УНИВЕРСИТЕТ ИТМО

**О.В. Кононова, Д.Е. Прокудин**

# ТЕХНОЛОГИИ ИЗВЛЕЧЕНИЯ И ИНТЕЛЛЕКТУАЛЬНОГО АНАЛИЗА ДАННЫХ В НАУЧНЫХ ИССЛЕДОВАНИЯХ

**Санкт-Петербург**
**2021**

МИНИСТЕРСТВО НАУКИ И ВЫСШЕГО ОБРАЗОВАНИЯ РОССИЙСКОЙ ФЕДЕРАЦИИ

УНИВЕРСИТЕТ ИТМО

**О.В. Кононова, Д.Е. Прокудин**

# ТЕХНОЛОГИИ ИЗВЛЕЧЕНИЯ И ИНТЕЛЛЕКТУАЛЬНОГО АНАЛИЗА ДАННЫХ В НАУЧНЫХ ИССЛЕДОВАНИЯХ

УЧЕБНОЕ ПОСОБИЕ

РЕКОМЕНДОВАНО К ИСПОЛЬЗОВАНИЮ В УНИВЕРСИТЕТЕ ИТМО
по направлению подготовки 09.04.03 Прикладная информатика
в качестве учебного пособия для реализации основных профессиональных образовательных программ высшего образования магистратуры

УНИВЕРСИТЕТ ИТМО

**Санкт-Петербург**
**2021**

Кононова О.В., Прокудин Д.Е. Технологии извлечения и интеллектуального анализа данных в научных исследованиях. – СПб: Университет ИТМО, 2021. – 133 с.

Рецензент(ы):
Борисов Николай Валентинович, доктор физ.-мат. наук, профессор, старший научный сотрудник, профессор, заведующий Кафедрой информационных систем в искусстве и гуманитарных науках, Санкт-Петербургский государственный университет.

Учебное пособие посвящено вопросам анализа текстовых данных в рамках проведения научно-исследовательской работы, методам и средствам обработки и экспликации контекстного знания. Рассматриваются подходы и технологии информационного поиска, содержательного и интеллектуального анализа текстов. Учебное пособие может быть использовано как в учебных целях, так и в профессиональной научно-исследовательской деятельности.

Учебное пособие предназначено для магистрантов, обучающихся по образовательной программе «Цифровые технологии умного города» (09.04.03 «Прикладная информатика», специализация «Управление государственными информационными системами»), дисциплина «Компьютерные технологии в научных исследованиях» и образовательной программе «Умный город и урбанистика» (09.04.03 «Прикладная информатика», специализация «Управление государственными информационными системами»), дисциплина «Технологии извлечения и интеллектуального анализа данных в научных исследованиях».

Содержание

# Введение

Фундаментальность проблемы результативного поиска, экспликации и анализа контекстного знания связана с постоянно возрастающим объемом доступной информации, ее неформализованностью и слабой структурированностью, высокой скоростью обновляемости научных знаний, многозначностью используемой терминологической базы, переносом терминов из одной предметной области в другую без необходимой интерпретации и адаптации. Существующие инструменты поиска, экспликации и анализа контекстного знания мало востребованы из-за отсутствия содержательной информации как о самих инструментах, так и о методах и алгоритмах их применения в науке и на практике, а также в связи с ограниченностью предоставляемых аналитическим ПО наборов сервисов.

Учебное пособие «Технологии извлечения и интеллектуального анализа данных в научных исследованиях» направлено на формирование исследовательских и аналитических компетенций магистрантов. В пособии уделено внимание вопросам рационального отбора информационных источников, подходам и методам проведения тематического поиска в различных информационно-поисковых системах, анализу текстовых данных, методам и технологиям представления результатов научно-исследовательской деятельности.

Методологическая и содержательная новизна содержания пособия заключается в описании разработанного авторами комплексного подхода (синтетического метода) к поиску и выделению контекстов в междисциплинарных исследованиях и построению на его основе методики анализа и экспликации контекстов, методики построения и интерпретации трендов, что позволяет проводить исследования развития различных предметных областей и практик человеческой деятельности через изучение их понятийно-терминологического аппарата.

При подготовке магистрантов наиболее важным является формирование у них компетенций, связанных с применением информационно-коммуникационных технологий для поиска, отбора, извлечения и интеллектуального анализа научной и профессионально значимой информации в научно-исследовательской и проектной деятельности. В результате освоения дисциплины с использованием материала пособия магистранты смогут осознанно применять в научно-исследовательской практике синтетический метод в задачах поиска и анализа контекстного знания, использовать на базе этого метода аналитическое ПО и среды со встроенными сервисами экспликации, кластеризации и статистической обработки научных текстов, выявлять и соотносить контексты текстовой и нетекстовой модальности с областями знаний и объектом исследования, проводить интеллектуальный анализ данных, строить и интерпретировать тренды, демонстрирующие динамику развития междисциплинарных направлений исследований.

Использование подхода к интеллектуальной обработке данных на базе систем продвинутого полнотекстового и мультимодального поиска, методов и инструментов извлечения контекстного знания с одновременным овладением широким спектром аналитического инструментария позволит магистрантам, аспирантам и другим категориям исследователей повысить результативность аналитической и исследовательской деятельности.

Учебное пособие состоит из трех глав, которые необходимо рассматривать как дополняющие друг друга. В **первой главе** рассматриваются вопросы использования информационных ресурсов и данных из различных источников в аналитической и исследовательской работе ученого, задачи и перспективы информатизации научной деятельности, информационная культура ученого, применение информационно-коммуникационных технологий в научной коммуникации. В главе дается понятие и описание информационного пространства научных исследований. Особое внимание уделяется научным электронным ресурсам сети Интернет как одним из основных информационных ресурсов для научного использования.

**Вторая** глава учебного пособия описывает общие подходы к определению знаний, рассматривает последовательно континуумы Информация – Данные – Знание и Неявное знание – Контекстное знание – Контекст, определяет роль и виды контекста как аналитической единицы знаний, место контекста и контекстных знаний в отдельных предметных областях, новейших междисциплинарных направлениях исследований на базе применения ИКТ.

**В главе 3** рассмотрены подходы к работе с научными источниками информации, позволяющими повысить результативность поисковой и аналитической деятельности за счет применения информационных технологий поиска и обработки данных. Задачи первого этапа любого научного исследования, в том числе и науковедческого, базируются на выполнении информационно-поисковых задач по первичному вычленению текстов релевантной исследованию тематики из массивов информационных ресурсов большого объема, например, полнотекстовых Интернет-ресурсов, электронных библиотек и документальных баз данных. В настоящее время эти массивы информации доступны в цифровой форме и аккумулируются в различных сетевых информационных источниках. Поэтому актуальной является задача организации эффективного поиска релевантных цифровых информационных ресурсов с целью последующего экспертного отбора, анализа и использования. Осуществление этой задачи возможно за счет использования научных методов и подходов, а также поддерживающего их программного инструментария, обеспечивающего различные виды поисковых запросов; функции выявления контекстов, анализа и экспликации контекстного знания; визуализацию результатов. В главе рассматриваются автоматизированные методы и подходы к интеллектуальному анализу научных текстов, которые выглядят более

привлекательными применительно к задаче самостоятельного извлечения релевантных тем из больших корпусов текстов, прогнозирования развития научной тематики в рамках выбранной предметной области, в том числе разработанный авторами синтетический метод – комплексный подход к экспликации и анализу контекстного знания. Описание подхода приводится на примерах развивающихся междисциплинарных направлениях исследований, связанных с цифровой экономикой и цифровыми трансформациями отдельных сфер деятельности.

Учебное пособие предназначено для магистрантов, обучающихся по образовательной программе «Цифровые технологии умного города» (09.04.03 «Прикладная информатика», специализация «Управление государственными информационными системами»), дисциплина «Компьютерные технологии в научных исследованиях» и образовательной программе «Умный город и урбанистика» (09.04.03 «Прикладная информатика», специализация «Управление государственными информационными системами»), дисциплина «Технологии извлечения и интеллектуального анализа данных в научных исследованиях».

Учебное пособие может быть использовано в учебных целях, в научно-исследовательской работе магистрантов и в профессиональной научно-исследовательской деятельности. Пособие ориентировано также на аспирантов и руководителей научно-исследовательской работой магистрантов и аспирантов.

# Глава 1. Информационно-коммуникационные технологии в научной деятельности[1]

## 1.1. Информатизации научной деятельности. Информационная культура ученого

Развитие современного информационного общества характеризуется тотальным проникновением информационно-коммуникационных технологий (ИКТ) во все области человеческой деятельности, общественного бытия и все пространства человеческого существования. Процессы целенаправленного внедрения ИКТ в человеческую деятельность (прежде всего профессиональную) называют информатизацией. ИКТ находят своё применение, в том числе, и в самой научной деятельности, являясь одним из инструментов, необходимых при проведении научных исследований. При этом информатизация научной деятельности является внутренней потребностью научного сообщества и развивается в соответствии с логикой внутреннего развития научных исследований, потребностями учёных и исследователей. Информатизацию научной деятельности можно разделить на два основных направления:

– информатизация научных исследований;

– информатизация процессов управления и организации научной деятельности в обществе (на уровне государственных институций).

В исследовательской работе значимым и востребованным является первое направление. Однако информатизация предполагает не только внедрение ИКТ в научную деятельность, но и подготовку будущих учёных к их применению. При подготовке специалиста к участию в научно-исследовательской деятельности одной из основных целей является формирование практических умений использования информационно-коммуникационных технологий в своей научно-исследовательской деятельности. При этом основной целью является повышение эффективности профессиональной направленности этой деятельности. В современном информационном обществе достижение этой цели направлено на формирование информационной культуры как профессиональной компоненты учёного и исследователя, которая качественно выражается в формировании определённого уровня готовности к использованию информационно-коммуникационных технологий в своей профессиональной деятельности – научной деятельности. Формирование готовности возможно за счёт решения следующих задач:

– формирование знаний основ применения ИКТ в своей дальнейшей научно-исследовательской деятельности;

[1] Прокудин Д.Е. Информатика в гуманитарных науках: Учебно-методическое пособие. — СПб.: Фонд развития конфликтологии, 2016. — 52 с.

– формирование умений использования ИКТ в информационно-библиографическом поиске;

– формирование умений использования ИКТ в поиске, извлечении и анализе необходимых для проведения исследования данных (в том числе и научных и иных публикаций);

– формирование основ использования ИКТ для научной коммуникации;

– активизация учёных и исследователей для решения задач информатизации своей предметной области.

Готовность – это некоторое интегративное качество личности, прежде всего его психологическая характеристика. В рамках профессиональной деятельности «готовность личности — это регулятор и предпосылка эффективной и творческой деятельности. Наиболее значительными ее факторами являются мотивация, подготовка, самомобилизация знаний, установка на деятельность, свойства личности, удовлетворенность трудом»[2]. Иначе, готовность как интегральное состояние включает в себя совокупность мотивационных, эмоциональных, операциональных, интеллектуальных и волевых качеств[3].

Опираясь на эти трактовки, можно выделить следующие элементы готовности:

– мотивационные (потребности успешного использования ИКТ в научно-исследовательской деятельности, интерес к работе с ИКТ, стремление добиться профессионального успеха и т.д.);

– познавательные (понимание значимости использования ИКТ в научно-исследовательской деятельности, оценка значимости их применения, сформированность знаний и умений в данном виде деятельности и т.д.);

– эмоционально-волевые (ответственность учёного и исследователя за вопросы информатизации научной сферы, уверенность в успехе научной деятельности и т.д.).

Формирование готовности к использованию ИКТ в научно-исследовательской деятельности означает образование необходимого отношения к информатизации научной сферы, приобретение первоначального опыта и начал мастерства научно-исследовательской деятельности, опыта отбора и использования ИКТ в своей научно-исследовательской деятельности.

В зависимости от сформированности тех или иных элементов выделяются три уровня готовности специалиста к использованию ИКТ: начальный уровень пользователя, устойчивый уровень пользователя, квалифицированный пользователь. В качестве основного критерия рассматривается уровень мотивации.

---

[2] Дурай-Новикова К. М. Формирование профессиональной готовности студентов к педагогической деятельности: автореф. дис. … д-ра пед. наук / К. М. Дурай-Новикова. — М., 1983. — 18 с.

[3] Иванников А.Д., Тихонов А.Н., Цветков В.Я. Критерии готовности к использованию информационных технологий // Международный журнал прикладных и фундаментальных исследований. 2009. №3. С. 84-85.

*Начальный уровень пользователя* – уровень начинающего пользователя, при котором он имеет неустойчивый интерес, несмотря на то, что осведомлен о роли и месте ИКТ, об их функциональных возможностях и даже имеет навыки выполнения отдельных операций. Однако этого недостаточно для самостоятельного решения задач;

*Уровень пользователя* – пользователь самостоятельно справляется с типовыми задачами, аналогичными ранее решенным; его деятельность имеет устойчивый характер и в целом репродуктивна. Он имеет устойчивый интерес к ИКТ. Однако деятельность пользователя на этом уровне не простирается за рамки общепринятых подходов и методов.

Характерным является различие в квалификации разных пользователей уже на этом уровне. Психологически на втором уровне специалист опирается на стереотипный подход. Его умение определяется числом стереотипов, которые он освоил. Чем больше им освоено стереотипов, тем более квалифицирована его работа.

*Уровень квалифицированного пользователя* – пользователь решает как типовые задачи, так и новые, отличные от решенных ранее, видит возможности новых точек приложения ИКТ; в целом деятельность продуктивная. На этом уровне специалист опирается на стереотипный и аналитический подходы. Он может решать задачи и ставить новые, никем не решенные.

Распределение учёных и исследователей по этим уровням можно охарактеризовать с помощью следующих критериев:

*Уровень первый*. Осведомлённость в предметной области ИКТ, знакомство с терминологией, но недостаточно навыков по применению технологических приемов, методов и подходов.

*Уровень второй*. Компетентность в области ИКТ. Знание методов решения практических задач, знание технологий, но безотносительно к приложениям информатики и ИКТ. Это уровень знаний прикладной информатики. На этом уровне учёный-исследователь способен применить известные ему методы использования ИКТ в своей научно-исследовательской деятельности, а также адаптировать аналогичные подходы и методы.

*Уровень третий*. Универсальность или креативность. Самостоятельность не только в решении, но и в анализе и оценках. Знание теории, выходящее за рамки технологии. Этот уровень характерен способностью к творческой деятельности. На нем объем знаний соответствует уровню общей и прикладной информатики. Учёный-исследователь способен формировать собственные подходы к использованию ИКТ в своей научно-исследовательской деятельности, разрабатывать новые методики, участвовать в разработке соответствующего программного обеспечения.

Иначе можно охарактеризовать эти уровни готовности как:

1) элементарная готовность;

2) функциональная готовность;

3) системная готовность.

На формирование того или иного уровня готовности влияет также система подготовки будущих исследователей (как правило, в магистратуре или аспирантуре). Можно установить следующие соответствия формирования уровня готовности к использованию ИКТ системе подготовки будущих исследователей (по аналогии с системой подготовки будущих педагогов)[4]:

– элементарная готовность: система массово-репродуктивной подготовки;

– функциональная готовность: система массово-репродуктивной подготовки с элементами творческой деятельности;

– системная готовность: система индивидуально-творческой подготовки.

## 1.2. Информационно-коммуникационные технологии на различных этапах научно-исследовательского цикла. Общие методы использования информационно-коммуникационных технологий в научных исследованиях

Существуют различные подходы к определению структуры научно-исследовательской деятельности, «жизненного цикла исследования». Во многом они нацелены на формирование так называемого «бренда» учёного, что характерно для западного подхода к организации научной деятельности, в основе которого лежит не столько концепция «наука как общественное благо», сколько сугубо меркантильное отношение к научной деятельности как сфере, деятельность в которой обязательно должна приносить прибыль. Анализ этих подходов, а также анализ институализации научно-исследовательской деятельности в обществе позволяет выделить основные виды этой деятельности, регламентирующие её структуру и являющиеся инвариантными по отношению к предметной области, области знания, содержанию, методам и подходам конкретного научного исследования. К ним относятся[5]:

– информационно-поисковый, направленный на поиск и отбор источников исследования, анализ которых позволяет сформулировать цели и задачи исследования, выбрать подходы и методы, необходимые для решения задач и т.п. При этом формальным результатом этого этапа является составление библиографии по теме исследования;

---

[4] Кручинина Г.А. Готовность будущего учителя к использованию новых информационных технологий обучения (теоретические основы, экспериментальные исследования): Монография. Под ред. В.А. Сластенина. — М., Изд-во «Прометей», 1996. 176 с.; Печерская С. А. Теоретико-методологические основы готовности студентов к использованию информационных технологий. - Сочи: НОЦ РАО, 2007.

[5] Прокудин Д.Е. Проектирование и реализация комплексной информационной системы поддержки научных исследований // Технологии информационного общества в науке, образовании и культуре: сборник научных статей. Труды XVII Всероссийской объединенной конференции «Интернет и современное общество» (IMS-2014), Санкт-Петербург, 19 – 20 ноября 2014 г. С. 31-36.

– «констатирующий», на котором в результате мыслительной деятельности рождается новое научное знание, формально представляемое в форме научного текста;

– обнародование результатов научной деятельности в виде публикаций или выступлений с докладами на различных научных мероприятиях;

– научная коммуникация, дающая, с одной стороны, возможность получения «обратной связи» исследователю, а с другой – служащая установлению академических контактов, распространению научного знания в обществе, проведению совместных, коллективных исследований (научная коллаборация, collaboration (анг.) – сотрудничество).

Эти инвариантные и, следовательно, общие для всех исследований виды деятельности в развивающемся информационном обществе наиболее эффективно организуются и выполняются на основе всестороннего применения ИКТ. Говоря в общем об ИКТ, можно констатировать, что основным инструментом является электронно-вычислительная техника (компьютер), а в качестве самого доступного аппаратного обеспечения выступает персональный компьютер. При этом функционал его использования определяется тем или иным набором установленного программного обеспечения (ПО). С помощью того или иного прикладного ПО решаются конкретные задачи, в том числе и в рамках осуществления научной деятельности. В связи с этим, рассматривая роль и место применения ИКТ в структуре научной деятельности, можно выделить следующее ПО, методы использования которого также являются общими для любого исследователя:

1. Информационно-поисковая деятельность: обозреватели сети Интернет (задачи поиска), сетевые базы данных (специализированные научные ресурсы – электронные картотеки библиотек, полнотекстовые и реферативные базы научной информации, репозитории, агрегаторы, сетевые научные журналы и т.п.).

2. Констатирующая деятельность, в результате которой происходит подготовка научного текста: системы обработки текстовой информации (текстовые редакторы, системы оптического распознавания текстовой информации, системы компьютерного перевода, электронные словари), системы обработки графической информации (для подготовки иллюстраций), электронные таблицы (для представления результатов исследований) и т.д.

3. Предъявление результатов научной деятельности через публикацию в печатном или электронном издании, или выступление с докладом на научном мероприятии: системы обработки текстовой информации, программы подготовки и показа электронных мультимедийных презентаций, сетевое ПО.

4. Научная коммуникация: всё многообразие сетевого ПО.

В качестве интегрирующего и универсального можно выделить библиографические менеджеры (например, EndNote, Mendeley, Zotero), которые

используются во всех видах научной деятельности: для составления библиографии; для вставки в создаваемый в текстовом редакторе текст сносок, цитат и формирования списка цитируемой литературы; для организации совместной работы с библиографией (за счёт наличия облачной инфраструктуры). Также наличие функций социальной сети позволяет использовать библиографические менеджеры для установления академических контактов и поиска научных публикаций.

### 1.3. ИКТ в научной коммуникации. ИКТ в организации и проведении научных конференций

Важнейшей составляющей научной деятельности любого исследователя является коммуникация как в рамках научного сообщества, так и в обществе в целом. По своей сути она является информационной деятельностью и реализует различные функции, среди которых можно выделить как наиболее значимые:

– донесение до научного сообщества, общества в целом результатов научных исследований;

– распространение научных знаний в обществе (в том числе – популяризация научного знания);

– обмен мнениями, знаниями с коллегами;

– установление академических контактов;

– получение обратной связи на результаты своей научной деятельности;

– организация и проведение совместных научных исследований.

Только через коммуникацию можно решить и такую важную задачу, как профориентация молодёжи, из среды которой затем будут приходить в науку новые ломоносовы и ньютоны.

Современное общество предоставляет достаточно большое разнообразие ИКТ, которые используются для осуществления коммуникации. Среди них можно выделить:

1) системы интерактивного общения в режиме реального времени, такие как Skype, Oovoo. Они позволяют общаться одновременно с несколькими пользователями (например, Oovoo бесплатно поддерживает телеконференцию между 12 пользователями), а также передавать файлы, мультимединый контент и включать режим «белой доски» для одновременного показа содержимого окна какого-либо приложения (например, презентацию, электронные таблицы или другой иллюстративный материал). Эти возможности востребованы при проведении он-лайн научных семинаров, распределенных интернет-конференций, публичных лекций и докладов. А поддержка этих решений на различных платформах (MS Windows, Android, iOS и др.) и различных устройствах (планшетные компьютеры, смартфоны) расширяет возможности их применения и создают условия для повышения академической мобильности;

2) интернет-сервисы, представляющие собой специализированные социальные сети, ориентированные на научное сообщество. К ним относятся:

– Academia.edu (http://academia.edu) – научная социальная сеть, позволяющая устанавливать научные контакты для обмена опытом и организации совместной деятельности; искать исследователей по их научным интересам и публикациям, размещенным в открытом доступе на этом ресурсе; отслеживать научную активность по публикуемым участниками новостям; делиться информацией со своими коллегами; оставлять комментарии на статьи и другие материалы; завязывать дискуссии;

– ORCID (http://orcid.org) – сервис, присваивающий исследователю уникальный цифровой идентификатор и позволяющий размещать информацию о публикациях автора, участии его в грантах и других исследованиях. Поисковый механизм позволяет производить поиск и устанавливать контакты с исследователями на основании их публикаций, участия в научных исследованиях или ключевым словам их научных интересов.

– Помимо приведённых в этих целях используются и другие интернет-сервисы: ResearcherGate (http://www.researchgate.net), Linkedin (https://www.linkedin.com).

3) Онлайн системы для проведения заочных интернет-конференций. Таких систем достаточно много. Как примеры можно привести следующие ресурсы: научные конференции издательского дома «Научное обозрение» (http://russian-science.info/conferences); система виртуальных научных конференций Pax Grid (http:// www.paxgrid.ru); заочные электронные конференции Российской академии естествознания (http://econf.rae.ru); заочные электронные конференции консалтинговой компании Юком (http://www.ucom.ru/conf.html). Помимо публикации докладов участников, на сайтах электронных конференций зачастую организована система публикации отзывов (комментариев) на тексты докладов любым пользователем. Поэтому такие системы позволяют ликвидировать дефицит в общении между участниками очных конференций, которые в силу формата и регламента проведения (2 – 3 дня, ограниченное число участников, ограниченность во времени вопросов-ответов выступающему с докладом и т.д.) остаются за чертой внимания многих заинтересованных учёных – не все из них имеют возможность поучаствовать в очных конференциях, а особенно, если они проводятся в других городах или странах (отсутствие средств; невозможность оформить командировку). Таким образом, участники электронных заочных конференций могут получать обратную связь на результаты своих исследований, устанавливать академические контакты, проводить дискуссии.

Научная коммуникация в парадигме постнеклассической науки с тенденциями к глобализации и междисциплинарности играет огромную роль при организации и проведении совместных научных исследований. В этом плане сетевые информационные технологии являются самым востребованным

средством в научной среде. Всё чаще в совместной распределённой научной деятельности используются: облачные хранилища информации (Google Drive, Microsoft Sky Drive, DropBox, SugarSync, Яндекс.Диск и т.п.), которые позволяют обеспечить совместный сетевой доступ к хранимым данным; библиографические менеджеры (EndNote, Mendeley, Zotero, Citavi и т.п.), позволяющие создавать тематические библиографии на серверах соответствующих сервисов и предоставлять её в открытый доступ своим коллегам; интернет-лаборатории, предназначенные для организации информационного пространства при проведении совместных научных исследований[6].

Только коммуникация позволяет как доносить результаты научных исследований до научного сообщества, так и способствовать распространению научного знания в обществе. В основном это достигается через выступление с докладами на различного рода публичных мероприятиях. При подготовке выступлений используются средства создания и показа мультимедийных электронных презентаций. Каждому исследователю необходимо учитывать следующие принципы создания презентаций:

– использование режима ручного управления;

– презентация не должна отвлекать аудиторию от выступления;

– докладчик должен уделять презентации минимальное внимание (презентация для докладчика, а не наоборот);

– учёт эргономических и эстетических норм предъявления электронного изображения: освещённость помещения, размер и стиль используемых шрифтов, цветовое сочетание, использование средств дистанционного управления и т.п.

Также практика показывает, что при средней продолжительности выступления 7-10 минут для акцентирования внимания на основных положениях доклада количество содержательных слайдов не должно превышать 8-10 (исключение могут составлять данные экспериментов, графики, диаграммы, иллюстрации и т.п.).

## 1.4. Информационное пространство научных исследований

Эффективность научно-исследовательской деятельности зависит от информационного окружения исследователя, обеспечивающего оперативный доступ к необходимой для проведения исследования информации и способствующего научной коммуникации и коллаборации. В процессе своей деятельности исследователь формирует собственное информационное пространство. Однако, хотя для каждого оно специфично и уникально, всё же это пространство обладает общими характерными признаками, которые определяются общими для всех информационно-коммуникационными

[6] Прокудин Д.Е. Тенденции научной коммуникации в информационном обществе // Информация – Коммуникация – Общество (ИКО–2014): Материалы XI Всероссийской научной конференции / Санкт-Петербург, 23 –24 января 2014 г. – СПб., 2014. – 168 с. С. 132-135.

технологиями и методами их использования. К тому же институализация научной деятельности в обществе порождает её бюрократизацию и стандартизацию. Поэтому исследователи и учёные, осознанно или нет, формируют вокруг себя во многом стандартное информационное пространство.

