Извлечение знаний и фактов из текстов диссертаций и авторефератов

Леонова Ю.В. , Федотов А.М.

Институт вычислительных технологий СО РАН

Новосибирск

Цель работы

- Целью данной работы является изучение связей научных сообществ, в рамках которых осуществляется научная деятельность, основанное на анализе диссертаций и авторефератов.
- Научное сообщество - 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 цели

- 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.

Выделение фактографической информации

- Значительную часть знаний о связях научных сообществ можно получить в результате анализа и синтеза информации из разобщенных фактов, размещенных в текстах диссертаций и авторефератов.
- При обработке документов процесс автоматического структурирования текстовой информации заменяет экспертный процесс выделения фактографической информации и объектов, выполняемый вручную.

Проблемы выделения фактографической информации

- Отклонение от работы с исключительно формальными признаками, формализуемыми текстовыми последовательностями и подстроками. Это заставляет обращаться к работе с некоторыми объектами (сущностями), отсутствующими в тексте в явном формализованном виде, но описанными автором.
- В обобщенной и абстрактной формулировке стоит задача выделения смысла текста. В конкретной формулировке стоит задача восстановления отдельных объектов и их взаимосвязей, которые были описаны, либо упомянуты, либо подразумевались автором неявно.

Выделение фактографической информации

- Диссертация является формальным квалификационным сочинением и её оформление должно соответствовать всем правилам, которые устанавливает ГОСТ для диссертации, что обеспечивает фиксированную структуру данного типа документов и позволяет реализовать методы извлечения фактов из диссертаций.

Анализ диссертаций 1

- Выделение тематически близких к криминалистике дисциплин из ряда иных наук
- Используются методы информационного анализ текста, применяя названия авторефератов диссертаций (АРД). Так установлена близость к криминалистике ряда юридических наук «криминального блока» по некоторым аспектам рядов динамики количества тем АРД, дескрипторов, информационной плотности названий АРД.

Результаты

- 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;

Результаты

- Индикаторы «эксцесс» и «асимметрия относительно среднего» могут служить для выделения весьма близких дисциплин (в данном 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).

Результаты

- 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.

Выводы

- 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.

Анализ диссертаций 2

- Анализ базы диссертаций центральной библиотеки Пущинского научного центра РАН.
- Проведен количественный и фактографический анализ диссертаций, защищенных в институтах ПНЦ РАН и в профильных организациях.

Результаты

- 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, разрабатывающие 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.

Анализ диссертаций 3

- Анализ 888 диссертационных исследований российских (советских) социологов на соискание ученой степени доктора социологических наук с момента открытия научной специальности - 1990 по июнь 2010 г. Анализ тем диссертаций докторов социологических наук показал, что авторами диссертаций использовано 1327 слов (понятий) или их производных (прилагательное, глагол и т.д.). После исключения предлогов, союзов осталось 1271 слово (понятие). Выделены наиболее часто употребляемые слова (понятия), которые использованы в массиве докторских диссертаций.

Результаты

- 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).

Анализ диссертаций 4

- 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.

Результаты

- 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.

Результаты

- 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).

Анализ диссертаций 5

- Анализ списков цитирования 49 диссертаций Индийского института менеджмента г. Ахмедабаб за период 2004-2009 г.г. Из 4319 ссылок цитирования были извлечены названия журналов и было произведено их ранжирование. Цель 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

Анализ диссертаций 6

- Из ссылок цитирования диссертаций за 1999–2003 г.г., полученных из ProQuest's Dissertations and Theses databases 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.

Анализ диссертаций 7

- Анализ 415 диссертаций за период 1987-2010 из Национальной базы данных диссертаций и авторефератов, содержащих словосочетание "Раннее детство" в качестве ключевых слов. Диссертации анализировались по параметрам: год, университет, подразделение и тематике.

Выводы

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

Информационная модель фактов

- Согласно «Логико-философскому трактату» Л.Витгенштейна, 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».

Информационная модель фактов

- 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.

Информационная модель 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).

