — М.: Наука, 1983.

[14] Элементы математической теории организации //Портал Cadmium http://cadmium.ru/content/view/832/45/

[15] 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 [ISBN 978-5-905569-05-0].

[16] Л.М. Ермакова Методы извлечения информации из текста // Вестник Пермского университета. Сер.: Математика. Механика. Информатика. - 2012. - Вып. 1 (9). - С. 77-84.

[17] Manning, C. D., and Schtze, H. 1999. Foundations of Statistical Natural Language Processing. The MIT Press.

[18] 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/

[19] Сегаран Т. Программируем коллективный разум. – Пер. с англ. – СПб: Символ-Плюс, 2008.

Extraction of knowledge and facts from texts of theses and abstracts for studying of communications of scientific communities

Yuliya V. Leonova, Anatolii M. Fedotov

In this work a research of theses and abstracts for the purpose of studying of structure of scientific communications of a scientist (a scientific environment of a scientist), structure and dynamics of development of research teams (schools of sciences), statistical research of the text of theses is undertaken. Such researches give the chance of studying and estimations of trends of development of various scientific directions, to identify persons, scientific centers and the organizations, schools of sciences, to study interrelations between separate communities.