Многообразие информации в различных научных информационных сетевых ресурсах зачастую не позволяет однозначно соотнести информацию с тем или иным учёным. В связи с этим для устранения неоднозначности необходимо идентифицировать личность учёного. Для этого разработан и используется универсальный идентификатор учёного, реализованный в открытом проекте ORCID (https://orcid.org). Этот проект является общедоступным и создан при поддержке ведущих мировых университетов, научных сообществ, научных издательств и других организаций, ориентированных на научную деятельность или её поддержку. В своём профиле пользователь может заполнить биографическую информацию, информацию об обучении, об опыте работы; информацию об участии в выполнении поддержанных фондами исследований, а также список своих публикаций, которые могут быть получены автоматически из основных баз данных научных публикаций и реферативных баз (например, CrossRef, PubMed, ResearcherID, Scopus и др.), что говорит об интеграции сервиса ORCID с внешними информационными системами. Определение в своём профиле списка ключевых слов своих научных интересов является основой механизма поиска персоналий из реестра сервиса и установления академических контактов. Наличие в ORCID сервиса для организаций позволяет:

– для научно-исследовательских организаций – анализировать информацию о своих сотрудниках через доступ к реестру;

– для издательств – управлять базой данных авторов и производить поиск потенциальных авторов и т.п.;

– для научных фондов – вести учёт грантов и публиковать информацию о них[7].

Этот универсальный идентификатор уже достаточно широко используется, например, для идентификации авторов публикаций и участников научных мероприятий.

В целом, характеризуя такое понятие, как «информационное пространство научных исследований», можно выделить основные функции, которые выполняют те или иные его элементы:

– поиск и доступ к научной информации: публикациям, материалам, библиографической информации, набором данных и т.п. Для этого используются различные цифровые научные ресурсы – электронные каталоги библиотек, полнотекстовые базы научной информации, реферативные базы научных публикаций, репозитории и открытые архивы научной информации, электронные

[7] Мбого И.А., Прокудин Д.Е., Чугунов В.А. Формирование информационного пространства междисциплинарного научного направления: подходы и решения // Межотраслевая информационная служба. 2015. №1. С. 36-44.

платформы научных издательств, сетевые агрегаторы и каталоги научной информации, электронно-библиотечные системы, научные социальные сети и т.п. Необходимо отметить, что во многих цифровых научных ресурсах есть функция подписки, по которой на указанный адрес электронной почты учёный может получать уведомления о новых публикациях по указанной тематике или по научным интересам, указанных при регистрации;

– обнародование и презентация результатов исследований. Здесь можно использовать всё многообразие открытых цифровых сетевых ресурсов: институциональные и тематические репозитории, открытые сетевые архивы, научные социальные сети. Помимо этого, научные публикации или метаданные о них без приложения усилий учёных размечаются на электронных платформах издательств, в наукометрических ресурсах, на сайтах конференций и т.д.;

– установление академических контактов и научная коллаборация. Эта функция реализована в научных социальных сетях, специализированных сетевых сервисах для организации совместных проектов (сетевые лаборатории) и других сервисах общего назначения (например, сетевые блокноты или сетевые сервисы библиографических менеджеров).

В зависимости от решаемых научных задач, тематики исследований и других индивидуальных особенностей научной деятельности конкретного учёного его информационное пространство может включать различные элементы, которые могут варьироваться в любых необходимых сочетаниях. К тому же с развитием информационно-коммуникационных технологий в это пространство могут добавляться новые элементы, которые добавят новый функционал. В дальнейшем в этой главе основные элементы информационного пространства научных исследований будут рассмотрены более подробно.

## 1.5. Информационно-коммуникационные технологии в научном информационном поиске

### 1.5.1. Основы научного информационного поиска

При поиске научной информации, как и для поиска любой другой, размещённой на электронных сетевых ресурсах, используются информационно-поисковые системы (ИПС). Основой поисковых систем являются так называемые поисковые машины, или автоматические индексы. Основным инструментом поиска в ИПС является запрос. Специальные программы-роботы в автоматическом режиме периодически обследуют Интернет на основе определенных алгоритмов, проводя индексацию найденных гипертекстовых страниц и документов. Созданные индексные базы данных используются поисковыми машинами для предоставления пользователю доступа к размещенной на сайтах сети Интернет информации. Пользователь в рамках соответствующего интерфейса формулирует запрос, который обрабатывается

системой, после чего в окно браузера выдаются результаты обработки запроса. Механизмы обработки запросов постоянно совершенствуются, и современные поисковые системы не просто перебирают огромное число документов. Поиск ведется на основе оригинальных и весьма сложных алгоритмов, а его результаты анализируются и сортируются таким образом, чтобы представленная пользователю информация в наибольшей степени соответствовала его ожиданиям. В каждую ИПС заложена конкретная технология представления и структурирования информации.

В поисковых системах при составлении запросов используется язык запросов, в который входят логические операторы («И», «ИЛИ», «НЕ»), метасимволы (например, «*», «?») и скобки, предназначенные для расстановки приоритета выполнения логических команд. Еще используются кавычки для указания ИПС вести поиск в точном соответствии с фразой, заключенной в кавычки.

Запросы бывают:

– простые. В таких запросах указываются фразы с использованием языка запросов, и после выдачи результата поиска приходится самостоятельно его анализировать и выбирать релевантный документ;

– сложные, которые реализуются через язык запросов и возможность расширенного поиска в рамках конкретной ИПС. При этом сужается область поиска, и в результатах поиска ИПС выдают меньше ссылок, но с большей вероятностью релевантности документов.

Как правило, в каждой ИПС есть достаточно разветвлённая система помощи, в которой помимо описания операторов языка поисковых запросов есть примеры составления сложных поисковых запросов, что позволяет эффективно использовать ИПС в том числе и для научного поиска.

Для поиска научной информации целесообразно использовать специализированный ресурс Академия Гугл (https://scholar.google.ru). Ядром этой информационной системы является информационно-поисковая система, в индексной базе которой содержаться ссылки только на научную и специализированную литературу.

### 1.5.2. Качество информации. Релевантность

Важным обстоятельством при информационном поиске является рациональный отбор из всех полученных результатов выполнения поисковых запросов максимально соответствующие потребностям пользователя, его ожиданиям. Степень соответствия полученной по запросу пользователем информации его ожиданиям характеризуется понятием релевантности. По-другому можно сказать, что релевантность – это степень соответствия документа запросу. Однако запрос редко может точно выразить информационную потребность. Поэтому ответственность за определение релевантности целиком и

полностью лежит на пользователе, который составлял поисковый запрос. А отсюда можно сделать вывод, что результаты поиска напрямую зависят от грамотности составления поисковых запросов.

Но даже полученная релевантная информация не всегда может быть использована. Особенно это относится к научной информации, на которую опираются учёные в своих исследованиях (например, ссылаются на неё или цитируют). При поиске информации в сети Интернет необходимо учитывать её качество. К основным критериям, отвечающим за качество информации, относятся достоверность, актуальность, точность и полнота.

Кроме этого, при поиске информации необходимо учитывать и другие аспекты:

– наличие большого объема недостоверной, некачественной информации, что связано с почти полным отсутствием цензуры в сети Интернет;

– дублирование информации, электронный плагиат, что увеличивает информационный шум и затрудняет отбирать первоисточники;

– наличие в сети Интернет огромного объема неактуальной и бесполезной информации. Это одна из самых насущных проблем современных сетевых информационных ресурсов, потому что нет разработанных эффективных методов контроля и утилизации информации. На сегодняшний день Интернет – это огромная информационная свалка.

Также в научном поиске необходимо учитывать репутацию и авторитетность как источника информации (автора), так и информационного ресурса, на котором он расположен. Перед использованием найденной информации (цитирование, ссылки и т.п.) необходимо найти первоисточник, так как в сети Интернет обычной практикой является так называемый «перепост» (размещение на ресурсе информации, взятой с другого ресурса), при котором зачастую происходит искажение первоначальной информации, а ссылки на первоисточник могут отсутствовать. Так, например, очень осторожно надо подходить к использованию в научных работах ссылки на статьи из электронной энциклопедии Википедия (http://wikipedia.org), которая является свободно редактируемой любым пользователем сети, а администрация этого ресурса официально отказывается от ответственности за качество и достоверность размещённой информации (https://ru.wikipedia.org/wiki/Википедия:Отказ_от_ответственности).

### 1.5.3. Возможности информационного писка при помощи ИПС. Понятие Deep Web

Далеко не всю размещённую в цифровых информационных ресурсах информацию возможно получить при поиске через ИПС. Это связано с тем, что не все гипертекстовые страницы в сети Интернет могут быть проиндексированы. Поисковые системы используют специальных роботов (англ. web crawler),

которые переходят по гиперссылкам и индексируют содержимое гипертекстовых страниц, на которых они оказываются. В принципе, всё информационное пространство можно разделить на (рис. 1.1):

– публичное (Public Web), которое индексируется поисковыми роботами. Поэтому информацию из этого пространства всегда можно найти через ИПС;

– «тёмную сеть» (Dark Web). Это также индексируемое пространство сети Интернет, но для доступа к нему необходимо использовать специальные технические средства (например, браузер Tor);

– «глубокая сеть» (англ. Deep Web) или «глубинный Интернет». В этом пространстве находятся гипертекстовые страницы, которые в принципе не могут быть посещены поисковыми роботами и, соответственно, ссылки на них не содержатся в индексной базе ни одной ИПС общего назначения.

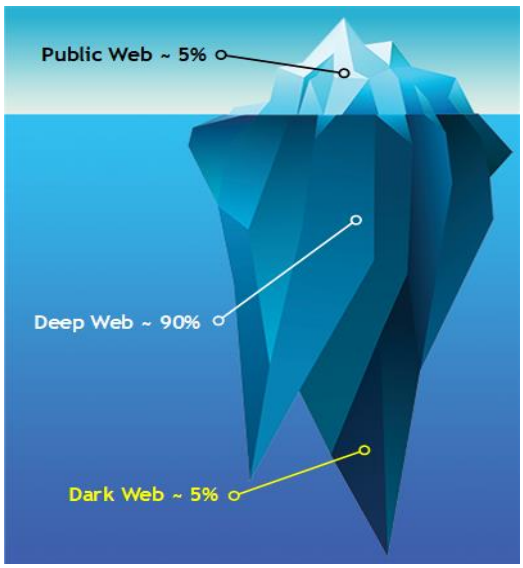



Рис. 1.1. Структура сети Интернет (источник: The Dark Web, Where Access to Your Business can be Bought, https://www.momentumit.com/dark-web-monitoring/)

В пространство Deep Web входят следующие типы гипертекстовых страниц или ресурсов (как группы связанных страниц):

– сайты, доступ к которым открыт только для зарегистрированных пользователей. Эти сайты, как правило, защищены механизмом аутентификации (например, использование логина и пароля);

– защищённые от поисковых роботов страницы (например, механизм CAPTCHA);

– страницы, динамически создаваемые по запросам к базам данных. Информация из баз данных, доступная пользователям через поисковые веб-формы (но не по гиперссылкам), остается недоступной для робота, не способного в режиме реального времени правильно заполнить форму значениями (другими словами, сформировать запрос к базе данных);

– содержимое информационных ресурсов, обладающих собственными поисковыми механизмами;

– страницы, переход на которые происходит после заполнения интерактивных формы;

– страницы с динамически формируемым содержанием.

Таким образом, значительная часть информационного пространства сети Интернет оказывается скрыта от поисковых роботов.

Помимо этого, владелец каждого веб-сайта может сам определить, какая часть его контента должна попасть в индексную базу данных ИПС и должна ли попасть туда вообще. То есть владелец может закрыть доступ на свой ресурс для поисковых роботов. Следовательно, информация с этого ресурса не появится в списках результатов поиска даже в том случае, когда на нее есть ссылки с сайтов, проиндексированных поисковыми системами.

В следующем параграфе будут рассмотрены категории электронных сетевых ресурсов, большая часть которых является представителями пространства Deep Web. Поэтому важным для учёного и исследователя является знание этих ресурсов и способов доступа к информации, размещённой на них.

## 1.6. Электронные сетевые научные ресурсы и другие источники данных[8]

Для поиска научных публикаций во всём мире используются цифровые научные ресурсы. Прежде всего, их можно разделить по доступности в них размещаемых текстов на общедоступные или открытые (как ресурсы, поддерживающие инициативу «Открытые архивы» и «Открытый доступ»), коммерческие (в основном платформы коммерческих издательств, а также реферативные базы WoS и Scopus) и смешанные, в которых есть как документы с открытым доступом, так и предоставляемые на платной основе.

Немаловажным аспектом для использования цифровых ресурсов в научных целях является степень детализации предоставляемой информации – в ресурсах могут содержаться либо только библиографические записи, либо библиографические записи могут сопровождаться метаданными, либо к этому ещё есть наличие полных текстов публикаций.

Систематизировать цифровые электронные ресурсы можно также и по типу размещаемых в них документов: монографии, периодические и сериальные издания, диссертации, сборники статей, материалы конференций, патенты, отчёты, депонированные рукописи, статьи и т.д.

Важной характеристикой для исследователей является степень охвата научных направлений, представленных в цифровых ресурсах:

[8] Прокудин Д.Е., Левит Г.С. Методы отбора цифровых информационных ресурсов на примере исследовании влияния научных идей Г.Ф. Гаузе на развитие науки // Научный сервис в сети Интернет: труды XIX Всероссийской научной конференции (18-23 сентября 2017 г., г. Новороссийск). М.: ИПМ им. М.В.Келдыша, 2017. С. 389-499. DOI:10.20948/abrau-2017-75

узкоспециализированные (ориентированные на несколько научных направлений) или политематические (которые не ограничены по тематике научных направлений).

### 1.6.1. Электронные каталоги библиотек

Повсеместное развитие глобальных сетей и доступность информационно-коммуникационных технологий дали толчок к разработке электронных каталогов и картотек классических библиотек. Электронные каталоги библиотек содержат основную долю библиографической информации в сети Интернет. С середины 1990 годов библиотеки стали постепенно переходить на использование для доступа к каталогам web-технологий, а также протокола Z39.50. В настоящее время электронные каталоги библиотек функционируют на основе баз данных и технологиях клиент-сервер, поэтому поисковые запросы к ним выполняются быстро.

Основными характеристикам электронных каталогов являются:

– осуществление поиска по всем значимым полям с возможностью усечения терминов запроса или вариативностью их употребления (любой из терминов, все термины только вместе, точная фраза);

– возможность ограничения поиска по формальным критериям (годы издания, вид издания, место издания, язык документа и т.д.);

– наличие словарей с возможностью автоматического переноса термина словаря в форму запроса (поисковое предписание);

– возможность формирования списка релевантных записей из общего перечня результатов поиска;

– возможность отправки отобранных результатов поиска по электронной почте или сохранения на локальном диске пользователя;

– наличие гипертекстовых ссылок на такие элементы библиографической записи, как авторы (индивидуальные или коллективные), предметные рубрики и название серии.

В наиболее мощных программных разработках существует возможность показа текущего состояния источника (выдан или находится на полке), а также ряд сервисных функций (сохранение истории поиска, представление записи в виде каталожной карточки, MARC-формата и т.д.).

Доступ к таким каталогам позволяет более оперативно получать информацию о библиотечных фондах, следить за их наполняемостью и даже заказывать через Интернет книги в библиотеках.

Помимо этих функций, электронные каталоги обладают дополнительными возможностями, из которым необходимо выделить:

– возможность электронного заказа;

– доступ к электронным текстам публикаций (далеко не всех, что связано с соблюдением авторских и издательских прав);

– поиск по содержаниям составных публикаций, например, сборников статей или материалов научных конференций. Такой функционал реализован в электронных каталогах, работающих на программном обеспечении ИРБИС.

Электронный каталог позволяет в случае отсутствия издания в своей библиотеке сразу направить читателя к месту хранения издания в другой. Среди электронных каталогов библиотек выделяются каталоги национальных библиотек (прежде всего — по объему каталогизированных в машиночитаемой форме фондов) — такие, как Библиотека Конгресса США, Британская библиотека, Национальная библиотека в Париже. Также важны для библиографического поиска каталоги университетских библиотек (например, Оксфорда), особенно при поиске редких старопечатных изданий и малотиражных научных монографий. При поиске книг на немецком языке незаменим каталог Библиотечного консорциума Карлсруэ. Электронные каталоги отечественных библиотек находятся в фазе бурного развития и пока не отражают в полной мере фонды, но многие из них уже имеют весьма значительный массив записей и предоставляют дополнительные поисковые возможности. Здесь можно выделить ГПНТБ России (на ее сайте находится не только каталог самой библиотеки, но и сводный каталог научно-технической литературы, позволяющий установить наличие книги в других библиотеках) и Российскую Государственную библиотеку, которая кроме описания новых поступлений переводит в форму электронного каталога книги XIX века, что позволяет установить наличие старой книги также в РНБ или ГПИБ. Сводный каталог иностранных книг по общественным наукам, поступивших в библиотеки России, ведется на сайте ВГБИЛ. Отдельно необходимо выделить электронные каталоги Российской национальной библиотеки и Библиотеки Российской Академии наук, функционал которых в полной мере отвечают запросам учёных и исследователей.

### 1.6.2. Полнотекстовые базы научной информации (электронные библиотеки)

Наиболее востребованными цифровыми ресурсами в научной среде являются полнотекстовые базы научной информации (называемые также «электронными библиотеками»), которые следует определить как упорядоченные коллекции разнородных электронных документов, снабженные средствами навигации и поиска. Одной из важных отличительных черт электронных библиотек является обязательное наличие заданной структуры и навигационно-поисковых средств, обеспечивающих ориентирование в документах. Это дает возможность исключить из понятия «электронная библиотека» беспорядочные файловые массивы, не объединенные единой иерархической структурой и системой индексации. Термин "разнородные" позволяет отделить электронные библиотеки от архивов периодических изданий, в которых представлены выпуски лишь одного издания.

Одно из главных преимуществ цифровых документов – поиск в них может производиться не только по сведениям титульного листа, но и по всему тексту.

Интенсивное развитие электронные библиотеки получили с началом эры глобальных компьютерных сетей. Доступность расположенных в глобальных сетях документов для огромной аудитории повсеместно привела к появлению электронных версий классических библиотек, а также специализированных полнотекстовых электронных библиотек, которые можно разделить на бесплатные электронные коллекции текстов и коммерческие полнотекстовые базы данных.

Наполнение полнотекстовых баз данных производится двумя способами. Первый заключается в сканировании печатных оригиналов и получении электронных копий документов, выполненных в большинстве случаев в формате PDF. Второй метод - покупка электронных копий книг, газет или журналов непосредственно в издательствах.

Наиболее качественный сервис и наиболее полную информацию представляют в сети Интернет коммерческие полнотекстовые базы данных (иначе можно их назвать цифровыми библиотеками). В них содержатся тексты книг, статей из журналов, газет и сборников, сообщения информационных агентств, аналитические отчеты различных учреждений и другие документы. Коммерческие цифровые библиотеки, в отличие от бесплатных коллекций, придерживаются намного более четкой политики отбора источников, обладают более высокой степенью полноты и оперативностью актуализации материалов.

Компании, непосредственно предоставляющие доступ к полнотекстовым собраниям, располагают сразу несколькими десятками или даже сотнями баз данных от разных производителей. Кроме этого, подобного рода полнотекстовыми цифровыми платформами обладают крупнейшие издательства, выпускающие научную литературу. На них они представляют цифровые копии своих изданий.

Наполнение полнотекстовых баз данных производится двумя способами:

– сканирование печатных оригиналов и получение электронных копий документов, выполненных в большинстве случаев в формате PDF. Сканирование печатных оригиналов применяется при оцифровывании существующих тематически и логически законченных собраний, хранящихся, как правило, в фондах крупнейших библиотек или архивов;

– покупка электронных копий книг, газет или журналов непосредственно в издательствах. По предварительному договору издательства передают электронную версию документа (чаще всего готовый оригинал-макет) поставщику и получают определенный процент от средств, полученных за обращение к поставленным файлам. Загрузка документов в систему осуществляется, как правило, в момент опубликования печатного оригинала или даже ранее.

Ну а цифровые платформы издательств наполняются самими издательствами.

Коммерческие цифровые библиотеки, в отличие от бесплатных коллекций, отличает намного более четкая политика в отборе источников, высокая степень полноты и оперативность актуализации материалов. Поисковый механизм позволяет осуществлять многоаспектный поиск с возможностью сочетания данных из разных полей. Поиск может осуществляться по отдельным словам, словосочетаниям и точным фразам. Результаты поиска выдаются в виде списка библиографических записей с указанием всех необходимых элементов. Существует возможность формирования из общего перечня списка релевантных документов.

Доступ к коммерческим базам осуществляется по предварительной подписке, которая оформляется, как правило, на один год. Подписку могут оформить организации, в которых проводятся научные исследования, в том числе высшие учебные заведения. Технически доступ осуществляется по предварительно выданным имени пользователя и паролю или по IP-адресу. Последний способ наиболее выгоден для организаций, так как позволяет предоставлять пользование такими базами данных с любых компьютеров, расположенных в учреждении. Практически все коммерческие базы данных позволяют в той или иной мере воспользоваться своими ресурсами без оформления платной подписки. Некоторые базы можно весьма эффективно применять в качестве библиографических источников, без возможности получать полные тексты документов.

Примеры наиболее значимых коммерческих полнотекстовых проектов:

– JSTOR (http://jstor.org) – электронные архивы, включающие научные материалы, опубликованные в более чем тысяче самых высококачественных академических журналах по гуманитарным, общественным и естественным дисциплинам, а также монографии и другие материалы, которые могут потребоваться при выполнении научной работы.

– LexisNexis (http://www.lexisnexis.com) - одна из крупнейших информационных корпораций мира. Комплекс баз данных LexisNexis включает огромный массив файлов, многие из которых представляют собой полнотекстовое содержание ведущих периодических изданий.

– ProQuest (http://www.proquest.com). Линия продуктов ProQuest включает большое число основных баз данных, среди которых отраслевые и тематические полнотекстовые собрания, реферативные и библиографические базы данных, электронные архивы известнейших газет и журналов.

– ScienceDirect (http://www.sciencedirect.com) включает полные тексты научных журналов и книг, базы данных рефератов, фундаментальные энциклопедические и справочные издания, изданных голландским издательством Elsevier.

– Springer Link (https://link.springer.com) – электронная платформа одного из крупнейших мировых научных издательств Springer, на которой размещены научные публикации начиная с 1840-х годов.

– EBSCO Information Services (https://www.ebsco.com) имеет в составе полнотекстовые базы данных, в числе которых материалы практически по всем отраслям знания.

– Интегрум (http://www.integrum.ru). В настоящее время – одна из самых крупных информационных онлайновых служб России. Среди представленных баз данных – архивы центральной, региональной и зарубежной прессы, сообщения агентств новостей, текстовые транскрипты передач радио и телевидения, тексты законов, данные Госкомстата России, электронные каталоги библиотек, сведения о патентах, адресные справочники, фотоархив и многие другие источники. Следует, однако, помнить, что значительный массив этих записей составляют библиографические описания книг и статей из библиотечных каталогов. Все источники интегрированы в единый информационный массив.

– Публичная библиотека (http://www.public.ru). Проект предназначен прежде всего для библиотек, которым предлагается оформить подписку на электронные версии российских центральных и региональных периодических изданий. Публичная библиотека дает возможность бесплатного библиографического поиска (опция «Открытый доступ») и возможность пользования полными текстами статей (опция «Профессиональный поиск»).

– EastView (http://www.eastview.com). В составе базы данных – центральные и региональные российские газеты, государственные стандарты, журналы Российской Академии наук, художественно-публицистические (толстые) журналы России, карты, статистические источники, материалы агентств новостей.

Среди отечественных научных цифровых библиотек можно выделить две самые основные:

– Научная электронная библиотека (http://elibrary.ru) – проект Российского Фонда фундаментальных исследований. На этой платформе представлены основные отечественные и зарубежные научные журналы (в том числе и журналы открытого доступа), а также научные монографии, сборники статей и труды научных конференций. Доступ к размещённым публикациям смешанный – здесь есть и свободно доступные публикации (для их получения только надо быть зарегистрированным пользователем), есть и публикации, которые можно получить на коммерческой основе.

– Научная электронная библиотека КиберЛенинка (https://cyberleninka.ru) – проект МГУ. Библиотека предоставляет доступ к статьям из журналов открытого доступа.

В нашей стране подписка на тематические и политематические полнотекстовые базы научной информации осуществляется всеми ведущими

университетами и научно-исследовательскими институтами и центрами. Помимо этого, существует национальная подписка на основные коммерческие научные ресурсы, которую осуществляет Минобрнауки России и предоставляет на конкурсной основе доступ к ним российским вузам.

### 1.6.3. Электронные научные журналы

Развитие издательского дела и, прежде всего, научной периодики привело к развитию электронных научных периодических изданий (электронных научных журналов). Электронные научные журналы выпускаются на коммерческой основе и доступны, как правило, на электронных платформах издательств. Также существуют журналы открытого доступа (Open Access), которые предоставляют свои издания свободно и без каких-либо ограничений.

В настоящее время существует три основных формы издания научной периодики с использованием информационно-коммуникационных технологий:

1. Журнал является печатным, но у него есть официальный сайт в сети Интернет (и на сайте даже могут быть размещены статьи и выпуски журнала в электронной форме). У такого журнала есть один ISSN (печатной версии).

2. У печатного научного журнала есть электронная версия, которая существует в виде сайта в сети Интернет. Тогда у такого журнала, как правило, есть два номера ISSN – для печатной (обозначается: Печ., Print, ISSN) и электронной версии (обозначается: Эл., Online, E-ISSN). Пример: печатное и электронное научное периодическое издание «Логико-философские штудии». Официальный сайт с представленной электронной версией: http://ojs.philosophy.spbu.ru/index.php/lphs.

3. Журнал был учреждён как электронный и существует только как электронный. У него есть один номер ISSN – это номер электронного журнала. Пример: электронное научное издание «Аналитика культурологии» (http://www.analiculturolog.ru), электронный мультимедийный журнал «Культура и технологии» (http://cat.itmo.ru).