Информационная модель фактов

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

Модель документа в системе

- Информационная система представляет собой множество связанных различными отношениями документов, описывающих некие сущности (объекты, факты или понятия).
- Информация о той или иной сущности содержится в системе:

- непосредственно в виде документа, который ее представляет, описывает или моделирует,
- в виде упоминаний об этой сущности, которые имеются в других документах, т. е. содержат опосредованную информацию об этой сущности.

Модель документа в системе

- Структура и содержание документа описываются 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).

Модель документа в системе

- Документ $d_i = <S_i, V_i>$,

  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.
- Информационная система представляется в виде набора коллекций.

Модель документа в системе

Метаданные:

- 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.).

Модель документа в системе

- Элемент схемы данных 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.

Модель документа в системе

- Структура документа - это набор структурных элементов.
- Содержание документа - объединение значений экземпляров элементов, составляющих документ.
- Информационная система содержит коллекции:

1) Персоны и организации, диссертационные советы

Модель документа в системе

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.
- Лингвистическая информативность реализуется в автореферате или диссертации в атрибуте «Текст».

Модель документа в системе

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. Примером терминов являются ключевые слова, описывающие содержание диссертации.

Модель отношений между документами в системе

- В основу нашей модели отношений между документами в информационной системе легла модель RDF.
- Выстраиваемые отношения переносятся на уровень элементов, определяющих структуру документов.
- Связи между документами устанавливаются путем задания на множестве документов бинарных отношений A(R,V): объект R имеет атрибут A со значением V.
- 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.

Модель отношений между документами в системе

- 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.

Модель отношений между документами в системе

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

Модель отношений между документами в системе

Виды связей:

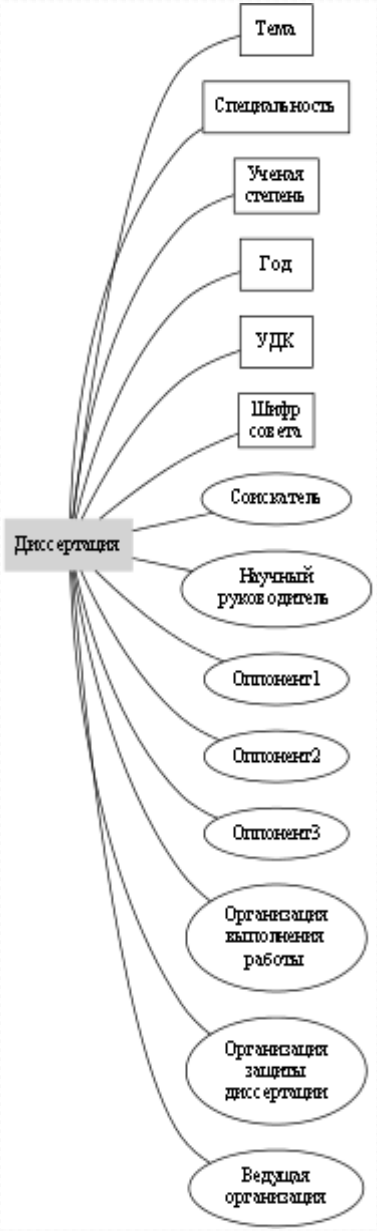
- **Прямые**. В этом случае есть факт о связи двух объектов, например, отношение соискатель-оппонент;
- **Нечеткие** (не представленные фактом):
  - **по общему месту и времени** у пары различных фактов различных объектов, например, дата и место защиты диссертации позволяет установить соискателей, защитивших 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

Модель отношений между документами в системе

- Высказывание: "Преподаватель Иванов А.А, родился в 1962 году", 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.)

Модель отношений между документами в системе

- Пример информационной модели описания диссертаций - связи между документом и его элементами.
- 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».



Модель отношений между документами в системе

- 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 модели:

Диссертация (Соискатель, тема, год, специальность, ученая степень, организация выполнения работы, организация защиты диссертации, ведущая организация, шифр совета, научный руководитель, оппонент1, оппонент2, оппонент3, УДК).