Понимание того, что представляет собой сайт периодического издания, необходимо для грамотного оформления библиографической ссылки на статью из журнала. Так, например, на статью из печатного журнала можно ссылаться только как на печатный источник (даже если текст статьи найден в электронном виде на сайте в сети Интернет); на статью из журнала, у которого есть печатная и электронная версии, можно ссылаться как на печатный вариант, так и на электронный (правильнее будет ссылаться на тот вариант, который был прочитан); на статью из электронного научного журнала надо ссылаться только как на электронный источник. При этом необходимо помнить, что при формировании библиографической записи на электронный текст необходимо указывать полный адрес, по которому расположен текст статьи на официальном

сайте электронного журнала (а не на стороннем ресурсе, например, в Киберленинке или Научной электронной библиотеке).

### 1.6.4. Репозитории и агрегаторы научной информации

В развивающемся информационном обществе одной из важнейших являются задачи аккумуляции, архивирования и сохранения результатов научных исследований как на уровне организаций, так и на уровне научных сообществ. Эти технологии реализуют различные инициативы, основными из которых являются Будапештская инициатива «Открытый Доступ» (2001)[9], Берлинская декларация об открытом доступе к научному и гуманитарному знанию (2003, 2005)[10] и Международная петиция за гарантированный публичный доступ к результатам исследований, финансируемых Европейской Комиссией (2007)[11].

В рамках реализации подобных инициатив создаются и поддерживаются сетевые электронные научные репозитории, которые представляют собой электронные архивы для длительного хранения, накопления и обеспечения долговременного и надежного открытого доступа к результатам научных исследований и могут содержать следующие материалы: научные статьи, аннотации и диссертации, учебные материалы, книги или разделы книг, студенческие работы, материалы конференций, патенты, изображения, аудио- и видео-файлы, веб-страницы, компьютерные программы, статистические материалы, учебные объекты, научные отчеты и т.п. Как правило, подавляющее большинство электронных научных репозиториев придерживается принципа свободного доступа к результатам научных исследований, согласно которому работы предоставляются авторами на безвозмездной основе и могут быть свободно использованы другими учёными и исследователями. Размещением материалов в этих репозиториях занимаются сами авторы или уполномоченные лица, что полностью соответствует принципам самоархивирования, изложенным в Международном соглашении «Берлин-3», 2005 г., согласно которому организации должны создавать репозитории и предлагать своим сотрудникам (научным работникам) выкладывать в онлайновые архивы с отрытым доступом электронные копии всех опубликованных статей.

К основным преимуществам использования электронных научных репозиториев относят[12]:

---

[9] Будапештская Инициатива «Открытый Доступ» [Электронный текст] // Budapest Open Access Initiative. Russian Translation. URL: http://www.budapestopenaccessinitiative.org/translations/russian-translation

[10] Berlin 3 Open Access: Progress in Implementing the Berlin Declaration on Open Access to Knowledge in the Sciences and Humanities. Feb 28th - Mar 1st, 2005, University of Southampton, UK [Электронный текст] // http://www.eprints.org/events/berlin3/outcomes.html

[11] Земсков А.И., Шрайберг Я.Л. Системы открытого доступа к информации: причины и история возникновения [Электронный текст] // Науч. и техн. б-ки. - 2008. №4. URL: http://ellib.gpntb.ru/subscribe/ntb/2008/4/ntb_4_2_2008.htm

[12] Электронные репозитории [Электронный текст] // Научная библиотека Дагестанского государственного университета. URL: http://elib.dgu.ru/?q=node/739 (дата обращения: 24.04.2015); Электронные репозитории: какие

(а) для учёных и исследователей:

– возможность бесплатного размещения статей и препринтов;

– повышение индекса цитирования работ;

– постоянное и надежное хранение;

– сохранение авторских прав (путем их фиксации в институциональном информационном пространстве);

– бесплатный доступ к полнотекстовым статьям, содержащим результаты наиболее актуальных исследований;

– информационное пространство для коммуникации с авторами, работающими в близких областях;

– поиск соавторов и партнеров по исследовательским проектам;

(б) для научных организаций и учреждений:

– поддержка научной деятельности;

– повышение качества научных коммуникаций;

– повышение рейтинга научного учреждения или университета;

– обеспечение свободного доступа к результатам исследований;

– распространение результатов исследований сотрудников;

– повышение узнаваемости организации и её сотрудников в научном пространстве.

В настоящее время большое количество научных электронных научных репозиториев можно разделить на несколько основных категорий.

Первая группа – это полностью свободные как для авторов, так и для читателей электронные пополняемые архивы мирового масштаба. Ниже приведены основные ресурсы:

– ArXiv.org – крупнейший бесплатный архив электронных публикаций научных статей и их препринтов по физике, математике, информатике, астрономии, биологии, статистике, финансам, созданный в 1991 году в Лос-Аламосской национальной лаборатории (США), а в настоящее время координируемый Корнуэлльским университетом. При добавлении в архив публикация автоматически добавляется в базу цитирования Citebase (http://adsabs.harvard.edu), что, в том числе, позволяет оценить индекс цитирования авторов.

– SSRN – Social Science Research Network (http://www.ssrn.com) – один из самых крупных в мире открытых электронных репозиториев научных статей и препринтов, содержащий информацию по социальным наукам. Регистрация и публикация результатов исследований являются бесплатными, а каждый препринт и статья могут быть размещены как на личной странице автора в SSRN, так и в 12 тематических журналах, которые выходят только в электронном виде и рассылаются по подписке.

---

они и как в них зарегистрироваться? // Академическая среда. 2012. № 3(08). URL: http://www.hse.ru/data/2012/02/29/1265868424/Area_8.pdf

– SSOAR – Social Science Open Access Repository (http://www.ssoar.info) – один из самых крупных в мире открытых электронных репозиториев по социальным наукам (демография, социология, психология, пролитология, экономика и др.). Более 25 тыс. документов. Создатель - GESIS – Leibniz Institute for the Social Sciences (Германия, Кельн).

– DSpace@MIT (http://dspace.mit.edu) – открытый электронный репозиторий Массачусетского технологического института, состоящий из документов по различным отраслям знаний (наиболее полные коллекции по математике, физике, компьютерным технологиям).

– PhilSci-Archive (http://philsci-archive.pitt.edu) – открытый электронный архив по философии науки, созданный в 2000 г. специалистами Питтсбургского университета.

Перечисленные выше и другие многочисленные электронные репозитории и архивы открытого доступа вносят существенный вклад в развитие концепции «Открытой науки»[13] и поддержки информационного пространства научных исследований и научных коммуникаций в XXI веке[14].

Вторую группу электронных научных репозиториев составляют институциональные ресурсы, созданные и поддерживаемые различными научными организациями, учреждениями и университетами. Как правило, они полностью вписываются в концепцию «Открытого доступа» и предназначены для самоархивирования результатов научных исследований в виде статей, препринтов, научных отчётов, тезисов, авторефератов диссертаций научно-педагогическими сотрудниками соответствующих организаций. Такие репозитории в настоящее время имеются у практически всех зарубежных университетов, а также у ведущих российских университетов и научно-исследовательских институтов, например, у Санкт-Петербурского государственного университета (https://dspace.spbu.ru), Университета ИТМО (http://openbooks.ifmo.ru), Уральского государственного университета (http://elar.urfu.ru), Сибирского федерального университета (http://elib.sfu-kras.ru), Южно-Уральского государственного университета (http://dspace.susu.ac.ru), у Института вулканологии и сейсмологии Дальневосточного отделения РАН (http://repo.kscnet.ru).

Наличие огромного числа различных электронных научных репозиториев ставит проблему не столько поиска информации в них (как правило, все эти ресурсы имеют достаточно развитие поисковые возможности), сколько поиск самих этих ресурсов. Данная проблема решается за счёт развития агрегаторов,

---

[13] Паринов С.И. Развитие электронных библиотек – путь к Открытой Науке [Электронный текст] // Электронные библиотеки: перспективные методы и технологии, электронные коллекции: Труды XI Всероссийской научной конференции RCDL'2009. Петрозаводск: КарНЦ РАН, 2009. С. 225-234. URL: http://rcdl.ru/doc/2009/225_234_Invited-2.pdf

[14] Лагозе К. Связывая прошлое с будущим: Научные коммуникации в 21 веке [Электронный текст] // Электронные библиотеки. 2004. Т. 7, вып. 3. URL: http://www.elbib.ru/index.phtml?page=elbib/rus/journal/2004/part3/kl

которые представляют собой информационные системы, выполняющие роль сетевых каталогов научных электронных ресурсов. Агрегаторы не содержат самих текстов – они содержат сведения об электронных репозиториях со ссылками на них, а также в их электронных базах аккумулируются метаданные статей и других научных текстов, размещённых на соответствующих ресурсах. Эти системы обеспечивают упорядоченность и доступность многочисленных электронных репозиториев научной информации, поэтому их ещё называют электронными каталогами или регистрами репозиториев.

К самым известным агрегаторам научной информации относятся:

– The OAIster® database (http://oaister.worldcat.org) – один из самых мощных мировых агрегаторов, который содержит более 30 млн. записей метаданных публикаций более чем 1,5 тыс. организаций-участников, предоставляющих свои публикации по принципу «открытого доступа» 15. Эта разработка Мичиганского университета в настоящее время поддерживается и развивается Online Computer Library Center, Inc. (OCLC), являющейся научно-исследовательской организацией, общественной целью которой является расширение доступа к мировой информации и сокращение расходов на информацию[16];

– Registry of Open Access Repositories (ROAR - http://roar.eprints.org), созданный в Саутхемптонском университете. В настоящее время в ROAR зарегистрированы 3,3 тыс. архивов, в том числе всего 53 из России;

– Directory of Open Access Repositories (OpenDOAR - http://www.opendoar.org) создан и поддерживается Нотингемским университетом (Великобритания). В настоящее время OpenDOAR содержит 2845 архива, из которых 22 являются российскими;

– Directory of Open Access Journals (DOAJ - http://doaj.org) – каталог научных журналов открытого доступа, идея создания которого получила одобрение в 2002 г. на Первой Скандинавской конференции по проблемам научных коммуникаций в г. Лунде (Швеция). Создан совместно и при поддержке Лундского университета, Шведской национальной библиотеки и Королевской библиотеки (Стокгольм). В каталог включены рецензируемые электронные научные журналы свободного доступа. В настоящее время DOAJ предоставляет поиск на уровне статей из 10 тыс. журналов (около 2 млн статей).

Отдельного внимания заслуживает отечественная информационная система «Соционет», обеспечивающая информационную поддержку научно-образовательной деятельности в социогуманитарных, экономических и других научных дисциплинах. Эта система выполнена в рамках международных

[15] OCLC. OAIster Contributors [Электронный ресурс]. URL: http://www.oclc.org/oaister/contributors.en.html

[16] OCLC: союз библиотек и удобство читателей [Электронный текст] // Российская государственная библиотека. URL: http://www.rsl.ru/ru/s7/s409/2013/20137642

инициатив RePEc и Open Archives Initiative [17] и представляет собой платформу для создания информационных ресурсов и сервисов, адресованных профессиональным научным сообществам. [18]

Особенностями системы являются:

– поддержка протокола OAI-PMH на уровне сборщика метаданных;

– автоматизированный сбор метаданных с провайдеров (синхронизация) по протоколу OAI-PMH в соответствии с расписанием. Например, для ежеквартальных электронных сетевых научных журналов можно установить режим синхронизации один раз в три месяца;

– поддержка основных форматов метаданных (например, MARC, Dublin Core).

Это позволяет полностью автоматизировать сбор метаданных с подавляющего большинства электронных научных репозиториев. Администратору репозитория необходимо только зарегистрировать свой ресурс в выбранном агрегаторе и настроить процесс синхронизации.

### 1.6.5. Разнородные цифровые ресурсы для научного использования

Помимо специализированных электронных ресурсов, которые предназначены для предоставления учёным и исследователям доступа к научным публикациям и научной информации, существует довольно много сетевых ресурсов, которые в той или иной степени востребованы в научной деятельности. Большинство из них являются представителями публичного пространства сети Интернет и индексируются информационно-поисковыми системами (ИПС). Однако зачастую информация из этих ресурсов тонет в информационном шуме, который пользователи получают по поисковым запросам к ИПС. Поэтому выделим некоторые востребованные в научном сообществе категории таких разнородных ресурсов и опишем их основные достоинства.

1. Сайты научных и иных организаций. К такого рода ресурсам можно отнести сайты научно-исследовательских институтов и центров. В России, например, эти учреждения относятся к Российской Академии наук. На этих сайтах можно ознакомиться с научной жизнью данных учреждений, узнать анонсы предстоящих научных публичных мероприятий, найти библиографические списки основных научных трудов сотрудников этих организаций, а также полные тексты некоторых публикаций и изданий.

2. Сайты высших учебных заведений, факультетов и кафедр. На этих ресурсах, как правило, размещается следующая информация:

– анонсы научных мероприятий, проводимых указанными структурами;

---

[17] О базе данных системы Соционет [Электронный текст] // Соционет – научная информационная система. URL: http://socionet.ru/bd.htm

[18] Паринов С.И., Ляпунов В.М., Пузырев Р.Л. Система Соционет как платформа для разработки научных информационных ресурсов и онлайновых сервисов [Электронный текст] // Электронные библиотеки. 2003. Т. 6. № 1. С. 6-25. URL: http://elibrary.ru/item.asp?id=9121156

– объявления о защите диссертаций на соискание учёных степеней. Притом, в настоящее время в открытом доступе в этих объявлениях публикуются полные тексты авторефератов диссертаций и тексты самих защищаемых работ, а иногда и ссылки на интернет-трансляции защит;

– на сайтах факультетов и кафедр публикуются в открытом доступе научные тексты и публикации сотрудников, переводы зарубежных учёных и иные научные тексты;

– сведения о научной деятельности организаций и их подразделений и т.д.

3. Личные страницы учёных. Как правило, существует два основных варианта их представления – либо как самостоятельный интернет-ресурс, либо в качестве раздела на институциональном ресурсе. Эти ресурсы могут в той или иной степени полноты содержать информацию о личности учёного: его научные интересы, участие в научной деятельности (например, член редколлегии научного издания или участие с выступлением на научной конференции); основные научные публикации и т.п.

4. Сайты научных сообществ. Это могут быть как институционализированные сообщества (как, например, Российское философское общество, Российское географическое общество, Санкт-Петербургское философское общество и т.д.), так и сообщества, организованные на принципах совместной научной деятельности (различные научные центры, ассоциации и тому подобные сообщества), деятельность которых никак не регламентирована уставными документами. На этих ресурсах содержится различная информация о научной деятельности этих сообществ и учёных, ассоциированных с ними.

5. Сайты научных конференций (конгрессов, симпозиумов и т.п.). Здесь публикуются анонсы и объявления о предстоящих организующихся научных мероприятиях, что позволяет учёным принимать решение об участии в них. Также, как правило, через эти ресурсы происходит процесс регистрации участников и иногда подача материалов. На сайтах регулярных научных конференций можно обнаружить архивы материалов прошедших мероприятий (сборники тезисов или сборники статей), которые не всегда доступны даже в специализированных научных библиотеках.

6. Сайты научных издательств. Конечно, на данный момент библиографически грамотно в сети представлены преимущественно зарубежные издательства, публикующие книги научного характера. Как правило, на своих сайтах они дают не только сведения о новинках, но и аналоги своих издательских каталогов с довольно большим временным охватом. Если в данном издательстве выходят также журналы, то зачастую предоставляется и возможность поиска в них статей, причем аннотированных или даже с предоставлением полных текстов. Примером может служить издательство Wiley. При разыскании иностранной малотиражной научной литературы, в частности, материалов

различных конференций, плодотворным может быть обращение к сайтам различных институтов, проводивших те или иные конференции или издававшим монографии. При этом необходимо учитывать, что для обращения к научным журналам в сети чаще всего требуется оформление платной подписки. К таким, например, можно отнести: издательство «Наука» (http://www.naukaran.ru), издательство Санкт-Петербургского государственного университета (http://unipress.ru), издательство «Алетейа» (http://www.aletheia.spb.ru), издательский дом «Петрополис» (http://www.petropolis-ph.spb.ru). На этих ресурсах, как правило, можно ознакомиться с архивом изданных научных книг, а также с книжными новинками и планом издания на ближайшее время (как правило, на текущий год). Это позволяет не пропустить выходящие в издательствах научные издания, тиражи которых значительно меньше, нежели у ненаучной литературы.

## 1.7. Информационно-коммуникационные технологии при работе с библиографической информацией и в научном цитировании

Важным аспектом информационно-поисковой деятельности является составление библиографии по тематике проводимого исследования, источники из которой затем используются при оформлении результатов исследования в форме статей и иных текстов. При этом необходимо понимать, что библиографические ссылки должны быть грамотно оформлены как по форме, так и по содержанию.

При цитировании источников часто бывает необходимым ссылаться на тексты, размещённые в сети Интернет или на определённые сетевые ресурсы (сайты). Основываясь на существующих ГОСТах, можно предложить следующие варианты таких библиографических записей:

1. Статья из электронного научного журнала:

Микиртумов И. Б. Элементы логики Эрнста Малли: предметно-теоретические основания и проблемы интерпретации [Электронный текст] // Логико-философские штудии. Том 12. № 1. 2014. URL: http://ojs.philosophy.spbu.ru/index.php/lphs/article/view/38/38 (дата обращения: 01.01.2015).

2. Статья из сериального издания, которой присвоен DOI:

Возмищева Т.Г. Модифицированная модель войны или сражения и гонки вооружения на основе модели Лотки–Вольтерра как модель конфронтации государств: численный и качественный анализ // Информационное общество: образование, наука, культура и технологии будущего. Выпуск 4 (Труды XXIII Международной объединенной научной конференции «Интернет и современное общество», IMS-2020, Санкт-Петербург, 17 –20 июня 2020 г. Сборник научных статей). —СПб: Университет ИТМО, 2020. С. 72-91. DOI: 10.17586/2587-8557-2020-4-72-91.

3. Текст, размещённый в сети Интернет:

В Кисловодске открыт первый музей Солженицына [Электронный текст] // ТАСС: информационное агентство России. URL: http://tass.ru/kultura/2008247 (дата обращения: 01.06.2015).

4. Ресурс в сети Интернет:

Научная электронная библиотека [Электронный ресурс]. URL: http://elibrary.ru (дата обращения: 01.06.2015).

Для работы с библиографической информацией в настоящее время разработаны и используются библиографические менеджеры. Прежде всего это программное обеспечение, представляющее собой базу данных, предназначенную для хранения библиографической информации. При этом каждая библиографическая запись представляет собой набор метаданных, а наличие поискового механизма позволяет не только эффективно производить поиск по метаданным, но и осуществлять полнотекстовый поиск по загруженным текстам, которые ассоциированы с соответствующими библиографическими записями. Помимо решения задачи организации библиографии, её классификации и упорядочения, библиографические менеджеры также позволяют выполнять ряд других функций: автоматическое формирование библиографических описаний со страниц сетевых ресурсов, вставка в текстовом редакторе библиографических ссылок и автоматическое формирование списков использованной литературы, организация совместной работы над составлением библиографии и пр. Это возможно благодаря наличию динамически подключаемых к обозревателям сети и текстовым редакторам программных модулей, а также применению облачных технологий и элементов социальных сетей.

Одним из наиболее популярных является сервис библиографического менеджера Mendeley (http://mendeley.com). Основным инструментом является одноимённый свободно распространяемый программный продукт, который позволяет на персональном компьютере пользователя организовать упорядоченное хранилище различных тематических библиографий, производить по ним эффективный поиск и получать оперативный доступ к соответствующим ресурсам – файлам, хранящимся в локальной системе, или ресурсам сети Интернет. Наличие программных модулей для интеграции Mendeley с различными браузерами и популярными текстовыми редакторами позволяет расширить возможности рассматриваемой системы. В частности, использовать ее для оперативного импорта данных в локальное хранилище найденной в сети Интернет информации, для вставки в тексты библиографических ссылок, автоматического формирования пристатейных списков использованной литературы. При этом реализована возможность выбирать стили цитирования и представления списков литературы (например, по ГОСТу). Наличие облачного сервиса Mendeley расширяет возможности пользователей, в частности, позволяет им синхронизировать свои списки литературы с удалённым хранилищем, что

даёт возможность оперативного доступа к ней с любого устройства, подключенного к сети Интернет, а также содействует коллективной работе над общей библиографией для других пользователей сервиса. Правда, эти возможности в бесплатном базовом варианте очень ограничены, и за их расширение необходимо платить абонентскую плату, что накладывает известные ограничения на использование этого инструмента в академической среде. Указанные возможности рассчитаны, скорее всего, не на отдельных пользователей, а на использование сервиса Mendeley организациями для своих сотрудников.

В современной деятельности учёного большую роль играют формальные показатели его деятельности, которые отражаются в его публикационной активности (количестве публикуемых им научных работ). Кроме этого, важным показателем является факт цитирования его работ, который показывает о значимости его деятельности, востребованности результатов его исследований среди коллег. Из этого складывается индекс его цитирования, который в России рассчитывается в рамках проекта «Российский индекс научного цитирования» (РИНЦ). РИНЦ – это национальная библиографическая база данных научного цитирования, аккумулирующая огромное число публикаций российских авторов, а также информацию о цитировании этих публикаций из журналов, сборников статей, материалов конференций и монографий. Она предназначена не только для оперативного обеспечения научных исследований актуальной справочно-библиографической информацией, но является также мощным аналитическим инструментом, позволяющим осуществлять оценку результативности и эффективности деятельности научно-исследовательских организаций, ученых, уровень научных журналов и т.д.

Кроме этого, индексы научного цитирования рассчитываются и в зарубежных реферативных базах Scopus и Web of Science. Однако упомянутые ресурсы не могут представить объективную картину по публикационной активности всех учёных в силу своей неполноты. Тем не менее, они показательны для оценки самых авторитетных учёных и исследователей во всём мире.

## Контрольные вопросы к Главе 1

Релевантность в контексте информационного поиска – это:

1) Адекватность полученной пользователем информации общепринятым нормам и взглядам

2) Соответствие полученной пользователем информации его ожиданиям

3) Соответствие полученной пользователем информации уровню его осведомлённости

Укажите причины, по которым информация о ресурсах сети Интернет может отсутствовать в индексной базе ИПС:

1) владелец сайта не предоставил информацию

2) сайт ещё не проиндексирован поисковыми роботами

3) ресурс сети Интернет является представителем пространства «глубинного веба»

4) информация удалена по запросу владельца сайта

5) ресурс является запрещённым по решению ЮНЕСКО

Язык поисковых запросов представляет собой:

1) специальный язык для человеко-машинного общения

2) расширение языка программирования Java

3) набор команд и метасимволов, которые используются для модификации поисковых запросов в строке базового поиска ИПС

4) специальный язык для осуществления поиска в неиндексируемом пространстве сети Интернет

Что из нижеперечисленного не реализовано в электронных каталогах библиотек:

1) Базовый, расширенный и специализированный поиск

2) Доступ к рецензиям на научные публикации

3) Возможность заказа

4)Поиск по библиографическим спискам научных публикаций

5) Поиск по содержаниям сборников статей и коллективных монографий

6) Доступ к полным текстам всех публикаций

Выберите из списка основные типы периодических научных изданий:

1) печатное рецензируемое издание

2) печатное нерецензируемое издание

3) печатное рецензируемое издание с электронной версией

4) электронное рецензируемое издание
5) электронное издание с возможностью свободного размещения статей

Основными признаками периодических электронных сетевых научных изданий являются:
1) наличие сайта в сети Интернет
2) размещение информации о журнале в Википедии
3) обязательная научная экспертиза подаваемых в издание рукописей
4) наличие ISSN для электронной версии
5) обязательное предоставление номеров и выпусков на носителях информации (CD, DVD)
6) индексация в наукометрических базах
7) публикация на сайте всех номеров/выпусков издания с полными текстами статей

Доступ к полным текстам публикаций, размещённых в Научной электронной библиотеке Elibrary могут получить:
1) Только зарегистрированные пользователи
2) Все желающие
3) Оплатившие подписку
4) Сотрудники научных организаций

Укажите только те типы информационных ресурсов сети Интернет, на которых можно найти полные тексты научных публикаций:
1) Электронные каталоги библиотек
2) Полнотекстовые базы научной информации
3) Информационно-поисковые системы общего назначения
4) Сайты органов государственной власти
5) Информационные ресурсы научных и учебных организаций
6) Личные страницы учёных
7) Сайт Роскомнадзора
8) Сайты научных сообществ
9) Сайты новостных агентств
10) Сайты научных конференций
11) Сайты издательств

# Глава 2. Знание как научая категория

## 2.1. Общие подходы к определению знаний

Информация – это мера изменения во времени и в пространстве структурного разнообразия систем (согласно классическому определению Р. Эшби), а также – это устраненная неопределенность, отражение окружающего мира посредством сигналов и знаков; это сведения об объектах и явлениях окружающей среды, их параметрах, свойствах и состояниях, которые уменьшают имеющуюся степень неопределенности, неполноты знаний. Противоположным информации в теории информации является понятие энтропии, которое определяется как мера неопределенности, неотъемлемой от понятия вероятности. Чем больше информации, тем меньше энтропия системы, и наоборот. Р. Эшби в середине 50-х годов прошлого века осуществил переход от толкования информации как «снятой» неопределенности к «снятой» неразличимости, считая, что информация есть там, где имеется разнообразие (характеристика элементов множества, заключающаяся в их несовпадении), неоднородность. Понятие энтропии применялось первоначально только для систем, стремящихся к термодинамическому равновесию, т.е. к максимальному беспорядку в движении ее составляющих, а значит к увеличению энтропии. Понятие информации обратило внимание на те системы, которые стремятся к дальнейшему уменьшению энтропии. Таким образом, так как энтропия является мерой неупорядоченности, то информация может быть определена как мера упорядоченности систем[19].

Необходимость введения термина «информация» возникла на тех этапах развития материального мира, когда вслед за возникновением общества и общественных отношений появляется потребность изучать целенаправленные действия, процедуры принятия решений и их зависимость от внешних условий. Во всех остальных случаях можно обойтись без термина «информация» и протекающие процессы описывать с помощью законов естественно-научных дисциплин. Информация по отношению к окружающей среде (или к использующей её среде) бывает трех типов: входная, выходная и внутренняя. Другие признаки, выбранные для классификации, позволяют сформулировать дополнительные типы информации (рис. 2.1).

Наряду с понятием информация, теория информации и информатика включают такие понятия, как сведения, сообщение, данные и знания, каждое из которых применяется в определенных случаях и для определенных целей, но иногда используются и как синонимичные понятию информация. Например, информация может быть рассмотрена как последовательность сведений, знаний,

[19] Система – это совокупность объектов, находящихся в отношениях и связях между собой и образующих определенную целостность, единство, т.е. существующая (функционирующая) как единое целое, приобретающая новые свойства, которые отсутствуют у этих объектов в отдельности.

сообщений, выражаемых с помощью некоторого алфавита символов, жестов, звуков, сигналов. Соответственно и формы представления информации могут быть различны: символьная (основана на использовании различных символов), текстовая (текст — это символы, расположенные в определенном порядке), графическая (различные виды изображений), звуковая и т.п.

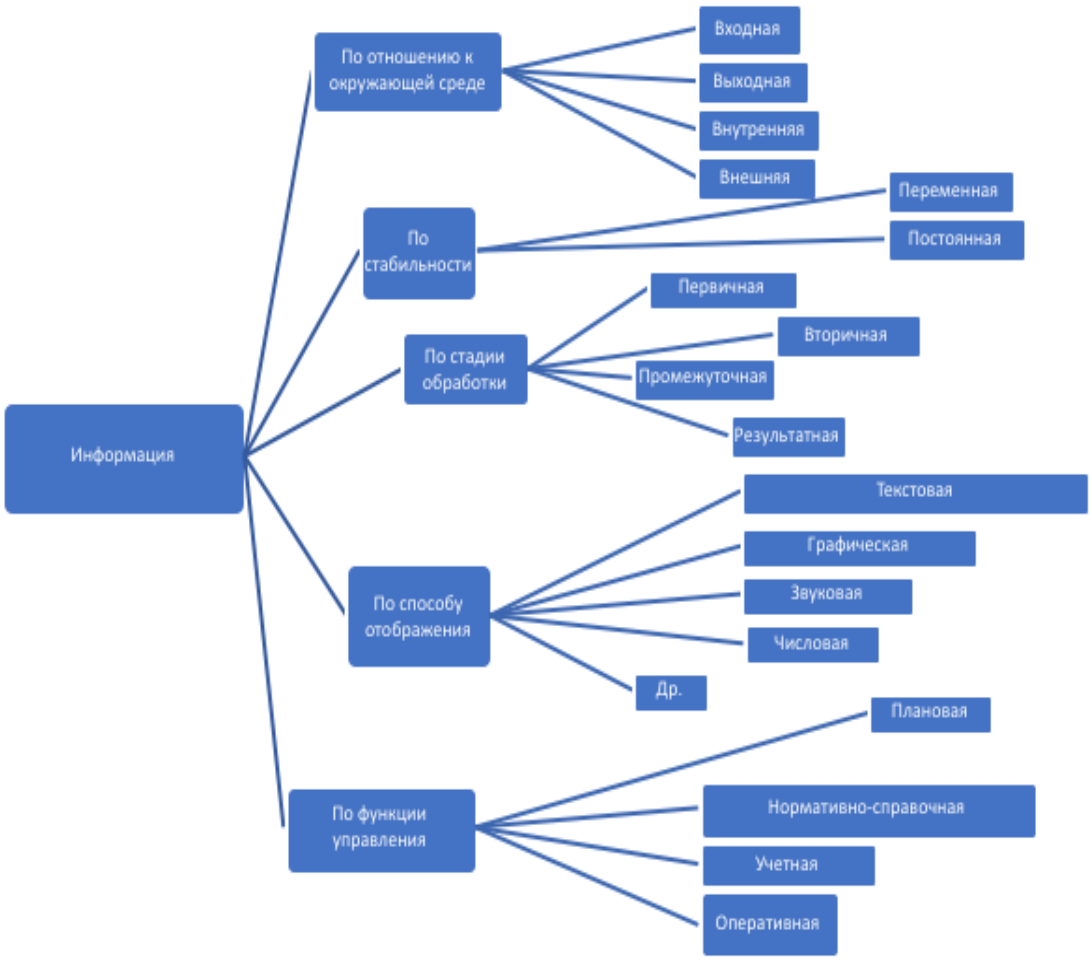



Рис. 2.1. Виды информации

Существует разные точки зрения на концепцию знания[20]. Знание может быть определено как форма существования и систематизации результатов познавательной деятельности человека, что означает наличие причинно-следственных и других видов связей между сущностями, относящимися к той или иной предметной области. Научному знанию присущи логическая обоснованность, доказательность, воспроизводимость результатов, проверяемость, стремление к устранению ошибок и преодолению противоречий. Эмпирические знания – это знания о видимых взаимосвязях между отдельными событиями и фактами в предметной области, получаемые в результате

[20] Раздел подготовлен с использованием материалов: Chu Keong Lee, Schubert Foo and Dion Goh On the Concept and Types of Knowledge Journal of Information & Knowledge Management. 2006. Vol. 5. No. 2. P. 151–163.

применения эмпирических методов познания: наблюдение, измерение, эксперимент. Знания констатируют качественные и количественные характеристики объектов и явлений. Знание в узком смысле – признак определённого объёма информации, определяющий её статус и отделяющий от всей прочей информации по критерию способности к решению поставленной задачи.

В информатике, теории искусственного интеллекта, базах знаний и экспертных системах знание это – совокупность данных, фактов, сведений и законов (у индивидуума, у общества или у системы ИИ) о мире, включающих в себя информационные свойства объектов, закономерности процессов и явлений, а также правила использования всей этой информации для принятия решений. Правила использования включают систему причинно-следственных связей. Главное отличие знаний от данных состоит в их активности, то есть появление в базе новых фактов или установление новых связей может стать источником изменений в принятии решений.

Знание в информатике и дисциплинах, связанных с информацией и информационными технологиями, человеко-машинным взаимодействием в рамках информационных систем, также может быть определено различным образом. Один из наиболее известных и востребованных взглядов предполагает наличие явного и неявного знания, теория вопроса является развитием теории информации (рис. 2.2).

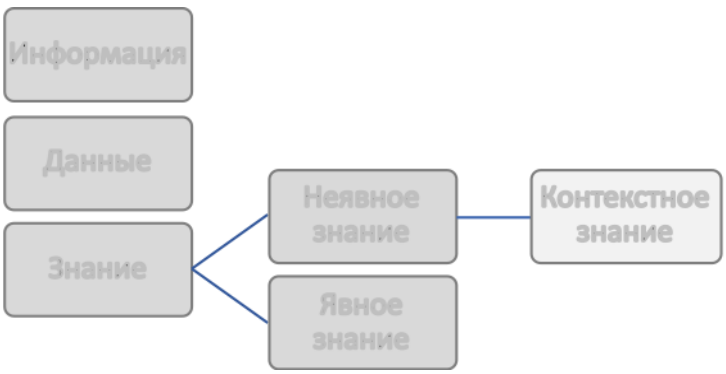



Рис. 2.2. От информации к знаниям

Описанный выше подход подразумевает вездесущность и независимость информации, ее наличие независимо от того, умеем или нет мы ее выявлять, различать и измерять. Знание в этой парадигме вторично, сосуществует наравне с данными, сообщениями и сведениями, и вместе они являются проекцией некой обозримой части информации на запоминающее устройство, информационный носитель (сознание, лист бумаги, память компьютера). Знания отражают наши представления об устройстве мира, тех или иных систем. Модель континуума в этом случае представляет собою трио, где информация является первой точкой континуума:

Информация–Данные–Знания
(Information–Data–Knowledge Continuum)

В основе альтернативного подхода лежит представления о континууме, точки которого располагаются в следующем порядке:

Данные–Информация–Знание
(Data–Knowledge–Information Continuum).

Получили известность три точки зрения на знание:

– знание как потенциал (knowledge as potential),

– знание как конечная точка в континууме (the endpoint on a Data–Information–Knowledge Continuum),

– знание как объект или как процесс.

**Знания как потенциал**

Знание является побочным продуктом исследования и может рассматриваться либо как коллекция информации, либо как деятельность, либо как потенциал. Однако представление знаний как совокупности информации «лишает понятие своей жизни» (Черчман, 1971) и не может изобразить его жизнеспособность и его способность производить огромные изменения в мире. Более того, знания принадлежат пользователю и используются пользователем, т.е. как бы находятся в пользователе, а не в собранной информации. С другой стороны, представлять знание посредством действий хоть и прагматично, но проблематично. Таким образом, данная точка зрения предполагает, что знание лучше всего рассматривать как очень мощный потенциал, который позволяет человеку корректировать свое поведение в соответствии с меняющимися обстоятельствами и дает ему возможность превращать данные и информацию в эффективные действия (Applehans et al. др., 1999).

**Знания как конечная точка в континууме. Данные–Информация–Знания (Data–Information–Knowledge Continuum)**

Данные могут быть определены как упорядоченные последовательности событий или статистика (Bell, 1999). Наличие значения (смысла, контекста) отличает информацию от данных и включает осведомленность о новостях, событиях и происшествиях. Значение смысла новостей, событий и происшествий, подтвержденное контекстом или теорией, составляет знание. Суждение (judgment), которое «возникает из желания переупорядочить, переставить и изменить то, что уже известно», участвует в процессе преобразования (рис. 2.3).

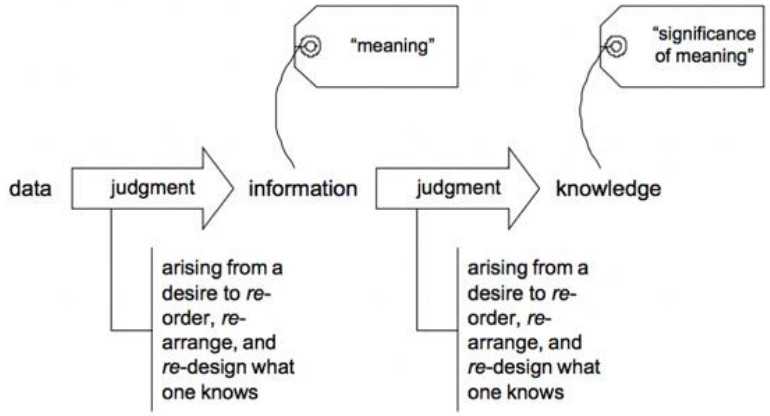



Рис. 2.3. Знания как конечная точка в континууме

Придерживаясь данной концепции, данные можно описать как набор дискретных, объективных фактов о событиях, обычно хранимых в памяти вычислительной системы (Davenport and Prusak, 1998). Информация извлекается из данных посредством контекстуализации, категоризации, вычислений, исправлений или уплотнения и может принимать форму текстового сообщения, визуальной или звуковой коммуникации. Знания, в свою очередь, получают из информации путем сравнения, обсуждения и установления связей с предыдущими знаниями (рис. 2.4). Знания определяются как «подвижное сочетание ранее сформированного опыта, ценностей, контекстной информации и экспертных оценок, которое обеспечивает основу для оценки и включения нового опыта и информации». Хотя знания исходят от людей и применяются ими, они встроены в документы, репозитории, организационные процедуры, процессы, практики и нормы.

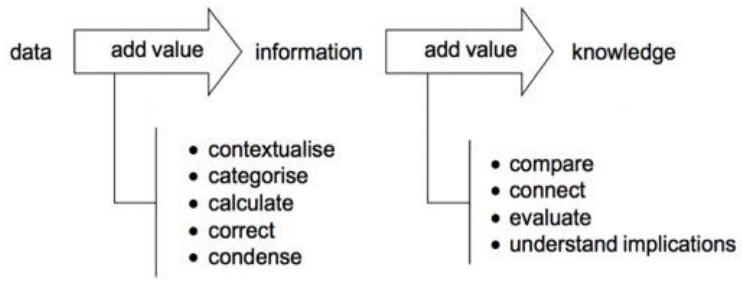



Рис. 2.4. Данные–Информация–Знания континуум (Davenport and Prusak, 1998)

Континуум может быть расширен за пределы знания включением в него концептов «мудрость» и «проницательность». Мысленное путешествие, посредством которого строится понимание, можно описать как путешествие, которое начинается с представления и сбора данных, проходит через ключевые пункты (stations) информации и знания, и далее движется к мудрости и пониманию (Haywood, 1995). Организационная мудрость, суждение, выбор и использование конкретных знаний для конкретного контекста – важная, хотя и

отсутствующая, конструкция в теории организации, основанной на знаниях (Bierly et al., 2000). Точка зрения на знание как на конечную точку в континууме Данные–Информация–Знания имеет ряд дополнительных интерпретаций[21]

В рассмотренных моделях Данные–Информация–Знания континуума данные являются наиболее объемными, информация занимает меньше места, а знания наименее объёмны.

**Знание как объект или как процесс**

Знание можно рассматривать как объект (object, stock), существующий независимо от знающего. В этом случае знания приравниваются к патентам, отчетам, математическим и химическим уравнениям или чертежам, которые можно собирать, распространять, измерять и управлять ими.

Знание также можно рассматривать как процесс (process, flow) познания и осмысления. Здесь знание неотделимо от знающего и действительно бессмысленно в отсутствие знающего. Только люди способны быть осведомленными, а не книги или базы данных. Из-за этого акт познания можно только поддерживать, поощрять, направлять и мотивировать. Акцент на накоплении знаний в ущерб потоку знаний – это распространенная ошибка в управлении знаниями (Cohen, 1998; Fahey and Prusak, 1998; Davenport and Prusak, 1998).

Таким образом, различное восприятие знания порождает различное восприятие обмена знаниями. Если знание рассматривается как потенциал особого рода, что-то, что рассматривается как источник силы и имеет способность придавать исключительность и уникальность человеку, то обмен знаниями будет восприниматься как разделение плодов знания, делая распространителя знаний менее мощным, менее эксклюзивным и менее уникальным. Ожидается, что это будет ярко выражено в культурах, которые высоко ценят индивидуализм, где ценятся достижения людей. Обмен знаниями в этом случае будет исключен. Если рассматривать знания как конечную точку в континууме, достижимую только после напряженных усилий по добавлению ценности формам меньшей ценности, можно ожидать таких же последствий. Обмен знаниями будет избегаться, поскольку обмен знаниями будет рассматриваться как предоставление чего-то, над чем человек трудился.

[21] Существует несколько вариантов модели «точка на континууме», подразумевающие вставки или изменения порядка иерархии. Например, это вставка элемента «понимание» (Ackoff, 1989), и вставка элемента «интеллект» (Por, 1997), между элементами «знание» и «мудрость». Изменения порядка иерархии обосновываются тем, что обычная модель данные – информация - знания может вводить в заблуждение во многих ситуациях. Альтернативная модель – это так называемая «обратная иерархия знаний», в которой порядок трех элементов обратный (Tuomi, 1999). При таком подходе знания должны быть на первом месте, поскольку они сначала используются для фиксации семантики (структуры значений). Затем эта структура используется для представления информации, и только после этого могут появиться данные. Здесь знания играют решающую роль в построении четко определенной семантики, без которой данные не могут быть созданы, и, следовательно, они должны существовать до того, как данные появятся.

Восприятие знаний как объектов означает, что обмен знаниями будет приравниваться к совместному использованию объектов. Акцент будет сделан на кодификации знаний, чтобы ими можно было делиться, сборе знаний, строительстве хранилищ знаний и изучении содержимого хранилищ. Здесь обмен знаниями будет приравниваться к внесению вкладов в репозитории документов и отправке электронных писем с вложенными файлами, поскольку объекты, воплощающие знания, становятся доступными для других и, следовательно, «общими».

Восприятие знания как процесса подчеркивает процесс познания, осмысления и конструирования реальности. Акцент будет сделан на среду, в которой происходит обмен знаниями, и это будет включать организационную культуру, опыт обмена знаниями, психологическую безопасность, необходимую для откровенного обмена, а также наличие времени и возможностей для общения.

## 2.2. Общие таксономии знаний

Чтобы понять роль знания в обществе и в организации, ученые в области практической эпистемологии провели различия между разными видами знания и предложили варианты таксономии знаний (Taxonomies of Knowledge).

### 2.2.1. Положительные и отрицательные знания

Знание о неудачах так же важно, как и знание об успехах. Точно так же знание о подходах, которые не работают, так же важно, как и знание о подходах, которые действительно работают. Открытие, т.е. положительное знание (positive knowledge) желательно, но следует также признать, что опыт зайти в тупик, отрицательное знание (negative knowledge), тоже может быть ценным, поскольку он может помочь в будущем направить распределение ресурсов в более перспективные направления. По этой причине отслеживание неудачных идей следует также считать полезным, поскольку это позволяет прийти к пониманию того, что они могут стать успешными в будущем (Davenport and Prusak, 1998). В этой связи стоит указать на важность и прагматическую значимость поиска и экспликации контекстов и контекстного знания в массивах научных публикаций, что заведомо даст положительный эффект при правильной постановке задач исследования.

### 2.2.2. Явное и неявное знание[22]

Согласно Полани (Polanyi, 1966), Кано-Кикоски и Кикоски (Kano-Kikoski and Kikoski, 2004) и Нонака (Nonaka, 1994), существует два типа знания:

[22] Раздел подготовлен с использованием материалов: Chergui, W., et al. An approach to the acquisition of tacit knowledge based on an ontological model. Journal of King Saud University – Computer and Information Sciences. 2020. Vol. 32. Iss. 7. P. 818-828. DOI: 10.1016/j.jksuci.2018.09.012; Kanygin G., Kononova O. The Expression of Tacit Knowledge by Actors of Smart Technologies // CEUR Workshop Proceedings. 2020. Vol. 2784 P. 98-112. http://ceur-ws.org/Vol-2784/rpaper09.pdf

– явное или эксплицитное (explicit) знание;

– неявное или неэксплицитное, латентное (tacit, implicit, incarnate), оно же скрытое знание.

Явное знание (explicit knowledge) кодифицировано (Arling and Chun, 2011), то есть организовано и передано в соответствии с формализмом, символом или подходящим естественным языком (organized and communicated according to a formalism, a symbol or appropriate natural language). Поэтому оно легко передается и может быть записано в структурированной форме при помощи артефактов, таких как процедуры, отчеты, стратегии, руководства и т. д.

Неявное знание (tacit knowledge), которым мы обладаем, выходит за рамки того, что мы можем выразить (Arling and Chun, 2011). Оно ориентировано на действие, опыт и обязательства участников в конкретном контексте (oriented toward an action, an experience and a commitment of actors in a specific context). Неявное знание, которое включает в себя субъективные озарения, интуицию и догадки, носит очень личный характер. Неявное знание глубоко укоренено в действиях и опыте человека, а также в идеалах, ценностях или эмоциях (когнитивный аспект), которые тот или иной человек принимает. Неявное знание (tacit knowledge) является скрытым (implicit) и обращается к опыту и пониманию «знаю как» человека (технический аспект), который им владеет; оно коренится в действиях, процедурах, распорядках, обязательствах, идеях, ценностях и эмоциях этого человека (Nonaka et al., 2000). Неявное знание сложно формализовать, сформулировать и передать в форме, которая может быть использована другим человеком, без использования специальных методов. Поэтому им трудно делиться с другими.

Основное различие между неявным и явным знанием состоит в том, что неявное знание приобретается через опыт, тогда как явное знание представлено материальным образом, например, на бумаге или в компьютерной программе, книге или уведомлении, которое сохраняется и поэтому легко передается, и носит в силу этого консервативный характер. Вместе эти два вида знаний образуют интерактивный набор, в котором два типа знаний взаимозависимы. Интерактивный набор повышает качество знаний индивидов и позволяет предлагать интерпретации, которые зависят от «багажа» интерпретатора (Alavi and Leidner, 2001).

Неявные знания – это результат многих лет обучения и приобретения опыта. Неявные знания не проявляются явным образом в научных результатах, но являются существенным условием для появления этих результатов.

Таким образом, все знания либо сами неявны, либо основаны на неявных знаниях, т.е. имеют неявное происхождение (Nonaka, Takeuchi,1995). Явное знание зависит от неявного знания и основывается на нем. Неявное знание может быть передано только в том случае, если мы можем преобразовать его в слова, числа или изображения, понятные другим. Но на сегодняшний день не удалось

создать методы выражения неявных знаний, которые, с одной стороны, дали бы возможность каждому человеку осмысленно изъяснять или выражать свой практический опыт, а с другой, будучи реализованными в составе ИКТ, позволяли бы социальным акторам коллективно строить и взаимно согласовывать результаты такого выражения (Dragicevic et al., 2020).

Нонака и Такеши (Nonaka and Takeuchi, 1995) формализовали модель создания знаний, выделив четыре способа создания и передачи знаний. Модель предполагает различные взаимодействия между неявным и явным знанием (implicit and explicit knowledge), как показано на рис. 2.5. Эта модель суммирует основной процесс преобразования знаний, который происходит через социальные и когнитивные процессы.

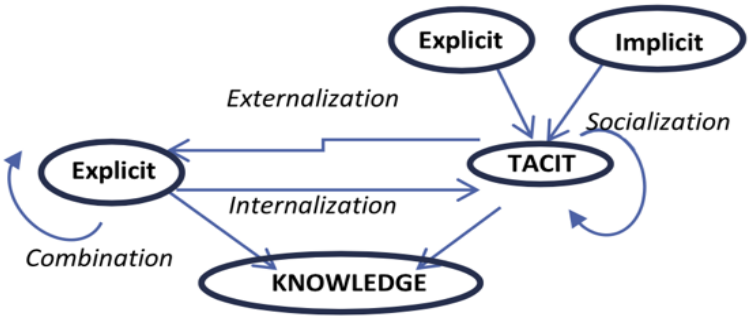



Рис. 2.5. Модель создания знаний (The Knowledge Creation Model by Nonaka and Takeuchi, 1995)

Для адаптации идей дуальности к задачам управления знаниями в организации, ориентированной на производство нового знания, Нонака и Такеши (Nonaka and Takeuchi, 1995) предложили спиральную (spiral) модель, которая объясняет функционирование знаний в организации через социальные взаимодействия, основанные на четырех процессах преобразования (knowledge conversion) двух видов знаний – явного и неявного. Модель Нонака и Такеши известна как SECI-модель – название образовано сокращением до первых букв режимов/процессов Socialization, Externalization, Combination, Internalization.

Таким образом, существует четыре режима (фазы) преобразования знаний.

Первый режим – социализация (socialization): переход от неявного знания (tacit knowledge) к неявному знанию (tacit knowledge), который возникает в результате взаимодействия между людьми в группе. Взаимный обмен знаниям, обучение, осуществляется путем наблюдения, имитации и обмена опытом. Социализация – это преобразование неявного знания (навыков, умений, обычаев, правил поведения и т.п.), существующего в обществе, в неявные же представления члена этого общества.

Второй режим – комбинирование (combination): переход от явного знания (explicit knowledge) к явному знанию (explicit knowledge), который позволяет создавать явное знание путем дедукции или индукции из реструктурированного

набора элементов явного знания, которые были получены ранее. Комбинирование – это известные операции с различными видами информации (преобразование текстов и изображений, формирование баз данных, программирование).

Третий режим – интернализация (internalization): переход от явного знания (explicit knowledge) к неявному знанию (tacit knowledge), который олицетворяет преобразование явного знания в неявное в процессе обучения типа «обучение на практике». Интернализация – это преобразование индивидом «внешней» информации в виде текста, таблицы, рисунка и т.п. в свой «внутренний код», доступный только ему самому.

Четвертый режим – экстернализация (externalization): переход от неявного знания (tacit knowledge) к явному знанию (explicit knowledge), который предлагает объяснение практик и убеждений путем формулирования неявного знания в явных концепциях, таких как аналогии, концепции, гипотезы или модели. Экстернализацию можно назвать выражением неявного знания индивидов в обозримой для всех, явной, форме (слов, рассказов, диаграмм и других средств).

Знания можно рассматривать как нематериальные элементы, несущие богатство; как интеллектуальный капитал, который отличает компанию от ее конкурентов (Ermine, 2000). С уходом ключевого сотрудника неявные знания для организации буквально теряются. Неявный человеческий капитал экспертов организации может быть выражен с помощью эксплицитных интервью на фазе экстернализация, на которой неявное знание экстернализируется как явное знание. Затем, с целью сохранения, знание представляется в виде онтологической модели. Впоследствии неявные и явные знания в онтологии обеспечивают справочную базу «знаю-как» (a referential knowledge base of know-how), которую можно использовать для измерения потенциала человека применительно к выполнению конкретной задачи. На второй фазе, фазе комбинирования, знания могут быть получены посредством умозаключений. Поэтому так важно создавать благоприятные условия для преобразования неявного человеческого капитала в явный структурный капитал, который должен быть получен от каждого компетентного сотрудника организации. Поэтому SECI-модель нашла многочисленных сторонников.

### 2.2.3. Концепции неявного знания

Участники современных дискуссий отмечают три взаимосвязанные ключевые особенности неявного знания:

– такие знания неотделимы от своего носителя;

– индивидуализированное неявное знание фактически определяется контекстом;

– неявное знание трудно формализовать.

Существует несколько концепций, касающихся неявного знания. Первым исследователем, обратившимся к неявному знанию, был Поланьи (M. Polanyi, 1958, 1966), который сказал, что мы можем знать больше, чем можем сказать. По словам Поланьи, все научные знания фактически основаны на личном опыте, который ближе к практическим, чем к теоретическим знаниям. Он привел пример езды на велосипеде, предполагающего неявное знание (сегодня это называется воплощенным знанием). Эти знания приобретаются путем обращения к нашим биологическим способностям. Каждый человек знает, что, выражая что-либо в виде слов, он оставляет многое недосказанным. Когда мы общаемся друг с другом, то видим только верхнюю часть айсберга наших мыслей, а их подводная основа в ее многообразии и хитросплетениях остается невыраженной. Такой дуальный подход к рассмотрению знания – неявное через явное – с середины прошлого века приобрел многочисленных сторонников, которые предприняли значительные усилия, чтобы превратить его в законченную науку управления знаниями.