Модель отношений между документами в системе

- Для конкретных значений аргументов этот предикат превращается в 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.

Статистическое исследование текста диссертации

- При исследование 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

Статистическое исследование текста диссертации

- 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.

Статистическое исследование текста диссертации

- 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).

Статистическое исследование текста диссертации

- Основой содержания диссертации является принципиально новый материал, включающий описание новых фактов, явлений и закономерно-стей, 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.

Статистическое исследование текста диссертации

- Стадии контент-анализа:
- 1) 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.

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

Статистическое исследование текста диссертации

- 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».

Статистическое исследование текста диссертации

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.;

3) 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).

Исследование структуры научных связей ученого и научных коллективов

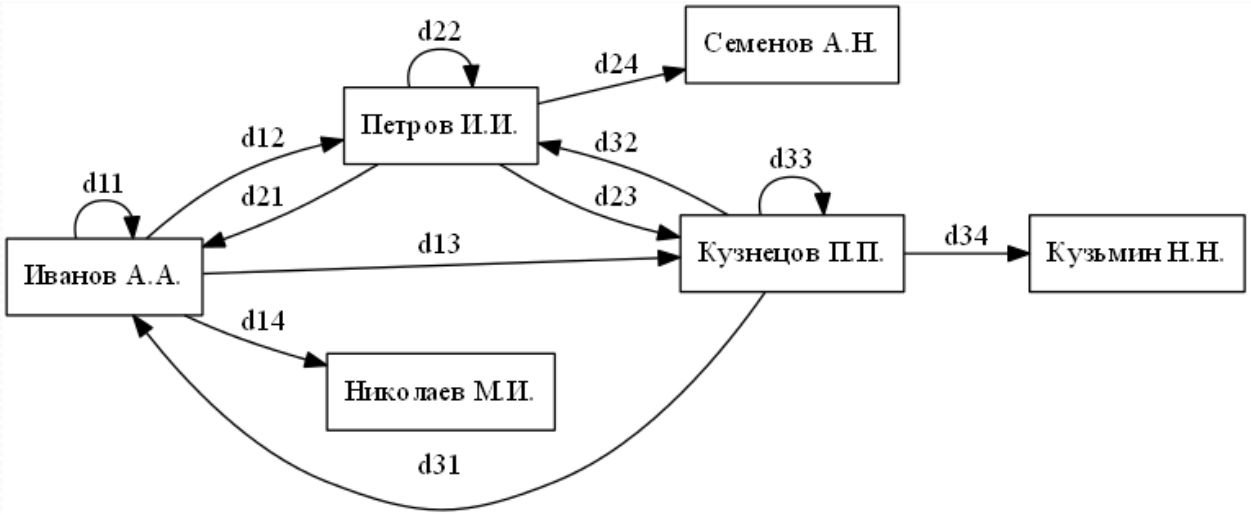
- Научное пространство учёного N определим как совокупность учёных {S}, связанных с N различными научными отношениями, как например, связи типа соискатель – научный руководитель, соискатель – оппонент, автор книги – редактор, автор книги – рецензент (не анонимный) и т.д.

Научные коллективы

- Коллектив - устойчивая во времени организационная группа взаимодействующих людей со специфическими органами управления, объединенных целями совместной общественно-полезной деятельности и сложной динамикой формальных (деловых) и неформальных взаимоотношений между членами группы. Коллектив имеет сложную структуру, спектр всевозможных отношений, связей и взаимосвязей его членов весьма широк.

Научные коллективы

- Звено сети характеризует степень влияние x на y, и может означать, например, что «y цитирует x» 10 раз.
- Если построить сеть взаимных ссылок, то можно выделить подграфы, элементы которых интенсивно связаны друг с другом. Такие подграфы образуют незримые коллективы (подграф (Иванов А.А., Петров И.И., Кузнецов П.П.) — научный неформальный коллектив).