Согласно концепциям дуального знания, основой всей информации, функционирующей в контуре современной ИКТ, являются неявные (скрытые, неэксплицитные, tacit, implicit) знания человека. Такое знание трактуется как «знание, которое укоренено (embodied) в человеке, т.е. оно не существует вне носителя знания; социально сконструировано (т.е. оно создается совместно индивидом и социально обусловленными смыслами); привязано к практике (т.е. неотделимо от взаимодействий людей); встроено в культуру и традиции общества (т.е. сформировано социально-культурной средой, в которой происходят взаимодействия носителей знания)» (например, Dragicevic et al., 2020; Brown & Duguid, 1991; Hislop, 2002).

Последующие исследования расширили концепцию неявного знания. Линде (2001) подразделяет неявные знания на более конкретные формы социальных, физических и других знаний и утверждает, что нарративы ликвидируют разрыв между неявным и явным знанием и особенно полезны для передачи социальной формы неявного знания. На рис. 2.6 представлена таксономия неявного знания как расширение таксономии Линде.

*Социальное знание* означает умение себя вести, контролировать события и взаимодействия. *Физическое знание* – сохранение равновесия. Эмоциональное знание как «разновидность неявного знания» – умение определять эмоции. Неявные знания – это больше, чем просто моторика, как при езде на велосипеде. Неявное знание в этом случае также включает эмоциональное измерение возможности жить такой жизнью, которая включает в себя возбуждение от езды на велосипеде (Spender, 2003).

Термин «*безмолвное знание*» (silent knowledge) обозначает знание, которое позволяет человеку взаимодействовать с другими в различных ситуациях – как рутинных, так и необычных. Безмолвные знания имеют решающее значение,

поскольку они позволяют человеку регулировать свое поведение, чтобы быть безопасным и полезным участником социальной жизни. Эта форма знания называется безмолвной (silent), поскольку она обычно не произносится. В результате его чаще «ловят» учащиеся, нежели «преподают» эксперты или читают по книге. Примерами такого знания могут служить: (1) знание приемлемого и недопустимого поведения и одежды при посещении похорон; (2) знание того, как сморкаться публично; (3) знание того, как просить и давать указания, и (4) знание того, как вести себя на первом свидании. Хотя некоторые из действий могут показаться простыми, они требуют от человека значительного количества знаний (Schwalbe, 2005). Безмолвное знание также подкрепляет понимание основных правил общественной жизни. Некоторые из этих правил являются нормативными, например, «всегда уважать чувства других», в то время как другие являются процедурными, например, «участвуя в обсуждении, не доминируйте в разговоре, говорите по очереди».

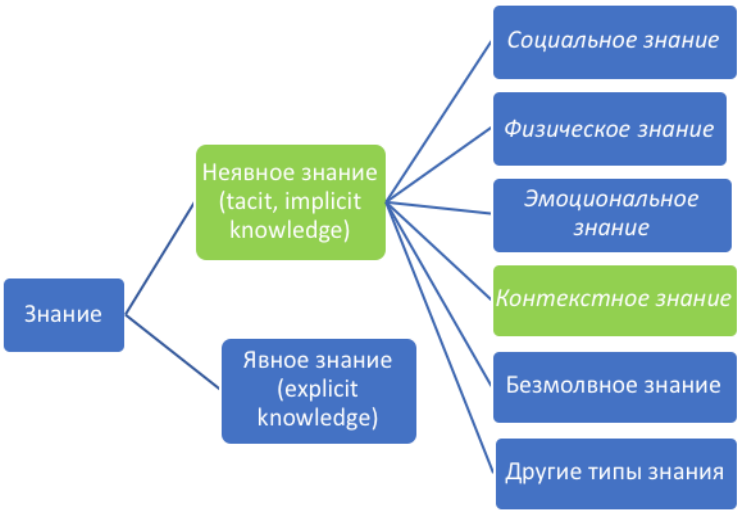



Рис. 2.6. Таксономия неявного знания как расширение таксономии Линде

*Контекстное знание* – умение понимать значения символов, звуков и трактовать рисунки в зависимости от их окружения, громкости, тона, интонации и т.п. Интерес, проявляющийся через указание на то, интересен ли данный документ или нет (согласно Stenmark, 1999), является примером неявного знания (tacit knowledge). Процессы распознавания контекстов моделируются с использованием различных методов и технологий, выбор которых зависит от класса задач и модальности, о которой пойдет речь в следующем разделе.

Понятию «неявное знание» в английском языке соответствует несколько терминов, основные из которых – это «tacit knowledge» и «implicit knowledge». Ряд исследователей проводят различия между ними. Так, например, «implicit knowledge» понимается как знание, которое, хотя и могло быть сформулировано (формализовано), но пока еще не сформулировано. Аналитик по задачам, инженер по знаниям или другой специалист, обладающий навыками определения

типа знаний, которые могут быть сформулированы как implicit knowledge, часто могут вытащить неявные знания из компетентного исполнителя. В противовес ему «tacit knowledge» часто не может быть сформулировано. Блок-схема, отражающая различия между двумя видами неявного знания – tacit knowledge и implicit knowledge – представлена на рис. 2.7.

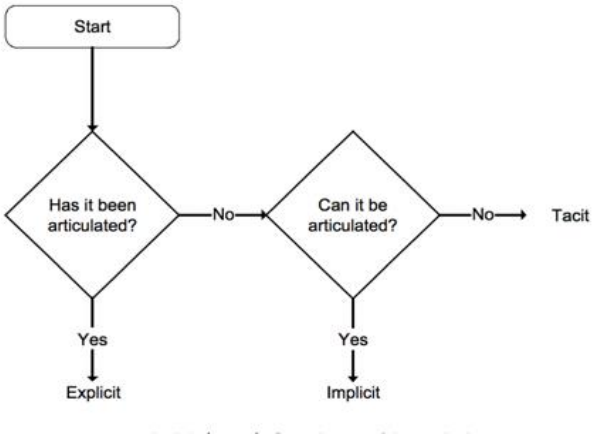



Рис. 2.7. Блок-схема, отражающая концепцию неявного знания в (Nickols, 2000)

Для того чтобы объяснить тот факт, что, хотя явная часть нашей базы знаний обширна, она представляет собой лишь малую часть, используется аналогия с айсбергом, где явное знание занимает верхушку айсберга, большая же часть знания, неявного знания, существует в негласной форме ниже уровня воды (рис. 2.8).

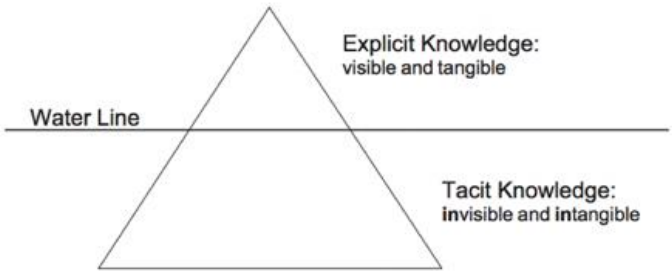



Рис. 2.8. Аналогия с айсбергом (Abdulai, 2004)

Еще одна концепция выделяет три «различных варианта» неявного знания. Во-первых, это относится к знаниям, которые были усвоены в течение длительного периода времени до такой степени, что теперь они принимаются как должное, потому что все их понимают. Во-вторых, это относится к знанию, которое не сформулировано, потому что никто не понимает его полностью. В-третьих, это относится к знаниям, которые не сформулированы не потому, что их никто не понимает, а потому, что их систематизация потребует больших затрат времени и усилий (Boisot,1998). Именно третий случай в наибольшей степени занимает умы и подвержен автоматизации, а также важен при проведении научных исследований независимо от выбранных научного направления и предметной области.

2.2.4. Пункты Бойсота в информационном пространстве (Boisot’s stations in the Information Space)

Альтернативные типы знания определяются как пункты (stations) или позиции (positions) в трехмерном эпистемологическом пространстве, которое также называют информационным пространством (information space) или И-пространством (I-space), рис. 2.9. Три оси: абстракция (степень, в которой знаниям можно придать структуру), кодификация (степень, в которой знаниям можно придать форму) и диффузия (степень, в которой знания становятся доступными для тех, кто хочет их использовать) – обозначают три измерения знания и определяют И-пространство. Концепция И-пространства отражает взаимосвязи между этими измерениями и последствия передвижения знаний по ним.

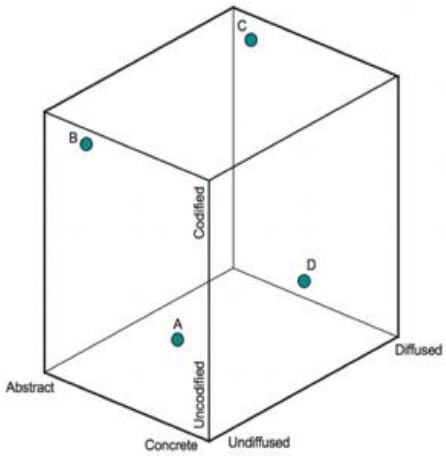



Рис. 2.9 Информационное пространство (Boisot, 1998)

2.2.5. Таксономия знаний Лундвалла и Джонсона

Лундвалл и Джонсон (Lundvall and Johnson, 1994) предлагают таксономию знаний, включающую know-what, know-why, know-how, and know-who знания. Знание know-what относится к фактам и характеризуется своей способностью разбиваться на части. Знание know-why относится к знаниям о причинно-следственных связях и устраняет необходимость в пробах, устраняет возникновение ошибок. Знание know-why особенно важно для технологического развития в наукоемких областях. Знание know-how означает навыки или способность что-то делать. Знание know-who включает в себя знания о том, кто что знает, и кто знает, что именно делать. Знание know-who также включает в себя социальные навыки, которые позволяют сотрудничать и общаться с коллегами и сотрудниками.

### 2.2.6. Типы полезного знания Мокира (Mokyr's types of useful knowledge)

Полезные знания определяются как знания, накопленные людьми в процессе наблюдения за природными явлениями и затем описанные и изученные в попытке установить в них закономерности (regularities) и модели (patterns). Полезные знания бывают двух типов (Mokyr, 2003). Во-первых, знание «что» ("what") или знание высказываний (prepositional knowledge; Ω-знание) связано с убеждениями о природных явлениях и их закономерностях. Во-вторых, знание «как» ("how") или обучающее, предписывающее знание (instructional, prescriptive knowledge; λ-знание), которое связано с методами.

### 2.2.7. Типология знания Агуайо (Aguayo's categories of knowledge)

Существует две категории знания:

– субстанциальные знания (substantive knowledge) относятся к знанию предмета, относящегося к конкретной области (например, знания о выращивании лосося для получения максимальной продуктивности и качества);

– предпринимательские знания (entrepreneurial knowledge) – знание того, как монетизировать или коммерциализировать существенные знания (например, знания о том, как продвигать и продавать с максимальной прибылью).

Организация, основанная на знаниях – это организация, в которой на каждого эксперта по предпринимательским знаниям приходится гораздо больше экспертов по основным знаниям. Типология Агуайо признает, что человеку, у которого есть хорошие идеи, часто требуется человек с хорошей деловой хваткой, чтобы получить от идей финансовое вознаграждение (Aguayo, 2004).

### 2.2.8. Уникальные и неуникальные знания Прайса (Price's unique knowledge and non-unique knowledge)

Следует проводить различие между знаниями, полученными в результате научной деятельности и культурной деятельности (Price, 1963). Научная деятельность дает знания, которые не являются уникальными (unique knowledge), если оценивать их относительно создавшего их ученого. По сути, одни и те же знания могли ли быть созданы разными людьми. С другой стороны, культурная деятельность – творческий вклад, который носит исключительно личный характер. Поэтому, если бы того или иного композитора, писателя, художника никогда не существовало, их произведения были бы заменены совсем другими произведениями. Таксономия знания Прайса унаследована от двух форм знания: интуитивного знания (уникальное знание Прайса) и логического знания (неуникальное знание Прайса). Интуитивные знания (intuitive knowledge) – это знания, полученные с помощью воображения, логические знания (logical knowledge) – это знания, полученные с помощью интеллекта (Croce, 1900).

Вагнер и Стернберг (Wagner and Sternberg, 1985) считают неявное знание аспектом практического интеллекта. Это знания, отражающие нашу

практическую способность учиться на собственном опыте и применять полученные знания для достижения личных целей. Поскольку неявное знание является аспектом практического интеллекта, концепция неявного знания предлагает уникальный взгляд на важный фактор, лежащий в основе успешного выполнения реальных задач. Неявное знание включает, с одной стороны, когнитивный компонент, а именно, ментальные модели (Johnson-Laird, 1993), которые люди формируют относительно мира (схемы, парадигмы, верования и точки зрения, которые обеспечивают точки зрения, которые помогают людям собирать и определять их видение мира), и, с другой стороны, технические компоненты, а именно ноу-хау и навыки, которые применяются в определенных контекстах (Dieng et al., 2000).

И сегодня, как и прежде, неявное знание считается одним из главных источников инноваций, которые ускоряют экономическое и социальное развитие (Nonaka, Takeuchi, 1995; Leonard, Sensiper, 1998; Virtanen, 2014). В последнее время задача выражения неявного знания приобрела особую практическую актуальность на фоне цифровизиции всех сфер человеческой деятельности на основе интернет-технологий и умных технологий. Для углубления понимания тенденций мировых и отечественных исследований по тематике неявного знания в контексте технологий были проанализированы результаты запросов в Российской библиотеке eLibrary и в интернациональной базе научных публикаций Science Direct (Каныгин, Кононова, 2020). Результаты графически представлены на рис. 2.10.

В работах по проблемам управления знаниями и другим смежным областям, например архитектуре предприятия и архитектурам информационных систем, нередко используются средства визуализации, такие как диаграммы (Paquette, 2002; Sliwa, Patalas-Maliszewska, 2015; Virtanen, 2010, 2011), которые часто называются концептуальными моделями (Chergui et al., 2020; Dragicevic et al., 2020). Такого сорта визуализация способствует прояснению неявного знания, которым обладают сами исследователи или архитекторы. Остается открытым вопрос о том, как подобную визуализацию могут использовать носители неявного знания, практически организующие свои социальные взаимодействия.

Модельный подход, представленный в работе (Chergui et al., 2020), предполагает, что неявное знание может быть выражено в виде модели с помощью уже существующих методов и подходов. Но это не решает проблемы построения такой модели, которая позволяла бы решать практические проблемы функционирования знания не в виде «словесных рекомендаций», разработанных под конкретную организацию или группу организаций, а посредством общенаучной идеи моделирования. Идея исходит из необходимости представлять знания в обозримом, операбельном, понятном и доступном виде (Paquette, 2002) При этом учет дуальности любого знания проводится в духе идей Нонака и

Такеши, т.е. с ориентацией на моделирование процессов обмена знаниями на уровне организации.

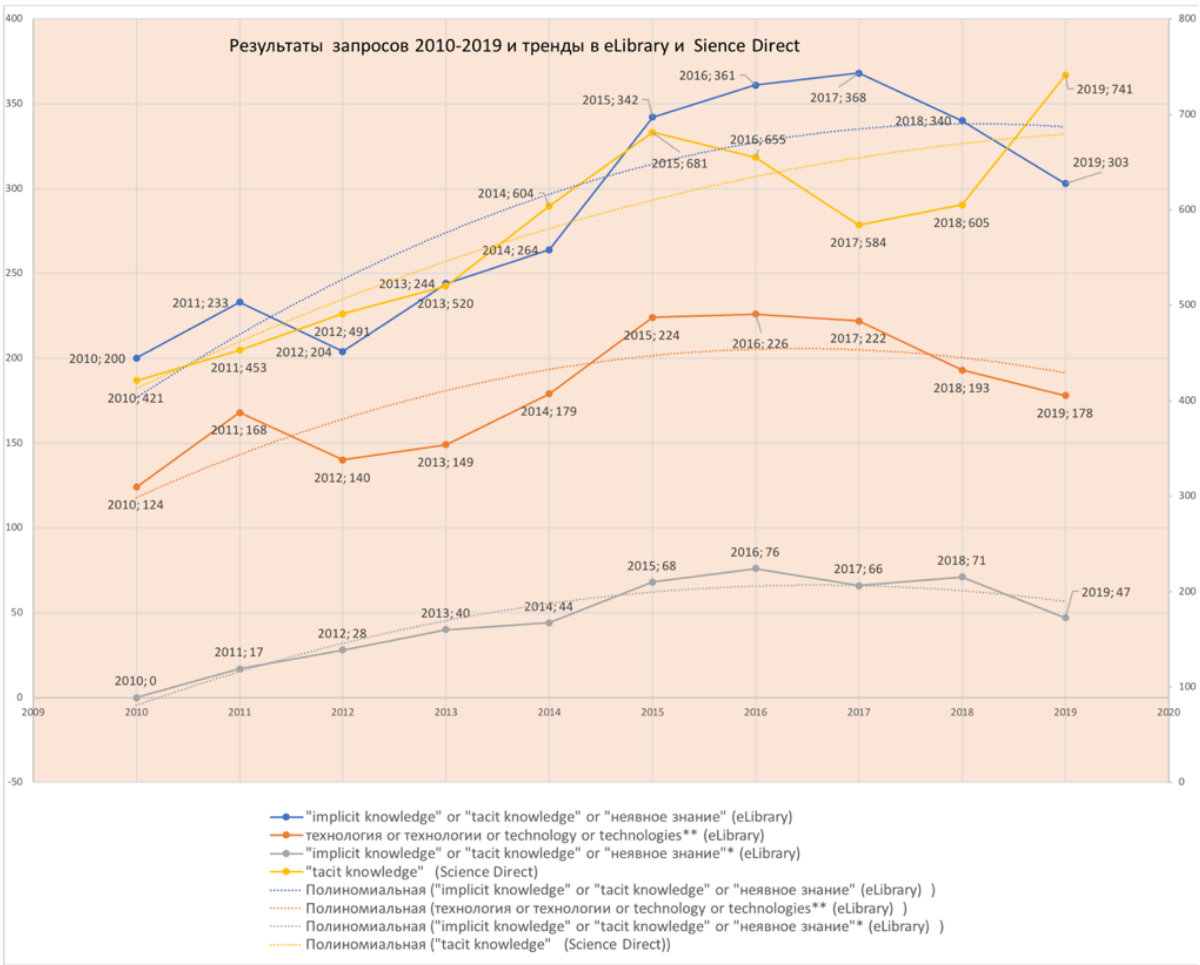



Рис. 2.10. Динамика интереса к теме неявных знаний на основе запросов *(*поиск только в ключевых словах и названии; ** поиск в результатах полнотекстового запроса: "implicit knowledge" or "tacit knowledge" or "неявное знание")*

Фактическим эталоном модели оказывается компьютерная онтология, накапливающая экспертные знания и делающая их доступными для дальнейшего использования (Mezghani, Expósito, Drira, 2016; Chergui et al., 2020). Однако существуют также концептуальные модели, основанные на представлении о неявном знании, но не доведенные до компьютерного воплощения (Sliwa, Patalas-Maliszewska, 2015; Оселедчик, 2017).

## 2.3. Контекстное знание и контексты

### 2.3.1. Континуум «Неявное знание – Контекстное знание – Контекст»

Континуум «Неявное знание – Контекст – Контекстное знание» требует отдельного рассмотрения. Логику, заложенную в данную цепочку рассуждений,

нельзя считать однозначной. Однако без разъяснения данной логики переход к рассмотрению и тем более практическому использованию в исследовательских целях как самих контекстов, так и платформ, сред и программ (в том числе класса BI), предназначенных для выявления, экспликации и анализа данных и знаний различной модальности, не представляется эффективным. Данный факт вызван тем, что каждый класс ПО, каждая отдельная ИС использует уникальный набор контекстов, даже если не определяет данное понятие явным образом. Поэтому так важно определиться с тем, что понимать под контекстом и как это понятие использовать в дальнейшем для осознанного выбора аналитического инструментария и для достижения требуемых и наиболее точных результатов анализа информационных массивов.

Контекст играет важную роль во многих областях, особенно для таких действий, как прогнозирование изменений контекста, объяснение непредвиденных событий и помощь в их преодолении, а также помощь в привлечении внимания. Понятие и теория контекста получает признание в 50-х годах 20-го века. 70-е годы ознаменованы возникновением контекстно-зависимых грамматик в рамках теории формальных грамматик. 80-е выявили связь контекстов с искусственным интеллектом. Исследователи пришли к выводу о необходимости учета контекстов как факторов, влияющих на эффективность задач формализации рассуждений и распознавания естественного языка. К 90-м контекст становиться предметом научного интереса специалистов многих направлений – не только философов, лингвистов, логиков, но и информатиков (the computer scientists). Сегодня понятие контекста имеет еще более широкое применение, затрагивая не только отдельные предметные области, но и новейшие междисциплинарные направления исследований на базе ИКТ. Тем не менее, высокая востребованность понятия не привела к формированию единого взгляда на понятие даже в рамках какого-то одного научного направления. Трактовки и определения термина изменяются от исследователя к исследователю, одновременно ширится спектр применения контекста в научных и практических задачах.

Контекст (context) и связанное с ним контекстное знание (contextual knowledge) относят к неявному знанию. Трудности или невозможность определения понятия контекста без учета людей, вовлеченных в ситуацию, связывают именно с тем, что «контекст включает в себя знания, которые не являются явными»[23]. В случае участия людей контекстное знание легко процедурно оформляется и становится неявной частью рассуждения, а затем может быть извлечено инженерами знаний.

Контекстное знание можно определить как:

[23] P. Brezillon and J.-Ch. Pomerol Contextual Knowledge and Proceduralized Context AAAI Technical Report WS-99-14. Compilation copyright © 1999.

– всё релевантное знание, которое может быть мобилизовано для понимания конкретной проблемы принятия решения;

– контекстное знание, вызываемое ситуациями или событиями: когда ситуации или события происходят, большая часть этого контекстного знания может быть процедурно обработана в соответствии с текущей направленностью принятия решения;

– знание, содержащееся в различных контекстах (полученных в результате полнотекстовых запросов), может существовать в различных видах и формах, может быть извлечено и изучено в результате полнотекстового поиска и последующей его обработки;

– форма интерпретации текстов и входящих в них концептов, находящихся в распределенной информационной среде и полученных в результате экспликации их смысла с помощью информационных технологий.

Таким образом, можно определить контекст как «набор всех знаний, которые может привлечь человек, столкнувшийся с некоторой ситуацией и имеющий неограниченное время, чтобы обдумать ее. Более того, контекст обладает временным измерением, что порождает некоторые проблемы при моделировании. Некоторые предположили, что контекст связан с взаимодействиями между агентами, в отличие от контекста как фиксированного понятия, относящегося к конкретной проблеме или области приложения. Это означает, что без взаимодействующих агентов не было бы контекста. Контекст появляется как общее пространство знаний. Однако каждая сущность, участвующая во взаимодействии, имеет свой собственный контекст, который может или не может согласовываться с некоторыми частями контекстов других. В рамках инженерии знаний термин «контекст» имеет некоторые общие черты со сценариями, фреймами или схемами, разработанными человеком в процессе познания. Контекст является кандидатом на то, чтобы храниться в долговременной памяти и восприниматься как единое целое, как жизнеспособная единица задачи, подходящая для некоторого шага в принятии решения»[24].

В отличие от контекста как фиксированного понятия, относящегося к конкретной проблеме или области приложения, существует понимание подвижного, процедурного контекста, возникающего при взаимодействии между двумя и более агентами, как программными, так и человеческими. Это означает, что без взаимодействующих агентов нет и контекста. Контекст проявляется как общее пространство знаний. Однако каждая сущность, участвующая во взаимодействии, имеет свой собственный контекст, который может или не может согласовываться с некоторыми частями контекстов других. Затем, когда фокус внимания перемещается, возникает переход от контекстного знания к процедурному контексту (рис. 2.11).

[24] Там же.

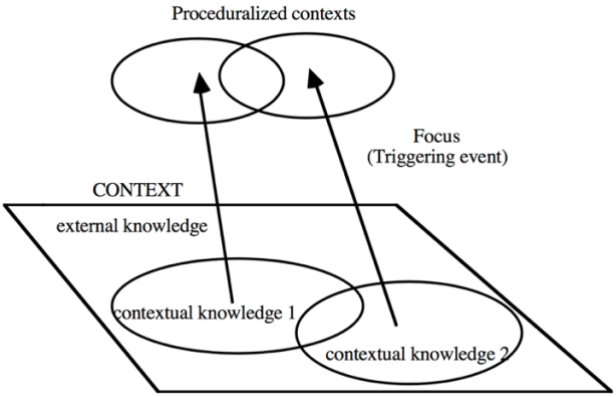



Рис. 2.11. Разные виды контекстов (P. Brezillon and J.-Ch. Pomerol, 1999)

Контекст используется в интерактивных системах, при этом, обладая временным измерением создает проблему представления и проблему границ при моделировании. Проблема границ возникает из-за того, что маловероятно учесть все, что может влиять на систему, включая контекстные данные незначительного или редкого влияния. Может возникнуть соблазн исключить такую информацию, что приведет к потере знаний, которые могут действительно иметь значение для моделируемых ситуаций. В то же время, не учитывая людей, вовлеченных в ситуацию, нельзя определить концепцию контекста, потому что контекст включает в себя знания, которые не являются явными, являясь одновременно общеизвестными знаниями (например, тот факт, что на станции легче организовать аварийные операции, чем в туннеле). Когда рассуждения приводят к такому типу знаний, они легко поддаются процедурной обработке и становятся неявной частью рассуждений, которые могут быть извлечены инженерами по знаниям и затем включены в операционную модель.