Научные коллективы

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

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

- d_{ij} — количество ссылок j на i (иначе говоря, мера неформального воздействия i на j).
- Можно ввести меру m(x) неформального (идейного, научного и пр.) статуса индивидуума x, например, следующего вида: $m(a) = \frac{d_{13}}{d_{31}} + \frac{d_{12}}{d_{21}}$ или $m(a) = \frac{d_{13} + d_{12}}{d_{31} + d_{21}}$
- Эти меры используют различные выражения отношения «влияния a на остальных» к «влиянию остальных на a».

Научные коллективы

- 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.

Научные школы

- 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».
- Информацию о научных школах можно получить на основе анализа реквизитов диссертации: учебное заведение, в котором выполнена работа, научный руководитель, ведущая организация, дата и время защиты, шифр совета и т.д.

Научные школы

Структура графа диссертаций

- 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.

Научные школы

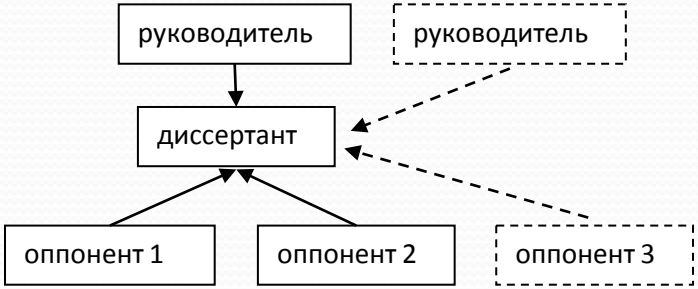
- **Вершины ориентированного графа диссертаций** соответствуют диссертантам, руководителям и оппонентам диссертантов. Бинарное отношение на парах вершин задается естественным образом: дуги выходят из вершин-руководителей и вершин-оппонентов и входят в соответствующую вершину-диссертант. Сохраняется информация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

Научные школы

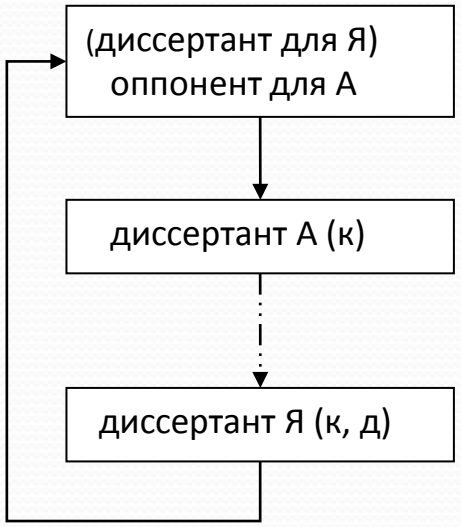
- **Из вершины-диссертанта дуга будет выходить**, если он в дальнейшем стал руководителем или оппонентом какой-либо диссертации.
- **Большие степени в графе** выявляют персон, оказавших большой влияние на формирование коллектива специалистов в данной области. Длинная цепь в графе показывает протяженный 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.

Научные школы

Фрагмент графа диссертаций



Контур графа диссертаций



Граф может иметь контуры. На рисунке 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.

Методы извлечение понятий из текста диссертации

Технологии получения информации в структурированном виде:

- 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.

Методы извлечение понятий из текста диссертации

- Методы автоматического извлечения понятий:
- 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.
- Наиболее эффективные – комбинированные методы

Извлечение именованных сущностей

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.

Для диссертаций и авторефератов выделяются: *люди* (PER), *места* (LOC), *организации* (ORG), *время* (TIME)

Извлечение именованных сущностей

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.

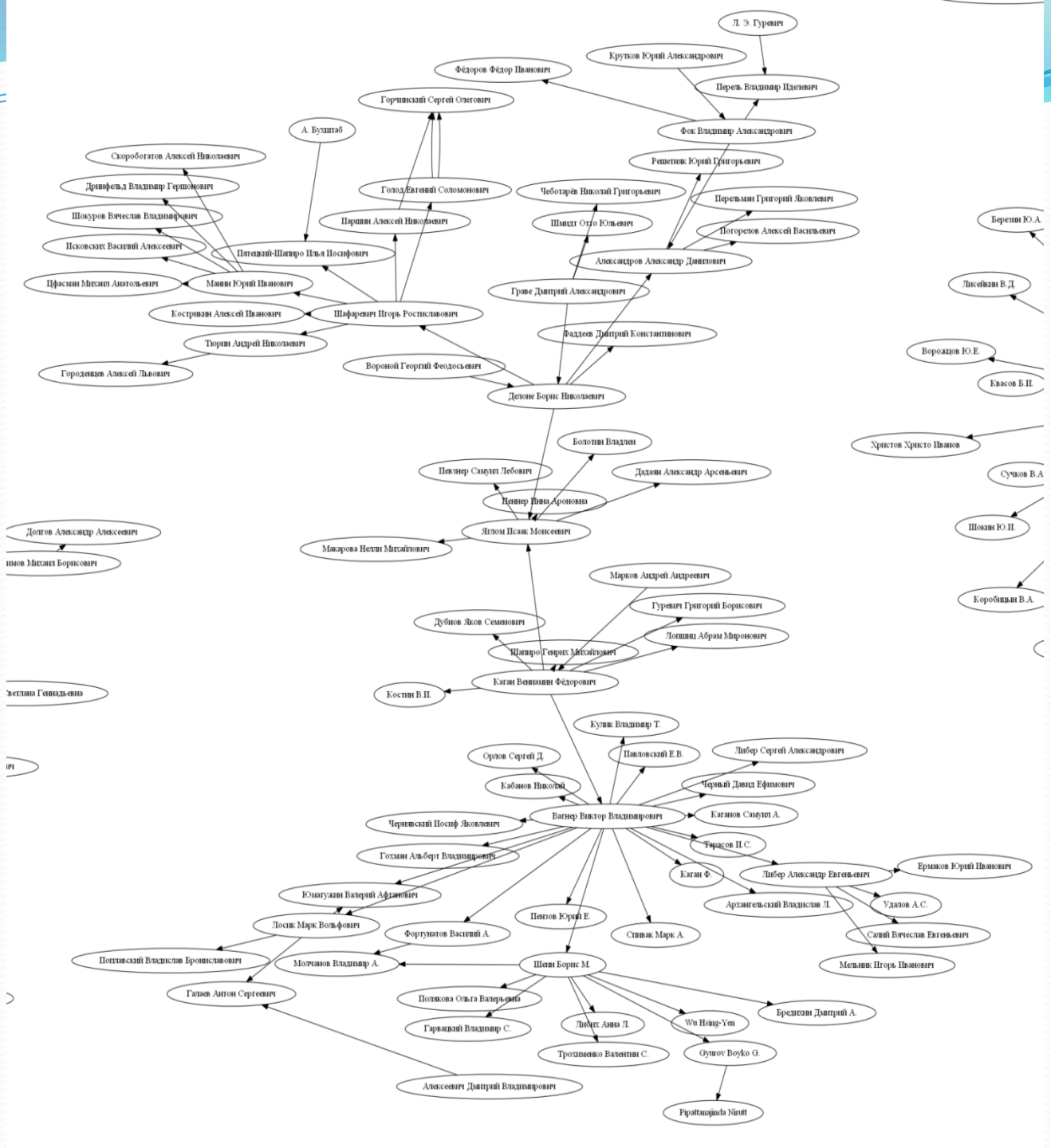
Извлечение именованных сущностей

Для извлечения именованных сущностей применяются несколько типов признаков:

- **признаки уровня слов** (N-граммы, суффиксы, префиксы, части речи и т.д.);
- **признаки уровня документа** (наличие акронимов в корпусе, позиция термина в предложении, наличие термина в заголовке или тексте и т.д.);
- **дополнительная информация** (слова указатели, например, Inc. ,Corp., списки стоп-слов, слов с капитализацией, которые не являются именованными сущностями и т.д.).