Каждый вовлеченный человек использует много знаний, объем которых отличается от одного человека к другому. Мы можем определить контекстное знание как все релевантное знание, которое может быть мобилизовано для понимания данного ситуативного решения проблемы (a given situated decision problem). Ситуативное решение подразумевает датированные, четко определенные обстоятельства, так как контекстное знание вызывается ситуациями и событиями и слабо связано с задачей или целью. Однако, когда задача становится более точной (например, происходит событие), большая часть этого контекстного знания может быть процедурно обработана в соответствии с текущей направленностью принятия решения. Хотя контекстное знание теоретически существует, на самом деле оно неявно и латентно и не может использоваться, если не возникает цель (или намерение).

В таком определении контекстные знания – это часть общего контекста (рис. 2.11). Таким образом, остальная часть контекста, не имеющая отношения к данной ситуации, называется внешним знанием. Процедурный контекст – это часть контекстного знания, которое вызывается, структурируется и располагается

в соответствии с заданным фокусом и является общим для различных людей, участвующих в принятии решений. Процедурный контекст может быть скомпилирован и, как правило, может быть обнаружен с помощью обычных методов приобретения знаний.

На некотором этапе принятии решения у человека есть: процедурный контекст, который представляет собой знание, широко известное участникам проблемы и непосредственно (но неявно) используемое для решения проблемы; контекстное знание, которое является знанием, которое не используется явно, но влияет на решение проблемы; и внешнее знание, не имеющее ничего общего с текущим этапом принятия решения, но известное многим участникам проблемы[25]

Применительно к научному исследованию, аналитической деятельности контекстное знание определяется как умение правильно «читать» контекст, извлекать и интерпретировать профессионально значимую информацию из любых источников. анализировать текст и нетекстовые модальности информации сначала на более высоком (и более абстрактном) уровне – уровне обобщенных структурных инвариантов контекстного знания, а затем редуцировать и специфицировать их применительно к конкретным дисциплинарным обстоятельствам и междисциплинарным отношениям (многоуровневый структурный анализ) контекстов и структур контекстного знания. Кластеры контекстного знания, соотнесенные с термин-концептами, позволяют определить тенденции формирования новых смыслов и новых контекстуальных значений научных терминов. Экспликация представляет уточнение понятий и утверждений естественного и\или научного языка с целью устранения выявленных в них неясностей и неточностей применительно к объекту научного исследования.

Уникальность проблем выражения контекста при коммуникации в сети Интернет заключается в их коротком, зашумленном, контекстно-зависимом и динамичном характере, а также в том, что употребляемые в социальных сетях сущности могут упоминаться как явно, так и неявно. Объекты, которые упоминаются, называются «неявными объектами». Проблема неявного связывания сущностей заметно отличается от явного связывания сущностей при выражении контекста. Особенности неявных упоминаний сущностей требуют нового подхода к решению этой проблемы, а именно моделирование сущности, используя фактическое и контекстное знание сущностей, чтобы дополнить контекст, выраженный в тексте сообщений. Контекстные знания должны

[25] В литературе есть несколько похожих взглядов. Например, теория Андерсона (1983, 1993) предполагает, что знания сначала приобретаются в декларативной форме, которая кодирует основные факты и примеры, содержащиеся в инструкциях (наши контекстные знания). После приобретения это знание используется общими правилами решения проблем для создания правил, специфичных для данного контекста (нашего процедурного контекста). Есть еще одна параллель с * global-context * и * local-context * в (Carenini and Moore 1993). * Global-context * указывает текущую обсуждаемую тему. В нем есть место в истории диалогов, где эта тема была начата. * Local-context * указывает на последнее высказывание. Теория Шенка предполагает, что контекст – это набор всех возможных историй, случай приблизительно соответствует нашему процедурному контексту, в то время как контекстуальное знание – это набор парадигматических случаев.

охватывать актуальные во времени темы и другие сущности, связанные с интересующей сущностью (Perera Sujan et al., 2016).

### 2.3.2. Определения и использование понятия контекста в разных предметных областях

Понятие контекста устанавливается через соотношение понятий «текст», «смысл» и «значение» и отражает специфику предметной области, в которой это понятие используется и интерпретируется неоднозначно в зависимости от предметной области, исследовательских задач (теоретических или практических) и объекта исследования. Проблема, связанная с выбором и употреблением правильных контекстуальных значений научных терминов, приобретает важное значение в междисциплинарных исследованиях или при переносе терминов из старых теорий в новые с изменением первоначальных значений.

Контексты отражают:

1) специфику предметных областей научного и ненаучного знаний;

2) виды носителей информации (текст, нетекстовые модальности информации);

3) метафорический перенос термина из области научного дискурса в сферы практической деятельности.

Определение контекста, общее для ряда наук, дается в энциклопедических и философских словарях. Контекст (от лат. contextus - соединение, связь) – «отрывок письменной или устной речи, в пределах которого можно уяснить значение отдельного входящего в него слова (или фразы). Только в контексте слово получает конкретное значение. Конкретизация понятия контекста и его видов осуществляется применительно к специфике ряда предметных областей науки. Обобщение подходов к пониманию сущности контекста находит отражение в функциональности электронных библиотек, аналитических платформ и компьютерных программ, терминологическом аппарате технологий интеллектуального анализа текстов, таких как машинное обучение.

В математической логике контекст определяется как «некоторая локализованная в пространстве и времени совокупность высказываний и терминов, в которую входит… исследуемый термин»[26]. Общее понятие контекста «дифференцируется на экстенсиональный контекст, в рамках которого эквивалентность и взаимозаменяемость выражений устанавливается по признаку объема, и интенсиональный, где логическая взаимозаменяемость определяется по критерию содержания»[27].

В лингвистике дается определение текстуального (текстовой модальности, текстового) контекста как совокупности лексических единиц, в окружении которых используется конкретная единица текста. Текстуальный контекст может

[26] Кондаков Н. И. Логический словарь-справочник. Изд. 2, доп. М.: Наука, 1975. С. 262.
[27] Можейко М. А. Контекст // Новейший философский словарь. Мн., 2001. С. 502.

быть представлен как оппозиция микроконтекста (понятие, предложение, фрагмент текста) и макроконтекста (широкий контекст без установления определенных рамок текста). Таким образом, микроконтекст – это непосредственное лингвистическое окружение единицы информации (например, слова) в словосочетании или предложении. Макроконтекст – это языковое окружение единицы информации (например, слова), выходящее за рамки предложения. Размер макроконтекста не ограничен: это может быть совокупность абзацев, глава, весь документ в целом. Следовательно, контекст можно определить как фрагмент текста, словесное окружение избранного для анализа элемента текста, определенную совокупность слов, грамматических форм и конструкций, в которых было использовано данное слово. Это определение важно для дальнейшего понимания трактовок понятия и их применения в аналитических целях.

В качестве альтернативы лингвистическому контексту выделяют экстралингвистический контекст, который определяют как совокупность фактов, способствующих правильной интерпретации текста (исторической эпохи, места, особенностей культуры и т.п.) в совокупности с опытом и знаниями участников коммуникации. Экстралингвистический контекст рассматривается как метафорический перенос понятия макроконтекста из лингвистики в сферу социальных дискурсов. При такой трактовке понятия, применительно к аналитическому инструментарию, речь идет уже не о формате реализации функционала этого инструментария, а о последующей интерпретации аналитических результатов, полученных с использованием этого инструментария.

В компьютерном программировании существует понятие контекста устройства (device context). Контекст устройства – это «системная структура данных, которая содержит характеристики устройства вывода и дескрипторы выбранных графических объектов и режимов рисования»[28].

Школой А.А. Вербицкого классическая доминирующая научная позиция рассматривает контекст как некоторый структурный фрагмент семиотической системы (в нашем случае текст), формирующий значение и смысл некоторого другого фрагмента (слова, ситуации, объекта). Данную позицию можно определить как статическую. Статическому определению контекста противопоставляется динамическая позиция, которая определяет выбранный термин как динамический когнитивный механизм, осуществляющий взаимодействие двух типов контекстов:

– «внутренних» контекстов, а именно, индивидуально-психологические особенности, знания и опыт человека, которые выполняют роль «различных

[28] Сучкова, Л.И. Win32 API: основы программирования: учебное пособие/ Л.И. Сучкова; АлтГТУ им. ИИ. Ползунова. -Барнаул, АлтГТУ, 2010. С. 138

психических функций и процессов для решения семиотических задач – порождения смыслов путем соотнесения различных психических содержаний»,

– «внешних» контекстов, а именно, предметных, социокультурных, пространственно-временных и иных характеристик ситуации, в которых действует человек (условий жизни, деятельности, образования, личного и профессионального опыта человека, «определяющих понимание и преобразование им конкретной ситуации, придающих смысл и значение этой ситуации как целому»)[29].

Выделяется несколько оппозиций контекста.

Оппозиция по степени вовлеченности исследователя:

– коммуникативный контекст существует независимо от исследователя и включает организацию средств и многообразных видов контекстов для передачи восприятия смысла сообщения;

– операционный контекст (с его помощью исследователь анализирует семантические единицы речевого текста для установления подтекста).

Оппозиция по объекту исследования:

– системно-структурный контекст;

– процессуальный контекст.

Оппозиция по модальности: контексты текстовой и нетекстовой модальности.

Оппозиция по специфике предметных областей:

– лингвистический контекст;

– экстралингвистический контекст (исторический, психологический, социальный, др. виды).

Оппозиция по способам экспликации: микроконтекст (в пределах авторского абзаца) и макроконтекст (в пределах документа или их совокупности), позволяющие выявить кластеры контекстного знания, соотнесенные с термин-концептами.

Оппозиция по области охвата:

– горизонтальный контекст (ближайшее окружение слова в предложении);

– вертикальный контекст (крупный фрагмент текста, в котором его значение может повергаться серьезным смысловым трансформациям).

Атрибутивными характеристиками контекста являются смыслы или значения терминов, термин-концептов (концептов) и ситуаций. «Концепты – это интерпретаторы смыслов, форма обработки субъективного опыта путем подведения его под определенные категории и классы», совокупность концептов организуются в иерархические семантические сети[30]. Совокупность тематически организованных терминов в рамках одной предметной области

[29] Вербицкий А.А. Контекст (в психологии) / Под ред. А.В. Петровского // Психологический лексикон. Энциклопедический словарь в 6 т. Т. 3. М.: ПЕР СЭ, 2005. С. 137-138.

[30] Микешина Л.А. Философия познания. Полемические главы. М.: Прогресс-Традиция, 2002. 624 с. С. 506.

(междисциплинарном направлении) составляет тезаурус предметной области или дисциплинарную онтологию.

2.3.3. Выводы

Современных знаний становятся все больше и больше. Огромный объем информации, которую необходимо хранить и управлять разумным образом, мотивировал интенсивные исследования в области управления большими базами знаний с учетом их модульности. Поэтому контекстуализацию (разложение знаний в соответствии с различными точками зрения) следует рассматривать как неотъемлемую часть концептуализации (мысленный процесс, который приводит к созданию описания некоторой области интереса). Ход многочисленных экспериментов российских и зарубежных исследователей показал, что путем дальнейшего развития контекстных методов можно создать универсальную структуру для сбора и эффективного управления знаниями, содержащимися в Интернет. Эта структура будет охватывать все аспекты управления большими данными и контекстно-ориентированными вычислениями, что позволит проводить интеллектуальную обработку информации и данных, производимых в настоящее время на повседневной основе.

## 2.4. Модальность и типология контекстов

2.4.1. Общие подходы к типологизации контекстов

Удовлетворение информационных потребностей в междисциплинарных научных исследованиях может достигаться применением технологий и методов извлечения контекстных знаний из больших массивов данных. Изучение структурно-организованных, автоматически извлекаемых из неструктурированных источников тематических и смысловых контекстов имеет общенаучное и научно-практическое значение, поскольку контекстное знание составляет важную часть знания, которым оперирует человек в научной, образовательной, социальной, культурной и предпринимательской деятельности.

Постоянно растущее число сайтов, социальных сетей и повсеместное распространение мобильных устройств привело к размещению в сети Интернет сведений, имеющих большой потенциал для исследователя, аналитика, представителя государства и бизнесмена. Однако поиск и экспликация данных без помощи инструментов автоматической фильтрации не позволяет пользователям обнаруживать необходимую им информацию, что обусловлено увеличением количества избыточной или неполной, дублирующей, недостоверной, низкого качества информации. Исходный текст, полученный в результате поиска, слишком велик, информация сложна, и присутствует много нежелательного шума. Возникла парадоксальная ситуация, связанная с поиском информации в сети Интернет – своего рода информационный кризис, вызванный невозможностью адекватно оценить, управлять информацией, потерей доверия к

ряду источников информации. Кроме того, публикуемый в сети контент зачастую частично или полностью не структурирован. Как следствие, возникают проблемы с его машинным распознаванием, что ограничивает возможности вычислительных машин. Выявление и затем использование контекстов для сужения области вторичного, повторного поиска и, главное, интеллектуального анализа текстов может значительно повысить релевантность получаемых результатов. Экспликация на основе контекстов позволяет создавать персонализированные результаты для пользователей – индивидуальные коллекции релевантных запросам контекстов. Переход от поиска по ключевым словам к контекстному поиску приводит к машинно-интерпретируемому контенту в сети, который может использоваться агентами для автоматического заключения о веб-контенте.

Экспликация контекстного знания в распределенной информационной среде с помощью информационных технологий ориентирована на выполнение исследовательских задач по формированию новых дисциплинарных и междисциплинарных онтологий, что позволяет сформулировать несколько основных задач его применения:

– типологизация контекстного знания на уровне контекстов для установления предмета исследования, границ и единиц его «измерения», обеспечения возможности выбора методов и инструментов анализа;

– получение контекстных знаний (acquiring contextual knowledge): контекстные знания могут быть извлечены из текстов, в которых явно упоминается объект, отраженный в поисковом запросе; первичная обработка информационных массивов – экспликация контекстов – позволит сократить объем информации, актуализировать информационное содержание, повысить степень адекватности массива поисковому запросу, выделить явные и неявные сущности, относящиеся к проблематике исследования;

– выявление и связывание неявных сущностей (implicit entities), определение ценности такого добавления применительно к результатам стандартной задачи связывания сущностей. Модели неявных сущностей, использующие как фактические, так и контекстные знания о сущностях, демонстрируют более высокую точность, чем современные подходы к связыванию сущностей (state-of-the-art entity linking approaches) [31];

– построение терминологического ландшафта и формирование тезауруса направления исследований.

В целях изучения структурно-организованных, автоматически извлекаемых из неструктурированных источников тематических и смысловых контекстов необходима систематизация многообразных видов контекста и контекстного знания, определяемых задачами исследования. Трактуемое в рамках пособия как

[31] Sujan Perera, Pablo N. Mendes, Adarsh Alex, Amit P. Sheth, and Krishnaprasad Thirunarayan Implicit Entity Linking in Tweets. http://www.internetlivestats.com/twitter-statistics.

независимая понятийная единица категориального аппарата, понятие контекста может быть положено в основу классификации научных текстов, визуализации иерархических и ассоциативных отношений между терминами, формирования типологии контекстного знания, которую можно в дальнейшем специфицировать применительно к изучению более определённых предметных областей.

В научных исследованиях приходится сталкиваться с контекстами как текстовой модальности, так и нетекстовой модальности, что нашло отражение в типологии контекстов. Структура типологии, представленная в таб. 2.1, выделяет две обобщенные группы – контексты текстовой модальности и контексты нетекстовой модальности. Следует отметить, что исследователи различают мономодальность и более сложный и интегрированный вариант подачи информации – мультимодальность данных, документа, сообщения. Анализ контекстов корпуса мультимодальной информации, как правило, осуществляется по каждой из модальностей отдельно и часто независимо, поэтому для каждой из обобщенных групп предложена своя типология контекстов.

Таб. 2.1. Структура типологии контекстов

| Обобщенная группа: Контексты текстовой модальности | Обобщенная группа: Контексты нетекстовой модальности |
|---|---|
| Группы контекстов | Группы контекстов |
| Виды или описание контекстов каждой группы | |

Типология контекстов текстовой модальности представлена контекстами, которые отличаются назначением, типом или характеристикой контекста и распадаются на отдельные виды. Использование тех или иных видов контекстов определяется задачами исследования и выбором технологий, а также используемым инструментарием для интеллектуального анализа данных. Например, рассмотрим тематические контексты (корпус, фрагмент, абзац, предложение). Электронная библиотека T-Libra рассматривает в качестве горизонтального контекста абзац текста, содержащий искомый термин-концепт; в качестве вертикального контекста фрагмент – совокупность абзацев: абзац, содержащий искомый термин-концепт и по 3 абзаца сверху и снизу от выделенного. В аналитической среде Voyant-Tools горизонтальный контекст – это предложение, содержащее термин-концепт; вертикальный контекст – фрагмент текста произвольного размера, в том числе полный документ. Как правило, набор контекстов текстовой модальности, как единиц анализа данных, задается внутренним стандартом и отражается на возможностях и особенностях реализации любой аналитической системы[32].

Знание контекста и предметной области – это знание области дисциплины, человеческой деятельности, которые имеют явно определенные сущности. Все

[32] T-Libra (http://77.234.221.107/bin/TauC.exe?DSN=tlibra)
Voyant-Tools (https://voyant-tools.org)

эти сущности можно отследить и устранить неоднозначность, когда пользователь выполняет запрос. Анализ запросов, используемых пользователем при поиске, может быть направлен на то, чтобы спрогнозировать ожидания пользователя. Сущности могут быть исследованы, и отношения между ними могут быть восстановлены и использованы при получении достаточных оснований для интерпретации выполненного запроса. В типологии, таб. 2.2, предложены новые виды контекстного знания: тематический запрос, тезаурус, предметно-тематический тренд, коллекции запросов и материалов.

Таб. 2.2. Типология контекстов текстовой модальности

| Обобщенная группа: Контексты текстовой модальности | | | |
|---|---|---|---|
| Группы | Типы *контекста* | | *Виды* |
| 1 Тематические | Вертикальный (задается размером) | Макро-контекст | Корпус |
| | | | Фрагмент |
| | | | Абзац |
| | Горизонтальный (задается размером) | Микро-контекст | Предложение |
| 2 Структурный | | | Термин-концепт |
| | Структурированный | | Тезаурус |
| | | | Мета-описание |
| 3 Запросы | Простой | | Подборка запросов |
| | Расширенный | | Подборка запросов |
| 4 Тренды | Динамический (задается на временном интервале) | | Единичный термин-концепт |
| | | | Семантическая группа |
| 5 Коллекции | Результаты поискового запроса | | Подборка публикаций |
| | Результаты каскадного запроса на подборке публикаций | | Тематические коллекции релевантных абзацев |
| | Содержательный | | Группа ключевых слов |
| | Междисциплинарный | | Перечень предметных областей |

Контексты нетекстовой модальности, с одной стороны, являются объектами научных исследований, а, с другой, подходы к их изучению образуют методологическую базу самих исследований. Поэтому для дальнейшего развития методов контекстного анализа в научных исследованиях представляется важным как изучение контекстов нетекстовой модальности, так и развитие методов и технологий их обработки и анализа. Контексты нетекстовой модальности – более сложная для изучения обобщенная группа, разнообразие которой полностью определяется свойствами, присущими той или иной модальности. Следует учитывать, что модальность задается принадлежностью данных к некоторому источнику, определяющему структуру, формат, структурно-функциональными связи данных, процедуры обработки и анализа данных. Выделяют следующие типы модальности: обонятельная, осязательная, вкусовая, зрительная и слуховая и, соответственно, мультимодальными явлениями считаются взаимодействия

между вербальными текстами и изображениями, видео, речью и жестами, размером и цветом текста (Кресс Г., 2009).

Таб. 2.3. Типология контекстов нетекстовой модальности

| Обобщенная группа: Контексты нетекстовой модальности | |
|---|---|
| Изображения | рисунки, фотографии, инфографика (infographics), схемы изображений (Image schemas), типографика |
| Аудио контент | звуки и сигналы |
| Видео контент | визуальные характеристики видеоконтента |
| Другие органы чувств | обонятельная, осязательная, вкусовая |
| Языки | многоязычный набор текста, языковые конструкции и стили, |
| Иные виды контента | метаданные документа и файла, токены, находящиеся внутри текста, встроенные в текст |

Контексты нетекстовой модальности могут быть рассмотрены и анализироваться сами по себе или как часть документа, содержащего одновременно данные (контексты) нескольких модальностей. В этом случае идет речь о мультимодальных и мультиязычных, документах, системах, поиске, моделях. Таким образом, мультимодальным является любой текстовый документ, объединяющий в себе две или более семиотических системы для создания смысла, различные семиотические коды, требующие актуализации сразу нескольких перцептивных каналов, визуального и аудиального, которые задействуются для составления и передачи сообщения. Технически мультимодальность – это концепция интеграции информации нескольких модальностей с целью прогнозирования. Мультимодальный поиск – это поиск, который позволяет осуществлять запросы, включая в запрос токены разной модальности последовательно или параллельно (гибридный запрос). Кроссмодальное взаимодействие элементов мультимодального документа осложняет процесс поиска и отбора контекстного знания по сравнению с мономодальным текстовым документом или мультимодальным документом, нетекстовые элементы которого распознаются посредством обработки мета-описаний. С другой стороны, интрамодальное моделирование и интрамодальное взаимодействие между текстовыми, визуальными и акустическими компонентами документов усугубляют проблемы поиска и идентификации контекстов[33].

Документ, можно определить как мультимодальный, если он объединяет две или более семиотических систем для создания смысла (рис. 2.12). Обычно семиотика – это исследование того, как создается значение и как значение передается. Семиотические системы можно разделить на следующие категории (The New London Group 2000):

[33] интрамодальность – принадлежность к одной модальности; кроссмодальность – принадлежность к разным модальностям

– Лингвистическая (языковая): словарный запас, структура, грамматика устного / письменного языка.

– Визуальная: цвет, векторы и точка обзора в неподвижных и движущихся изображениях.

– Слуховая: громкость, высота и ритм музыки и звуковых эффектов.

– Жестикуляционная: движение, выражение лица и язык тела.

– Пространственная: близость, направление, расположение макета, организация объектов в пространстве.

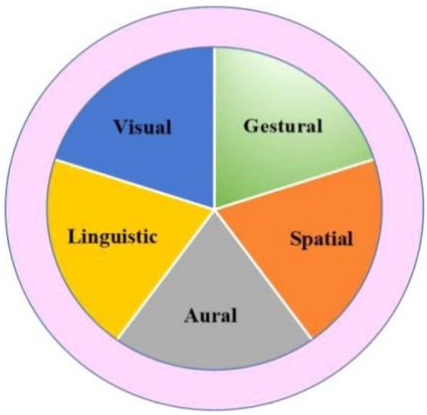



Рис. 2.12. Мультимодальность (согласно A. Kumar and al. [34]).

В качестве ориентира для восприятия сложности и разнообразия контекстов и как основа для развития типологии контекстов выделены токены нетекстовых модальностей применительно к модальностям базового, низкого уровня:

1. Метаданные документа – авторы, время, источник, рубрики, и т.д., метаданные файла (в зависимости от ситуации и исследовательского подхода могут рассматриваться как токены текстовой или нетекстовой модальности.

2. Токены, находящиеся внутри текста – ссылки, теги, словосочетания, именованные сущности, ссылки на изображениях, записи о действиях пользователей, и т.д.

3. Языки: многоязычный набор текста; в системах кроссязычного и мультиязычного тематического поиска запрос даётся на одном языке, а ответ может быть получен на других языках.

4. Типографика (художественная способ представления текста).

5. Рисунки, фотографии всех форматов.

6. Языковые конструкции и стили, которые включают использование символических языковых маркеров, таких как пунктуация, эмодзи (или другие пиктографические изображения), микротекст, мемы (вирусное изображение,

[34] Akshi Kumar, Kathiravan Srinivasan, Albert Y. Zomaya Hybrid context enriched deep learning model for fine-grained sentiment analysis in textual and visual semiotic modality social data / Information Processing & Management, Volume 57, Issue 1, January 2020, 102141 https://www.sciencedirect.com/science/article/abs/pii/S0306457319306934.

видео или словесное выражение для мимикрии или юмористических целей), анимированные GIF-файлы (формат обмена графическими данными, который объединяет несколько изображений или «кадров» в один файл для передачи движения).

7. Инфографика (infographics) (текст, встроенный вместе с изображением) – схемы, графики и т.п.

8. Схемы изображений (Image schemas) – скелетные концептуальные структуры, возникающие в результате восприятия, движений тела, манипулирования объектами, процессами и силами, такие как вверх / вниз, сила, противодействие или вход / выход. Схемы изображений в основном используются для объяснения связи между воплощенным опытом (особенно базовым пространственным опытом) и высшим познанием. Ощутимые различия схемы от рисунков состоят в том, как схемы формируют опыт, и в том, что передаваемое ими знание зависит от дополнительного концептуального элемента – точки зрения. Добавление точки зрения как характеристики схемы изображения означает, что один и тот же тип структуры может давать разные экспериментальные результаты. Таким образом, схемы отражают не только объекты, распознаваемые традиционными методами, но и мнения, представления с позиции определенного ракурса, а значит, содержат дополнительный контекст, выявление и распознавание которого требует неординарных подходов.

9. Аудио- и видеоинформация:

a. интонационные инструменты воздействия, жесты, мимика, паузы, смех и т.д. образуют знаковую систему, дополняют или заменяют средства вербальной коммуникации. «Внешность оратора», актора также можно рассматривать как токен и контекст;

b. субтитры, реплики, звуковые эффекты, речь, акустические события и т.п. из аудио и визуальных каналов;

c. движения камеры, методы съемки и тип действий считаются одними из наиболее важных визуальных характеристик видеоконтента, с помощью которых отличают различные жанры и техники съемок;

d. цвета – для моделирования цвета и освещения могут извлекаться следующие визуальные признаки: гистограммы RGB, гистограмма значений, гистограмма насыщенности;

e. лица и способ их изображения: характеристики, связанные с лицами: количество обнаруженных лиц в кадре и отношение площади ограничивающей рамки лица к общему размеру кадра.

Таким образом, экспликация и понимание контекста, особенно контекста нетекстовой модальности – это один из самых сложных аспектов модерации контента. Ситуация усугубляется особенностями распознавания и классификации модальности, налагаемые предметной областью и междисциплинарными направлениями исследований.

### 2.4.2. Мультимодальность. Кейс «Здравоохранение»[35]

В качестве примера предметной области, имеющей несомненную специфику, можно привести здравоохранение. Здравоохранению свойственно использовать классификацию изображений, контексты нетекстовой модальности базового уровня, применяемых в медицине. Модальность изображения – это фундаментальная визуальная характеристика изображения, которую можно использовать для повышения эффективности поиска. Однако аннотации или подписи, связанные с изображениями, часто не содержат информации о модальности и способе получения этого изображения.

Создаваемые большие коллекции медицинских изображений играют важную роль прежде всего в медицинских исследованиях и принятии клинических решений. Одна из основных проблем заключается в том, что размер коллекций постоянно растет из-за увеличения доступности оборудования для визуализации в больницах. Это создает огромные хранилища ценной информации, которую во многих случаях трудно обрабатывать и управлять ею должным образом. Это является основанием для разработки инструментов для эффективного и действенного доступа к этому типу информации, объединения методов визуализации и взаимодействия с больничными и ведомственными информационными системами для управления хранением и распространением изображений среди медицинского персонала, исследователей, клиник и центров визуализации. Задача индексации и каталогизации этих коллекций традиционно выполнялась вручную. Это дорогостоящая и трудоемкая процедура, к тому же она подвержена ошибкам, связанным с человеческим фактором. Следовательно, существует острая потребность в автоматизированном индексировании коллекций медицинских изображений, чтобы улучшить возможность поиска и извлечения соответствующих изображений. Системы поиска медицинских изображений традиционно основываются на тексте, относящегося к аннотации или подписям, связанных с изображением. Сочетание методов текстового и визуального поиска улучшает производительность по сравнению с их раздельным использованием. Поэтому в последние годы получили развитие методы мультимодального поиска, где в качестве входных данных выступают запросы, включающие последовательно или параллельно токены разной модальности (гибридный запрос). В этом случае запросы состоят из текстовой части (т.е. текстового подзапроса) и / или образцов изображений (т.е. визуального подзапроса). Например, запросы могут содержать информацию о демографических характеристиках пациентов, ограниченном наборе симптомов и результатах медицинских обследований, включая визуализационные исследования. Базы данных медицинских изображений, используемые для поиска

[35] Подготовлен на основе Dimitrovski I., Kocev D., Kitanovski I., Loskovska S., Džeroski S.Improved medical image modality classification using a combination of visual and textual features // Computerized Medical Imaging and Graphics. 2015. Volume 39. P. 14-26, DOI: 10.1016/j.compmedimag.2014.06.005.

или в учебных целях, часто содержат изображения, полученные различными методами (например, рентген, компьютерная томография, ультразвук и т.д.), а их аннотации разнородны и не систематизированы. Это особенно верно для изображений, которые можно найти в различных онлайн-ресурсах, включая те, которые обращаются к онлайн-контенту журналов.

Модальность изображения – это фундаментальная визуальная характеристика изображения, которую можно использовать для повышения эффективности поиска. Однако аннотации или подписи, связанные с изображениями, часто не содержат информации о модальности и способе получения этого изображения (рисунок 2.13).

**Классификация модальности**
**Modality Classification**

**Составные или многоканальные изображения (Compoud or multipane images)**

| Диагностические изображения (**Diagnostic images**) | | Общие биомедицинские иллюстрации (**Generic biomedical illustrations**) |
|---|---|---|
| **Радиология** | **Печатные сигналы, волны** | Таблицы и формы |
| Ультразвук | Электроэнцефалография | Список программ |
| Магнитный резонанс | Электрокардиография | Статистические цифры, графики, диаграммы |
| Компьютерная томография | Электромиография | Скриншоты |
| Рентген, 2D-рентгенография | **Микроскопия** | Блок-схемы |
| Ангиография | Световая микроскопия | Обзор систем |
| ПозитроЇнно-эмиссиоЇнная томограЇфия | Электронная микроскопия | Последовательность гена |
| Комбинированные методы в одном изображении | Просвечивающая микроскопия | Хроматография, Гель |
| **Фотография в видимом свете** | Флуоресцентная микроскопия | Химическая структура |
| Эндоскопия | **3D реконструкции** | Математика, формулы |
| Дерматология, кожа | | Неклинические фотографии |
| Другие органы | | Рисованные эскизы |

Рис. 2.13. Классификация модальности изображений (http://www.imageclef.org/2013/medical)

Для адекватного представления изображения необходимо уметь распознавать признаки, которые позволяют уловить различные аспекты изображения (например, текстуру, формы, распределение цвета, ...).

Характеристики локального изображения имеют основополагающее значение для интерпретации изображения: в то время как глобальные элементы сохраняют информацию обо всем изображении, локальные элементы фиксируют детали. Таким образом, они более разборчивы в отношении проблемы межклассовой и внутриклассовой изменчивости. Цель состоит в том, чтобы правильно классифицировать модальность изображений, используя визуальную информацию из изображений и текста статьи, в которой встречается это изображение.

Подход к классификации медицинских модальностей использует различные визуальные особенности в сочетании с текстовыми функциями, извлеченными из окружающего текстового содержания изображений. Также считается, что обработка данных доменно-специфична и зависит от типа данных (модальности и контексты) и источника данных. В качестве источника знаний помимо онтологий, таких как открытые биомедицинские онтологии (ОБО, англ. Open Biomedical Ontologies), или словарей, может использоваться любой источник медицинских данных, из которого можно сформировать иерархию знаний, например медицинские схемы-гайдлайны и т.п. Модальность данных определяется на основе следующих соображений: выделяются текстовая и нетекстовая модальности, последняя может быть декомпозирована сначала на базовые, «низкие» уровни, а затем по каждой из модальности – на уровни высшего порядка (контексты), общие и задаваемые направлением исследований и задачами, стоящими перед исследователями.

### 2.4.3 Мультмодальность. Кейс «Первазивная среда»[36]

Вычисления с учетом контекста устанавливают локальное взаимодействие между пользователем и вычислительными службами на основе доступного пользовательского, физического местоположения, вычислительного и окружающего контекстов. Интеллектуальные устройства реагируют на различные входные данные, которые идентифицируются и наблюдаются с использованием контекстов, которые обычно воспринимаются во всеобъемлющей среде. Устройства обмениваются контекстной информацией для того, чтобы улучшить взаимодействие друг с другом в интеллектуальной среде. Примеры соответствующей контекстной информации, используемой для восприятия первазивной среды (pervasive environment):

– Вербальный контекст.
– Роли партнеров по коммуникации.
– Цели общения, цели людей.
– Местная среда.

[36] Подготовлен на основе: Prabakaran N., Kannadasan R. Contexts enabled Decision Making using sensors to perceive pervasive environment // International Conference on Computational Intelligence and Data Science (ICCIDS 2018). Procedia Computer Science. 2018. Vol. 132. P. 477–485.

– Социальная среда.

– Физическая и химическая среда.

Контекстные устройства основаны на датчиках, фокусируются на данных и могут использоваться повторно. Распределенные гетерогенные устройства или узлы предназначены для распознавания крупномасштабной первазивной среды (pervasive environment) с использованием контекстов, которые развиваются внутри и вне среды, обеспечивая тем самым ее первазивность, т.е. общность свойств и характеристик, присущих одновременно физическому и виртуальному информационному пространствам. Технологии, используемые для получения контекста (technologies used for context acquisition), представлены в таб. 2.5.

Таб. 2.5. Технологии, используемые для получения контекста: сенсорные технологии (sensing technologies)

| | Context acquisition technologies | Технологии получения контекста |
|---|---|---|
| 1 | Light and Visualisation | Свет и визуализация |
| 2 | Hearing | Слух (детектор звука) |
| 3 | Motion and Acceleration | Движение и ускорение |
| 4 | Location and Locus | Геолокация и определение местоположения |
| 5 | Zero-Power Sensors | Датчики нулевой мощности |
| 6 | General Magnetic Field and Orientation | Общее магнитное поле и ориентация |
| 7 | Proximity | Бесконтактные технологии |
| 8 | User Interaction | Взаимодействие с пользователем |
| 9 | Touch | Прикосновение (Сенсорные технологии) |
| 10 | Pressure | Давление |
| 11 | Humidity | Влажность |
| 12 | Temperature | Температура |
| 13 | Weightiness | Вес |
| 14 | Movement Detection | Обнаружение движения |
| 15 | Electronic Noses | Электронный нос (детектор запахов) |
| 16 | Gas-Sensers | Газовые сенсоры |
| 17 | Bio-Sensers | Био-сенсеры |

Сбор мультимодальных и соответственно нетекстовых контекстов есть необходимое условие функционирования любой контекстно-зависимой системы. Обычно получение контекста, известное как процесс наблюдения за реальным фактами в окружающей среде, выявляет основные особенности, абстрактные иллюстрации, которые затем предоставляются компонентам системы для дальнейшего продвижения. Для получения контекстов нетекстовой модальности используются различные подходы, включая сенсорные системы, компьютерное зрение, отслеживание местоположения, моделирование пользователей и их поведения.

### 2.4.4. Методика автоматизированного извлечения и изучения контекстного знания к информационным ресурсам нетекстовой модальности

Применение технологий автоматизированного извлечения и изучения контекстного знания к релевантному массиву информационных ресурсов нетекстовой модальности предполагает следующую последовательность действий:

1) разработку или выбор подхода, основанного на извлечении метаданных контекстного знания нетекстовой модальности (автоматизированно, например, по QR-кодам, системам сканирования и интеллектуального распознавания визуальных объектов, нейросетям, системам искусственного интеллекта и пр.);

2) метаданные должны быть описаны по основе единого стандарта;

3) для описания метаданных необходимо использовать стандарт, например, тегов Dublin Core, что позволит создавать и интегрировать различные документационные базы данных и базы знаний с учетом мультимодальности источников:

– музейные коллекции;

– культурные объекты (архитектурные сооружения, памятники, скульптуры и т.п.);

– базы данных медицинской информации (рентгеновские снимки, компьютерные томограммы, кардиограммы и т.п.);

– и т.п.

4) определение вида контекстного знания нетекстовой модальности в соответствие с типологией, выбор соответствующего ему метода анализа и аналитического инструментария;

5) при отсутствии, недоступности специализированного аналитического инструментария изучение контекстного знания нетекстовой модальности осуществляется на основе методов интеллектуального анализа текстов и аналитических инструментов для анализа текстов. В качестве текста берутся метаданные или мета-описания нетекстовых элементов. Таким образом, сложная задача анализа мультимодальных документов и данных сводится к более простой – использованию устоявшихся и доступных широкому кругу исследователей методов и подходов.

**Контрольные вопросы к Главе 2**

1. Какие научные точки зрения на знания получили наибольшую известность и почему?

2. Какое практическое применение имеет теория явного и неявного знания?

3. В англоязычной литературе различают два вида и соответственно два термина для обозначения неявного знания – «tacit knowledge» и «implicit knowledge». В русскоязычных текстах иногда наряду с термином неявное знание используется термин имплицитное знание. В чем отличие этих двух видов знаний?

4. Какие классы задач требуют анализа неявного знания?

5. Какое знание называют контекстным? Назовите сферы применения этого знания?

6. В каких случаях используется понятие контекста, трактуемого как единица обработки и анализа информации?

7. Что такое модальность контекста или документа? Какие документы называют мультимодальными?

8. Какие виды контекстов используются в системах интеллектуального анализа текстовых данных?

9. Что такое горизонтальный и вертикальный контексты?

10. Что такое макро-контекст и микро-контекст применительно к системам анализа массивов текстовых данных и визуализации результатов этого анализа?

11. Укажите основное различие между неявным и явным знанием. Почему эти два знания взаимозависимы?

# Глава 3. Использование технологий извлечения и анализа контекстного знания в научно-исследовательской работе

## 3.1. Обзор методов интеллектуального анализа научных текстов

В авторской интерпретации интеллектуальный анализ научных текстов входит в направление исследований контекстного знания. Анализ настоящего состояния научных исследований показывает, что как в русскоязычном, так и в мировом научном дискурсе изучение контекстного знания ведется в основном по двум направлениям: теории и практике контент-анализа; теории и практике контекстного обучения[37].

1. Контент-анализ. Исследование контекстов различного типа, вида и уровня проводятся в рамках традиционного контент-анализа: метода и технологии качественно-количественного анализа документов в целях выявления или измерения социальных фактов и тенденций, отраженных этими документами. Контент-анализ изучает документы в их социальном контексте.

Все более широко распространяется контент-анализ сообщений средств массовой информации, основанный на парадигматическом подходе, в соответствии с которым изучаемые признаки текстов (содержание проблемы, причины ее возникновения, проблемообразующий субъект, степень напряженности проблемы, пути ее решения и др.) рассматриваются как определенным образом организованная структура[38].

Вместе с тем традиционный контент-анализ и его результаты существенно определяется заранее заданными и внешними по отношению к тексту категориями (целевыми задачами и ключевыми понятиями, задаваемыми исследователем).

2. Контекстное обучение. Для изучения контекста и контекстного знания характерным является также социально-психологический подход, наиболее распространенный в теории и практике контекстного обучения. Речь идет о проектировании и использовании обучающих социальных ситуаций и ролевых игр как форм контекстного обучения. Концептуальной основой такого подхода является теория контекстного обучения и воспитания, разработанная в научно-педагогической школе А.А. Вербицкого. Суть контекстного обучения – последовательное моделирование в формах учебной деятельности студента предметного и социального содержания его будущей профессиональной

[37] Кононова, О., Ляпин, С., Прокудин, Д. Синтетический метод извлечения контекстного знания в русскоязычной социально-гуманитарной сфере: комплексный подход. Информационное общество: образование, наука, культура и технологии будущего. 2017. Вып. 1. С. 52-67. DOI: 10.17586/2587-8557-2017-1-52-67.

[38] Манаев О.Т. Контент-анализ как метод исследования // Социология: энциклопедия. М., 2003. URL: http://psyfactor.org/lib/content-analysis3.htm.

деятельности[39]. Несмотря на практическую эффективность контекстного обучения, в рамках этой методологии и технологии само контекстное знание используется на интуитивном уровне, не формализуется и не исследуется его содержательная структура.

Кроме этого, разрабатываются и находят своё применение и другие направления в рассматриваемой области. Например, для систематического комплексного изучения контекстного знания на базе Университета ИТМО создан и функционирует виртуальный информационно-сервисный центр Humanitariana («Гуманитариана»). Его основой является многофункциональная электронная библиотека с возможностями гибкого тематизируемого полнотекстового поиска как в локальной, так и в распределенной среде T-Libra (http://77.234.221.107/tlibra). На его ресурсно-сервисной основе были реализованы пилотные проекты по изучению некоторых конкретных аспектов контекстного знания: разработка коллекции тематических полнотекстовых запросов для экспликации контекстного знания; экспликации понятийно-тематических трендов по тематике «Электронное правительство» на основе публикаций в электронных СМИ. Результаты этих и других исследований подтвердили предположение об актуальности изучения проблем контекстного знания, правомерности и эффективности использования для этих целей автоматизированных методов полнотекстового поиска. Электронная библиотека T-Libra была использована в рамках разработки комплексного подхода (синтетического метода) к анализу развития терминологической базы развивающихся междисциплинарных направлений научных исследований в распределённой сетевой среде и построения предметно-тематических трендов.

Также в исследованиях контекстного знания применяются и другие подходы:

1) технологический, направленный на разработку информационных систем и реализации в поисковых системах алгоритмов контекстного поиска;

2) семантический, связанный с разработкой и применением лингвистических подходов к анализу тестов и выявлению в них определённых смыслов;

3) содержательный, состоящий в прикладном применении алгоритмов поиска, анализа информации в функционирующих информационных системах для количественной обработки текстов и качественном анализе в них содержащихся смыслов из определённых предметных областей. Кроме этого, достаточно важным направлением исследований является анализ возможности обработки разнородных данных (применительно к текстам), распределённых в различных разнородных информационных системах с доступом как из

[39] Вербицкий А.А. Контекстное обучение в компетентностном формате (Компетентностный подход как новая образовательная парадигма) // Проблемы социально-экономического развития Сибири. 2011. № 4 (6). С. 67-73. URL: http://brstu.ru/static/unit/journal_2/docs/number6/67-73.pdf.

глобальных, так и из локальных сетей. Это особенно актуально в свете развития Grid-технологий распределения информации для возможности её систематизации и обобщения при постоянном её количественном приращении.

В последнее время наиболее динамично развиваются технологии машинного обучения, на использовании которых также строятся методы интеллектуальной обработки текстов. Однако в рамках данного учебного пособия они не рассматриваются, так как требуют отдельного рассмотрения.

Далее кратко изложены основные методы анализа контекстного знания, которые используются, в том числе, и для интеллектуального анализа научных текстов.

### 3.1.1. Контент-анализ

Контент-анализ в рамках исследования информационных потоков — достаточно новое направление, которое предусматривает анализ массива текстовых документов — результатов мониторинга информационного пространства. Понятие контент-анализа, берущего свое начало в психологии и социологии, сегодня пока не имеет однозначного определения. Это порождает ряд проблем, важнейшая из которых состоит в том, что программные системы, построенные на основе разнообразных подходов к контент-анализу, в общем случае несовместимы.

Существуют различные трактовки контент-анализа, к основным из которых относятся:

– это методика объективного качественного и систематического изучения содержания средств коммуникации (Д. Джери, Дж. Джери);

– это систематическая числовая обработка, оценка и интерпретация формы и содержания информационного источника (Д. Мангейм, Р. Рич);

– это качественно-количественный метод изучения документов, который характеризуется объективностью выводов и строгостью процедуры и представляет собой квантификационную обработку текста с дальнейшей интерпретацией результатов (В. Иванов);

– состоит из поиска в тексте определенных содержательных понятий (единиц анализа), выявления частоты их появления и соотношения с содержанием всего документа (Б. Краснов);

– это исследовательская техника для получения результатов путем анализа содержания текста о состоянии и свойствах социальной действительности (Э. Таршис).

Большинство из приведенных определений конструктивные, т.е. процедурные. Через разные начальные подходы они порождают разнообразные алгоритмы, которые временами противоречат друг другу. Существующие разнообразные подходы к пониманию контент-анализа поддаются целиком

оправданной критике. Наибольшие сомнения вызывает игнорирование роли контекста.

Большое прикладное значение методологии позволяет избежать многих противоречий. Объединение средств и методов, их естественный отбор путем многократной оценки полученных результатов открывают возможность выделения и подтверждения знаний, а также фактической силы и полезности данного инструментария.

Методы и процедуры процесса контент-анализа:

– описание проблемной ситуации, поиск цели исследования;

– точное определение объекта и предмета исследования;

– предварительный анализ объекта;

– содержательное уточнение и эмпирическая интерпретация понятий;

– описание процедур регистрации свойств и явлений;

– определение общего плана исследования;

– определение типа выборки, круга источников и т.п.

Методология контент-анализа разделяется на количественную (частота появления в документах определенных характеристик содержания) и качественную (базируется на самом факте присутствия или отсутствия в тексте одной или нескольких характеристик содержания).

Основа количественного контент-анализа:

– первый этап: выделяются единицы анализа и переводятся в форму, приемлемую для обработки (сегодня — в электронный вид);

– второй этап: подсчет частот единиц анализа с применением разнообразного математического аппарата для выявления взаимосвязей между ними;

– третий этап: интерпретация полученных результатов (при этом без привлечения искусственного интеллекта, объемных семантических формализаторов, даже экспертов как таковых, с использованием только математических методов могут быть получены содержательные, семантически наполненные результаты).

В качестве примеров можно привести автоматическое формирование дайджестов, автоматическое выявление взаимосвязи понятий (категорий), автоматическую кластеризацию взаимосвязей для выявления наиболее важных, автоматическое выявление окраски взаимосвязей, в простейшем случае – определения положительных и отрицательных взаимосвязей.

Основа качественного контент-анализа:

– в любой фазе для оценок результатов может быть привлечен эксперт, который может обнаружить определенные свойства части информации и проверить их относительно общего текстового потока, а общие свойства текстового потока распространить на его конкретную тематическую часть;

– метод призван обеспечить эксперта необходимыми средствами для выводов и дополнительных результатов.

Процесс качественного контент-анализа можно разделить на три основных стадии:

первая – сведение большого количества текстовой информации к конечному числу интегрированных блоков текста – единиц содержания, которые кодируются для дальнейшей обработки этих блоков. Основными единицами содержания являются категории, последовательности и темы;

вторая – реконструкция субъективных составных текстового потока – системы значений, мыслей, взглядов и доказательств каждого источника текста;

третья – формирование выводов и обобщений путем сравнения индивидуальных систем значений.

Одной из важнейших проблем в методологии контент-анализа является категоризация. Использование набора категорий задает концептуальную сетку, в терминах которой анализируется текстовый поток

Исследования текстового потока, если он достаточно большой, можно проводить двумя путями (необходимо отметить, что при любом из этих подходов происходит генерация новых категорий):

– определение конечной, но заведомо избыточной, совокупности категорий для получения количественных данных о встречаемости некоторых из них. При этом предполагается автоматическая или полуавтоматическая кластеризация (деление на группы и классы) неупорядоченной последовательности категорий и получение на ее основе новых обобщенных категорий;

– выявление в потоке с помощью количественных многоразовых оценок новых знаний с последующей квалификацией их как категорий. Это направление контент-анализа получило название Data Mining — дословно «раскопка данных».

### 3.1.2. Контент-мониторинг информационных потоков

Этот метод состоит в постоянном выполнении узко очерченного своими задачами контент-анализа беспрерывных информационных потоков. Контент-мониторинг имеет собственную проблематику и собственные пути решения прикладных задач, контент-анализ выступает здесь как составная часть.

Методы контент-мониторинга как эволюция идеологии контент-анализа получили значительное развитие на территории бывшего СССР. Так, наиболее интересными сегодня являются проекты М. Г. Крейнеса "Ключи от текста", Д. А. Поспелова "Интерактивное выявление семантических структур текста", проект "Оружие аналитика" компании "Инвента", проект ВААЛ и прочие.

Уникальность предложенной технологии состоит в объединении содержательных и количественных методов контент-анализа. Последовательность этапов содержательного анализа проблемы, которая исследуется конкретной информационной системой, условно можно поделить на

содержательный (качественный) анализ совокупности публикаций и формализованный (количественный) анализ информационных массивов: индексного, библиографического и массива текстов ключевых фрагментов публикаций.

К особенностям автоматизированной технологии контент-мониторинга относятся:

– использование ключевого фрагмента публикации как единицы формирования текстового информационного массива;

– формирование банка ключевых фрагментов публикаций является объединением двух взаимосвязанных автоматизированных процессов: аналитико-синтетической переработки и многоуровневой процедуры контент-анализа текстов публикаций;

– индексация ключевых фрагментов публикаций происходит при помощи многофасетной классификации.

### 3.1.3. Контент-анализ и проект «Контекстное знание»

Изучение структурно-организованных, автоматически извлекаемых из неструктурированных источников информации смысловых и тематических контекстов важно для задач использования информационных технологий в сферах науки и образования, управления и бизнеса. Контекстное знание (т.е. знание, содержащееся в различных контекстах, полученных в результате полнотекстовых запросов) может существовать в различных видах и формах; может быть извлечено и изучено в результате полнотекстового поиска и последующей его обработки. Как в русскоязычном, так и в мировом научном дискурсе изучение контекстного знания ведется на основе теории и практики контент-анализа.

Исследование контекстов различного типа, вида и уровня проводятся в рамках традиционного контент-анализа: метода и технологии качественно-количественного анализа документов в целях выявления или измерения социальных фактов и тенденций, отраженных этими документами. Контент-анализ изучает документы в их социальном контексте. Все более широко распространяется контент-анализ сообщений средств массовой информации, основанный на парадигматическом подходе, в соответствии с которым изучаемые признаки текстов (содержание проблемы, причины ее возникновения, проблемообразующий субъект, степень напряженности проблемы, пути ее решения и др.) рассматриваются как определенным образом организованная структура[40]. Вместе с тем традиционный контент-анализ и его результаты существенно определяется заранее заданными и внешними по отношению к тексту категориями (целевыми задачами и ключевыми понятиями, задаваемыми исследователем).

[40] Манаев О.Т. Контент-анализ как метод исследования // http://psyfactor.org/lib/content-analysis3.htm.

В традиционном контент-анализе первичными являются целевая функция и категории анализа, вторичными – получаемые «обобщенно-текстовые» единицы анализа; в подходе, характерном для настоящего проекта, первичен «обобщенный текст» (с элементами мультимодальной информации), вторичен получаемый «контент», т.е. структурированное описание контекстуального знания. Можно сказать, что традиционный контент-анализ и предлагаемый «нетрадиционный» анализ контекстного знания являются дополнительными друг другу методами и технологиями изучения содержательных и смысловых информационных контекстов. «Обобщенный текст» (текст + мультимодальная информация) в данном случае является генератором эксплицируемых контекстов и, соответственно, структур контекстного знания. Инструментом генерации являются гибкие функциональные структуры мультимодальных запросов.

### 3.1.4. Тематическое моделирование

Под тематическим моделированием понимается технология статистического анализа текстов для автоматического выявления тематики в больших коллекциях документов. Тематическое моделирование – это способ построения модели коллекции текстовых документов, которая определяет, к каким темам относится каждый из документов. Переход из пространства терминов в пространство найденных тематик помогает разрешать синонимию и полисемию терминов[41]. Метод также дает понимание того, какими словами описывается каждая тема (семантическая группа).

Тематическое моделирование не требует, как правило, ручной разметки текстов, обучение модели происходит без учителя. Построение тематической модели может рассматриваться как задача одновременной кластеризации и слов, и документов по их семантической близости. Кластеризация в тематическом моделировании не является строгой (нечеткая кластеризация), допускается отнесение документов сразу к нескольким темам (семантическим группам), составляющих разные кластеры. Тематическое моделирование не претендует на понимание смысла текста, однако оно способно помочь ответить на вопросы о содержании текста или выявить общие темы/контексты в нескольких текстах.