Извлечение именованных сущностей

- 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. Граф распадается на множество несвязанных компонент, в которых можно отыскать подграфы с длинными цепями с длинной 2, что позволяет говорить о наличии научной школы.


Л. Э. Гуревич
Крутков Юрий Александрович
Фёдоров Фёдор Иванович
Перель Владимир Ителевич
Горчинский Сергей Олегович
А. Бухштаб
Фок Владимир Александрович
Скоробогатов Алексей Николаевич
Решетняк Юрий Григорьевич
Дривфельд Владимир Гершонович
Голод Евгений Соломонович
Чеботарёв Николай Григорьевич
Перельман Григорий Яковлевич
Шокуров Вячеслав Владимирович
Паршин Алексей Николаевич
Шмидт Отто Юльевич
Погорелов Алексей Васильевич
Исковских Василий Алексеевич
Пятецкий-Шапиро Илья Иосифович
Александров Александр Данилович
Цфасман Михаил Анатольевич
Манин Юрий Иванович
Граве Дмитрий Александрович
Костривин Алексей Иванович
Шафаревич Игорь Ростиславович
Фаддеев Дмитрий Константинович
Тюрин Андрей Николаевич
Городенцев Алексей Львович
Вороной Георгий Феодосьевич
Делоне Борис Николаевич
Болотин Владлен
Певзнер Самуил Лебович
Дадаян Александр Арсеньевич
Певзнер Инна Ароновна
Долгов Александр Алексеевич
Яглом Исаак Моисеевич
Макарова Нелли Михайловна
Марков Андрей Андреевич
Гуревич Григорий Борисович
Дубнов Яков Семенович
Лопшиц Абрам Миронович
Шапиро Генрих Михайлович
Каган Вениамин Фёдорович
Костин В.П.
Кулик Владимир Т.
Орлов Сергей Д.
Павловский Е.В.
Либер Сергей Александрович
Кабанов Николай
Черный Давид Ефимович
Вагнер Виктор Владимирович
Черняевский Иосиф Яковлевич
Каганов Самуил А.
Гохман Альберт Владимирович
Тарасов И.С.
Каган Ф.
Либер Александр Евгеньевич
Ермаков Юрий Иванович
Юмагулин Валерий Афтанович
Архангельский Владислав Л.
Удалов А.С.
Лосик Марк Вольфович
Пензов Юрий Е.
Фортунатов Василий А.
Спивак Марк А.
Салий Вячеслав Евгеньевич
Поплавский Владислав Брониславович
Молчанов Владимир А.
Шеин Борис М.
Мельник Игорь Иванович
Галаев Антон Сергеевич
Полякова Ольга Валерьевна
Бредихин Дмитрий А.
Гарвацкий Владимир С.
Лихих Анна Л.
Wu Hsing-Yen
Трохименко Валентин С.
Gyunov Boyko G.
Алексеевич Дмитрий Владимирович
Pipattanajinda Nirutt
Березин Ю.А.
Лисейкин В.Д.
Ворожцов Ю.Е.
Квасов Б.И.
Христов Христо Иванов
Сучков В.А.
Шокин Ю.И.
Коробицын В.А.

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

Извлечение ключевых терминов из текста

Морфологические шаблоны:

- **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.

Извлечение ключевых терминов из текста

- 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.

Извлечение ключевых терминов из текста

- Преимущество наивных байесовских классификаторов заключается в том, что их можно обучать и затем опрашивать на больших наборах данных. 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.
- Обучение проводится инкрементно, – каждый новый предъявленный образец можно использовать для обновления вероятностей без использования старых обучающих данных.

Тестирование алгоритмов классификации. Методика

- Из обучающей выборки удаляются документы какой-то одной рубрики, они в обучении не участвуют. Однако, при тестировании документы этой рубрики присутствуют.
- Если документ из выкинутой рубрики определяется как «чужой», то считается, что это правильно найденный «чужой»

Тестирование алгоритмов классификации

- Варианты исходов для документа
- 1)«Прав»: документ из какой-то рубрики («Свой») правильно определился в свою рубрику;
- 2)«Чуж»: действительно «Чужой» документ определился как «Чужой»;
- 3)«Ошиб»: документ из какой-то рубрики определился не в свою рубрику;
- 4)«Св_чуж»: «Свой» документ ошибочно определился как «Чужой»;
- 5)«Чуж_св»: «Чужой» документ ошибочно определился как «Свой» — т. е. попал в какую-то рубрику
- 1 и 2 – правильная работа рубрикатора, остальные – ошибочные исходы

Тестирование алгоритмов классификации. Оценки

- Нахождение чужих:

Точность = Чуж / (Чуж + Св_чуж).

Полнота = Чуж / (Чуж + Чуж_св)

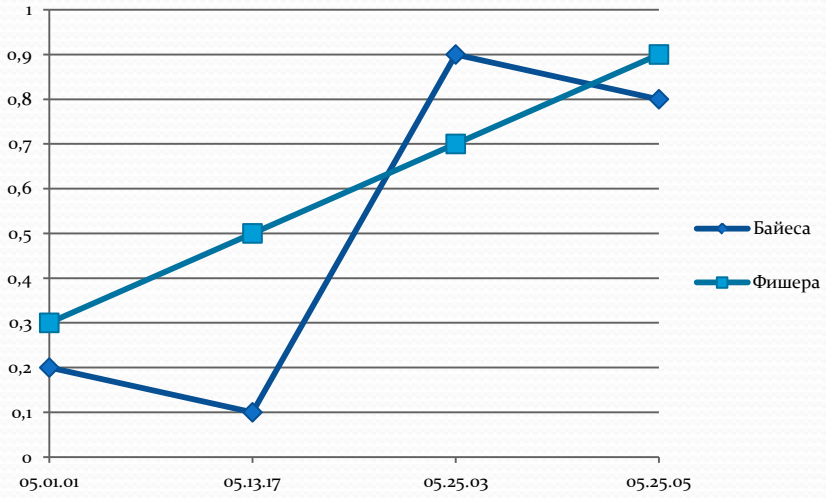
- Оценка классификации документов в рубриках

Точность = Прав / (Прав + Ошиб + Чуж_св)

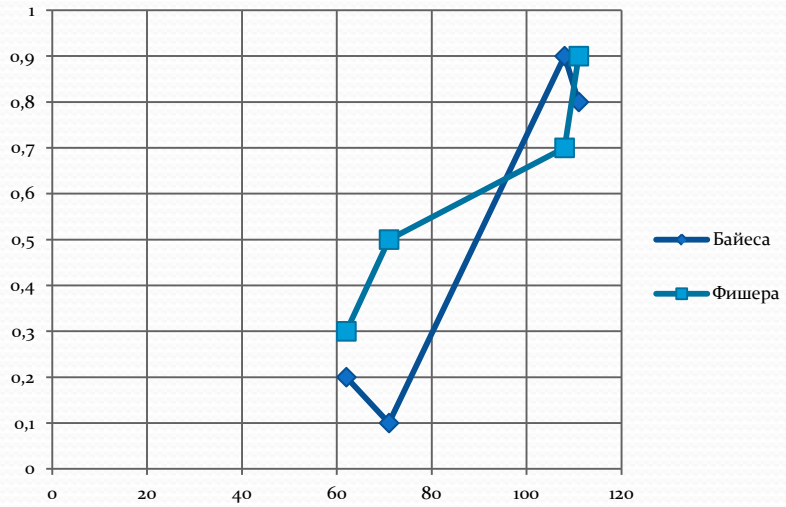
Полнота = Прав / (Прав + Ошиб + Св_чуж)

Результаты тестирования
Точность

Точность. Зависимость от категории



Точность. Зависимость от количества документов в рубрике



Результаты тестирования точности алгоритмов классификации терминов позволяют сделать выводы о точности алгоритмов классификации около 90%при количестве документов в рубрике более ста.