Задачи, решаемые тематическим моделированием:

– тематический поиск в электронных библиотеках – поиск по смыслу (контексту), а не по ключевым словам;

– отслеживание событий в новостных потоках;

– выявление тематических сообществ и тональности сообщений в социальных сетях;

[41] Глушков, Н. А. Анализ методов тематического моделирования текстов на естественном языке / Н. А. Глушков. — Текст: непосредственный // Молодой ученый. — 2018. — № 19 (205). — С. 101-103. — URL: https://moluch.ru/archive/205/50247.

– построение профилей интересов пользователей в рекомендательных системах;

– фильтрация спама;

– управление диалогом в системах разговорного интеллекта с возможностями выявления и отнесения к категории планов и интересов собеседника;

– поиск изображений по тексту и текстов по изображениям;

– поиск повторяющихся последовательностей элементов в данных, продуцируемых естественнонаучными экспериментами;

– аннотирование или рубрикация данных нетекстовой модальности;

– поиск аномального поведения объектов в видеопотоке;

– выявления поведенческих паттернов по транзакционным данным.

Перечисленные задачи находят применение в различных предметных областях.

При реализации тематического моделирования используются различные методы и подходы. Методы тематического моделирования можно разделить на две основных группы — алгебраические и вероятностные (генеративные).

К алгебраическим моделям относятся подходы – стандартная векторная модель текста (Vector Space Model, VSM) и латентно-семантический анализ (Latent Semantic Analysis, LSA), которые позволяют учитывать авторство, связи между словами, темами, документами и авторами, ряд других типов сущностей и метаданных.

Векторная модель текстов (VSM) является способом представления коллекций документов в виде векторов из общего для всей коллекции векторного пространства. VSM представляет документы и запросы как векторы весов. Каждый вес – это мера важности термина в документе или в запросе. Веса терминов вычисляются и ранжируются на основе частоты использования терминов в документе, запросе или коллекции. Данная модель используется для решения множества задач быстрого анализа документов; составления таблиц поиска, классификации и кластеризации. Также она используется в качестве основы для множества других алгоритмов. В данной модели документ рассматривается как неупорядоченное множество термов — слов и дополнительных элементов, из которых состоит текст (из этого множества элементов исключаются знаки препинания). Для каждого документа строится матрица терм-документ, где строка — это уникальное слово, а столбец — документ. Значением ячейки данной матрицы является вес данного слова в документе, способ вычисления которого может изменяться в зависимости от алгоритма.

VSM применяется для решения задач информационного поиска, классификации и кластеризация документов.

Также эта модель достаточно популярна для решения задач сравнения текстов между собой, однако в изначальном варианте не отличается достаточной быстротой для обработки больших объемов документов и занимает достаточно большой объём памяти. Развитием данного метода является латентно-семантический анализ (LSA).

LSA осуществляет определение слов, из которых состоят темы, а также определение тем, к которым принадлежит документ, с помощью применения к матрице «Слова-на-Документы» сингулярного матричного разложения SVD (Singular Value Decomposition). Основные недостатки LSA-модели – в значительном снижении скорости вычислений при увеличении объема входных данных, а также трудностями интерпретация результатов, вызванных использованием матричного разложения SVD.

Вероятностная тематическая модель представляет собою модель коллекции текстовых документов, которая определяет темы для каждого документа на основе специфических для этой темы терминов. Математическая модель позволяет, основываясь на статистических данных, предположить темы, к которым относится документ, а также соотношение этих тем в документе.

Основные предположения базовых вероятностных тематических моделей:

1) не важны для определения тематики документа

– порядок документов в корпусе,

– порядок слов в документе, документ рассматривается как «мешок слов»,

– слова общей лексики;

2) слово в разных грамматических формах считается одним и тем же словом.

К вероятностным методам относят:

– pLSA (1999) – вероятностный латентный семантический анализ (probabilistic LSA), более простая модель, предшественник модели LDA. Преимуществом pLSA модели по сравнению с предыдущей моделью LSA является базирование на статистических методах и, как следствие, лучшая применимость на практике. Основные недостатки pLSA – невозможность управлять разреженностью получаемых на выходе матриц вероятностей и линейный рост числа параметров pLSA с ростом числа документов в корпусе, что приводит к переобучению модели.

– LDA – (2003) – латентное размещение Дирихле (Latent Dirichlet Allocation) – применяемая в информационном поиске порождающая модель, позволяющая объяснить результаты наблюдений с помощью неявных групп. Модель принадлежит семейству порождающих вероятностных моделей, в которых темы представлены вероятностями появления каждого слова из заданного набора. Документы в свою очередь могут быть представлены как сочетания тем. Уникальность модели состоит в том, что темы не обязательно должны содержать уникальный набор терминов, т.е. одинаковые слова могут

встречаться в разных темах[42]. Тематические модели класса LDA (Latent Dirichlet Allocation) предполагают существование конечного множества скрытых тем T, и каждая коллекция документов D порождается дискретным распределением:

$$p\ (d, w, t) \qquad (а)$$

d – документ, $d \in D$; w – слово, $w \in Wd$ (наблюдаемая переменная в коллекции документов); t – тема, $t \in T$ (скрытая, латентная переменная). Коллекция документов Wd – это случайная, однородная и независимая выборка пар документ-слово – (d, w). Каждая пара (d, w) связана с некоторой неизвестной темой t. Построить тематическую модель корпуса документов означает найти множество скрытых тем T и определить условное распределение для каждой входящей в множество темы t:

$$p\ (w|\ t) \equiv \varphi\ (w, t) \qquad (б)$$

и для каждого документа d:

$$p\ (t \mid d) \equiv \theta\ (t, d) \qquad (в).$$

*Результаты моделирования*: матрица $\varphi$ (w, t) – распределение слов по темам; матрица $\theta$ (t, d) – распределение документов по темам. В ячейках матриц находятся вероятности принадлежности слов/документов к теме.

LDA – наиболее часто используемая тематическая модель. Недостаток модели в том, что хотя задача тематического моделирования имеет бесконечно много решений, LDA не даёт возможности определить лучшее решение. Это приходится делать на основе экспертной оценки по итогам анализа проведённой экспериментальной части. Основные метрики, применяемые для оценки тематической модели, представлены в таб. 3.1.

Таб. 3.1. Оценка тематической модели

| Метрика | Значение |
|---|---|
| Разреженность | Доля вероятностей, близких к нулю в матрицах $\varphi\ (w, t)$ и $\theta\ (t, d)$ |
| Чистота и контрастность | Оценка различности тем |
| Когерентность | Доля фоновых слов, отражает степень интерпретируемости темы; если доля велика, это может свидетельствовать о вырожденности модели |
| Перплексия | Ожидаемый размер словаря с равномерным распределением слов, который необходим модели чтобы сгенерировать слово из тестовой выборки. Рассчитывается через логарифм правдоподобия. |

– ARTM (2014) – теория аддитивной регуляризации является альтернативой байесовскому подходу и графическим моделям, которые использовались в PLSA

[42] Коршунов А., Гомзин А. Тематическое моделирование текстов на естественном языке // Труды Института системного программирования РАН. 2012. С. 215-242. https://elibrary.ru/item.asp?id=18361454.

и LDA. ARTM позволяет собирать модели из готовых модулей в стиле LEGO-конструктора, не требует при создании новой модели производить заново математические выкладки и программную реализацию. ARTM основан на идее многокритериальной регуляризации. Регуляризация позволяет улучшить качество классификации текстов, учесть нетекстовые данные, повышать точность и полноту поиска и т.д. заданием критерия. ARTM позволяет задавать несколько критериев-регулязаторов одновременно, совмещая регуляризаторы из разных моделей и тем самым создавая комбинированные модели с заранее заданными свойствами. Модульная технологии тематического моделирования с возможностью повторного использования кода реализована в инструменте BigARTM.

Тематическая модель, которая строится не только по словам, но и по любым элементам другой модальности, называется мультимодальной. Так, мультимодальная тематическая модель для текстов и изображений позволяет автоматически выделять темы из изображений на основе их описания и в дальнейшем использовать ключевые слова тем для описания новых изображений. На вход такой модели поступают набор пар – изображение и его описание. Мультимодальная тематическая модель находит применение в таких задачах, как аннотирование изображений или поиск тематических иллюстраций для текста. Для разных модальностей вероятностные распределения над терминами должны строиться раздельно.

– CorrLDA (Correspondence LDA) – модель на основе LDA для автоматической аннотации изображений. В данной модели изображения сначала сегментируются, затем для каждого сегмента извлекается вектор признаков, которыми выступают размер, позиция, цвет, текстура и форма, представленные в виде действительных значений. После этого формируются пары (вектор признаков изображения, набор слов описания).

– MixLDA и sLDA – модели на основе LDA для автоматической аннотации изображений. Для извлечения признаков изображений используется алгоритм SIFT. Разница между MixLDA и sLDA – в использовании алгоритмов извлечения тем.

– BigARTM (2015) – модель, расширение ARTM, являющаяся мультимодальным вариантом ARTM с открытой библиотекой. Необходимо учитывать, что пространство изображений вычислительно неперечислимо. Поэтому любое построенное над ним вероятностное распределение будет характеризовать только ту коллекцию, по которой оно было построено. Данное ограничение накладывает границы применимости мультимодальной ARTM к задачам тематического моделирования.

Инструменты, появившиеся в последние годы, такие как технологии глубокого обучения для распознавания образов и модель векторного представления слов для обработки естественного языка, позволяют учитывать

контекст. Применение этих технологий в научно-исследовательской работе может повысить качество построения тематических моделей для мультимодальных документов. Однако это требует специального изучения возможностей такого инструментария.

Перечислим основные этапы построения тематической модели:

1. *Препроцессинг данных*:

– Разбиение длинных строк текста на более мелкие части, токены (токенизация). К токенам относятся как слова, так и знаки препинания.

– Удаление стоп-слов, не несущих смысловой нагрузки. Перечень стоп-слов должен быть качественно составлен, чтобы исключить слова, которые не относятся к тематике текста.

– Выявление устойчивых словосочетаний (биграмма) – коллокаций;

– Приведение слов к смысловой канонической форме (лемматизация, стемминг);

– Подсчет частот слов (лемм), удаление слишком частых и слишком редких слов, не имеющих дискриминирующей силы.

2. Формирование словаря, состоящего из всех слов, которые встречались во всех описаниях и не были отброшены на предыдущем шаге. Каждому слову присваивается идентификатор — номер, под которым он записан в словаре.

3. Построение модели – определение количества тем (семантических групп), где каждая тема представляет собой комбинацию ключевых слов, и каждое слово вносит определенный вес в тему, и на основании значений когерентности и перплексии определяется оптимальное количество тем.

Для наглядности приведём алгоритм построения мультимодальной тематической модели.

Для построения мультимодальных тематических моделей (например, для распознавания изображений как одного из видов контекстного знания) используются нейронные сети, представляющие собой математическую модель, основанную на знаниях, полученных при изучении нейронных связей в мозге. Наиболее действенными при решении подобного рода задач являются сверточные нейронные сети (Convolutional Neural Network, CNN)[43], для которых сделано явное предположение, что на вход подается изображение.

Вначале необходимо «обучить» модель. Для этого на вход подаётся корпус, состоящий из изображений и подходящих к ним описаний. На описание не накладываются ограничения по длине и числу описаний, относящихся к одному изображению.

После обучения модель возвращает матрицу Φ, описывающую распределение векторов слов на темы, и матрицу Θ, описывающую распределение тем на образы изображений.

[43] Krizhevsky, A., I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolu- tional neural networks // Advances in Neural Information Processing Systems. 2012. P. 1097–1105.

Имея разложение «матрицы векторных представлений слов на векторы образов изображений» на матрицы Ф и Θ, можно относительно легко генерировать описания изображений.

На вход модели подается изображение, размер которого необходимо оптимизировать. После подачи этого изображения на вход CNN выдает вектор признаков данного изображения. Затем модель начинает поиск вектора, ближайшего к тому, который был подан на вход, среди всех векторов, известных модели. Для найденного вектора из матрицы Θ извлекается распределение тем, полученное во время создания модели.

Алгоритм нахождения слов, подходящих к описанию изображения, приведён на рис. 3.1.


Рис. 3.1. Схема алгоритма генерации аннотаций по изображению (Смелик Н. Д., Фильченков А. А)

В результате по найденным темам с помощью матрицы Ф извлекаются векторы слов, которые наилучшим образом характеризуют контекст данной темы. С помощью этих векторов можно найти слова, которые чаще всего употребляются в этом контекст, что как раз и представляет собой описание поданного на вход модели изображения.

Аналогичным образом решается задача поиска изображений по текстам, имея разложения матрицы F на матрицы Ф и Θ. Алгоритм нахождения изображений по текстовому описанию приведён на рис. 3.2.

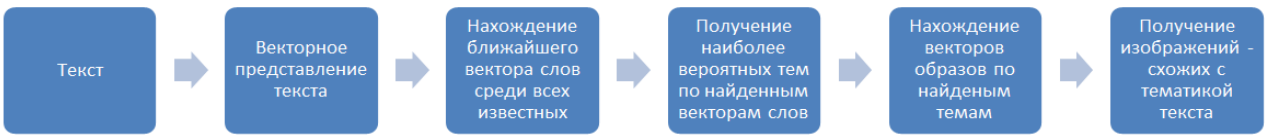



Рис. 3.2. Схема алгоритма поиска изображений по текстам (Смелик Н. Д., Фильченков А. А)

Теперь на вход модели подается текст, который предварительно обрабатывается аналогично предыдущему случаю. Затем для каждого термина в тексте находится его векторное представление. Далее текст представляется в виде нового документа, для которого нужно найти распределение тем, т. е. получить вектор θ распределений тем для нового документа. Это можно сделать с помощью алгоритма для обучения модели исходя из того, что матрица Ф уже известна.

После завершения распределения тем по текстовому документу для каждой темы с вероятностью в документе, отличной от нуля, производится нахождение

изображения, максимально соответствующего данной теме. Таким образом формируется набор изображений, представляющий собой иллюстрации к тексту[44].

В заключение перечислим основные инструменты, реализующие тематическое моделирование.

**Mallet** – это пакет для статистической обработки естественного языка, кластеризации, извлечения информации и других приложений машинного обучения для обработки текста. Инструментарий для моделирования тем Mallet содержит эффективные реализации на основе выборок скрытого распределения Дирихле, распределение Патинко и иерархического LDA.

**Gensim** – это пакет, в котором реализованы основные алгоритмы тематического моделирования: скрытое распределение Дирихле (LDA) и скрытое семантическое индексирование (LSI). Пакет показывает высокие скорость обработки, качество результата и оценки тематических моделей. Существенное преимущество Gensim – пакет позволяет обрабатывать большие текстовые файлы, не загружая весь файл в память.

**pyLDAvis** – построение интерактивной диаграммы модели LDA (круги представляют собой тему. Чем больше круг, тем тема считается распространенной).

**BigARTM**[45] – библиотека с открытым кодом для тематического моделирования больших текстовых коллекций и массивов транзакционных данных. BigARTM реализует несколько механизмов, которые снимают многие ограничения простых моделей типа PLSA или LDA и расширяют спектр приложений тематического моделирования. Приставка «big» в названии означает, что реализация модульной технологии ARTM позволяет эффективно обрабатывать большие данные. Основные особенности BigARTM:

– Регуляризаторы (Regularization), которые можно комбинировать в любых сочетаниях.

– Модальности (Modalities), которыми можно описывать нетекстовые объекты внутри документов.

– Тематические иерархии (Hierarch), в которых темы разделяются на подтемы.

– Использование данных о совместной встречаемости слов (Co-occurrence).

---

[44] Смелик Н. Д., Фильченков А.А. Мультимодальная тематическая модель текстов и изображений на основе использования их векторного представления // Машинное обучение и анализ данных, 2016. Т. 2. № 4. С. 421-441. doi:10.21469/22233792.2.4.05

[45] Документация по BigARTM – bigartm.org.
Теория (на русском языке): www.MachineLearning.ru/wiki/images/d/d5/Voron17survey-artm.pdf, (на английском): fruct.org/publications/fruct21/files/Koc.pdf.

– Intratext. Внутритекстовые регуляризаторы, обрабатывающие текст как последовательность тематических векторов слов.

– Тематизация транзакционных данных (Transaction).

**Rutermextract** – это библиотека для извлечения ключевых слов из текстов на русском языке.

**TopicMiner** – это профессиональный пакет тематического моделирования и визуального анализа для русского языка (http://linis.hse.ru/soft-linis).

**Scikit-learn** – пакет Python с открытым исходным кодом для Data Science и Machine Learning поддерживает: предварительную обработку данных, уменьшение размерности, выбор модели; решение задач: регрессия, классификация, кластерный анализ. Scikit-learn основан на библиотеках NumPy и SciPy.

### 3.1.5. Наукометрические методы и системы в анализе научного знания

Всё более широкое применение в различных научных исследованиях находят наукометрические и библиометрические методы. Это, прежде всего, связано с тем, что научная информация (научные публикации) представлена в цифровой форме и сконцентрирована в различных цифровых информационных ресурсах с доступом из сети Интернет (см. главу 1). Эти ресурсы представляют собой большие базы данных (Big Data) с мощным поисковым механизмом. Как правило, информация в этих базах представлена в виде сведений о научных публикациях (метаданные): название, аннотация, ключевые слова, авторы, аффилиация авторов, списки пристатейной литературы, сведения о цитировании и т.д. Поэтому наукометрические и библиометрические методы связаны с обработкой и анализом такого вида контекстного знания, как метаданные. Анализ метаданных позволяет решать такие задачи, как:

– выявление терминологии предметных научных областей;

– выявление тематик, входящих в междисциплинарные научные направления;

– определение важности отдельных исследований для развития науки;

– выявление различных концепций научных исследований в рамках определённого научного направления;

– географическая кластеризация разработки научных направлений исследований;

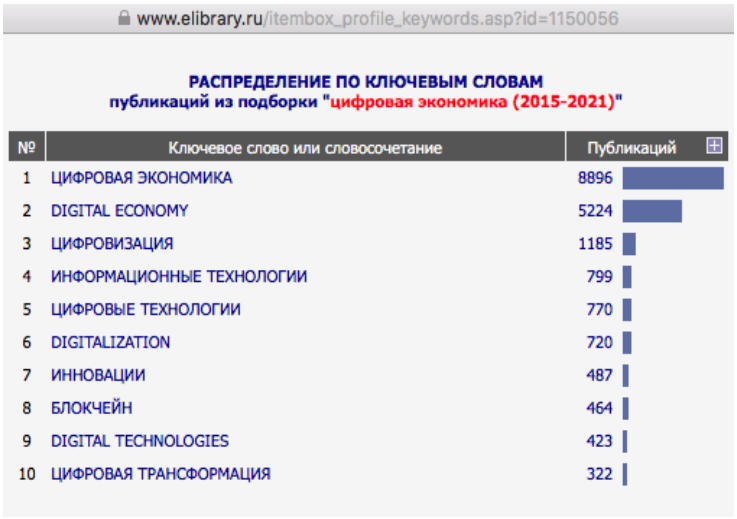
– определение динамики развития различных научных направлений.

Как правило, обработка метаданных производится с использованием статистических и количественных методов. Во многом эти методы реализованы в информационных системах или программных продуктах. Существует два основных пути использования наукометрических и библиометрических методов для анализа метаданных:

1) использование встроенного в цифровые научные ресурсы аналитического инструментария;

2) выгрузка (экспорт) метаданных из цифровых научных ресурсов и последующая их обработка с применением специализированного программного обеспечения.

Например, аналитическими инструментами обладают НЭБ Elibrary и реферативная база научных публикаций Scopus. По отобранным массивам документов в НЭБ Elibrary можно получить аналитический отчёт распределения их по тематике (рис. 3.3) или по ключевым словам (рис. 3.4), а также по организациям авторов (рис.3.5). Аналогичные аналитические отчёты можно сформировать в цифровом ресурсе Scopus (рис. 3.6 – 3.8).

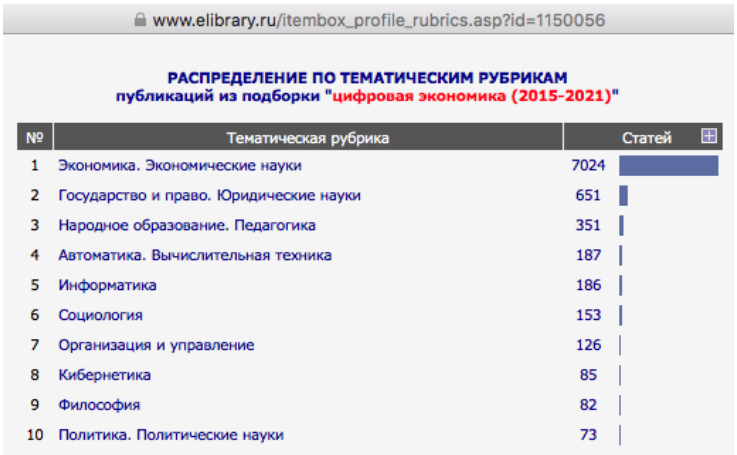


www.elibrary.ru/itembox_profile_keywords.asp?id=1150056

**РАСПРЕДЕЛЕНИЕ ПО КЛЮЧЕВЫМ СЛОВАМ**
**публикаций из подборки "цифровая экономика (2015-2021)"**

| № | Ключевое слово или словосочетание | Публикаций |
|---|---|---|
| 1 | ЦИФРОВАЯ ЭКОНОМИКА | 8896 |
| 2 | DIGITAL ECONOMY | 5224 |
| 3 | ЦИФРОВИЗАЦИЯ | 1185 |
| 4 | ИНФОРМАЦИОННЫЕ ТЕХНОЛОГИИ | 799 |
| 5 | ЦИФРОВЫЕ ТЕХНОЛОГИИ | 770 |
| 6 | DIGITALIZATION | 720 |
| 7 | ИННОВАЦИИ | 487 |
| 8 | БЛОКЧЕЙН | 464 |
| 9 | DIGITAL TECHNOLOGIES | 423 |
| 10 | ЦИФРОВАЯ ТРАНСФОРМАЦИЯ | 322 |

Рис. 3.3. Аналитический отчёт о распределении публикаций по ключевым словам публикаций, запрос «цифровая экономика» за 2015-2021 годы (НЭБ Elibrary)



www.elibrary.ru/itembox_profile_rubrics.asp?id=1150056

**РАСПРЕДЕЛЕНИЕ ПО ТЕМАТИЧЕСКИМ РУБРИКАМ**
**публикаций из подборки "цифровая экономика (2015-2021)"**

| № | Тематическая рубрика | Статей |
|---|---|---|
| 1 | Экономика. Экономические науки | 7024 |
| 2 | Государство и право. Юридические науки | 651 |
| 3 | Народное образование. Педагогика | 351 |
| 4 | Автоматика. Вычислительная техника | 187 |
| 5 | Информатика | 186 |
| 6 | Социология | 153 |
| 7 | Организация и управление | 126 |
| 8 | Кибернетика | 85 |
| 9 | Философия | 82 |
| 10 | Политика. Политические науки | 73 |

Рис. 3.4. Аналитический отчёт о распределении по тематическим направлениям публикаций, запрос «цифровая экономика» за 2015-2021 годы (НЭБ Elibrary)

www.elibrary.ru/itembox_profile_orgs.asp?id=1150056

**РАСПРЕДЕЛЕНИЕ ПО ОРГАНИЗАЦИЯМ**
**публикаций из подборки "цифровая экономика (2015-2021)"**

| № | Название организации | Статей |
|---|---|---|
| 1 | Финансовый университет при Правительстве РФ | 841 |
| 2 | Российский экономический университет им. Г.В. Плеханова | 655 |
| 3 | Российская академия народного хозяйства и государственной службы при Президенте РФ | 305 |
| 4 | Санкт-Петербургский государственный экономический университет | 289 |
| 5 | Московский государственный университет им. М.В. Ломоносова | 233 |
| 6 | Крымский федеральный университет им. В.И. Вернадского | 197 |
| 7 | Государственный университет управления | 188 |
| 8 | Санкт-Петербургский политехнический университет Петра Великого | 148 |
| 9 | Санкт-Петербургский государственный университет | 118 |
| 10 | Самарский государственный экономический университет | 114 |

Рис. 3.5. Аналитический отчёт о распределении по организациям авторов публикаций, полученных по запросу «цифровая экономика» за 2015-2021 годы (НЭБ eLibrary)

TITLE-ABS-KEY ( "digital economy" ) AND PUBYEAR > 2014

3 547 результатов поиска документов

| Страна | Документы |
|---|---|
| Russian Federation | 1745 |
| United States | 265 |
| United Kingdom | 194 |
| China | 152 |
| Australia | 104 |
| Ukraine | 101 |
| Germany | 90 |
| Italy | 78 |
| India | 74 |

Рис. 3.6. Аналитический отчёт о распределении по странам авторов публикаций, запрос «digital economy» за 2015-2021 годы (ИС Scopus)


Рис. 3.7. Аналитический отчёт о распределении по организациям авторов публикаций, запрос «digital economy» за 2015-2021 годы (ИС Scopus)

Также для анализа метаданных можно воспользоваться специализированным программным обеспечением (например, VosViewer или CitNetExplorer). Так, VosViewer (https://www.vosviewer.com) предназначен для построения и визуализации карт сетей данных библиометрической информации. Такие сети могут включать журналы, авторов-исследователей или отдельные публикации. Каждый из таких узлов сетей может опираться на цитирование другого узла (а для авторов – может быть отношение соавторства), что называется библиографической связью. Карты поддерживают технологии выделения кластеров.


Рис. 3.8. Аналитический отчёт о распределении по тематическим направлениям исследований публикаций, запрос «digital economy» за 2015-2021 годы (ИС Scopus)

В качестве данных могут использоваться либо экспортированные файлы метаданных (например, Web of Science, Scopus, Dimensions, PubMed), либо метаданные, полученные по API (Crossref, Europe PMC, Microsoft Academic, Semantic Scholar, OpenCitations, WikiData). VosViewer предлагает функции интеллектуального анализа текста, который может визуализировать совместное использование важных терминов – ключевых слов, извлеченных из массива научной литературы и основанных на их одновременном появлении в сети. На рис. 3.10 и рис. 3.11 приведены примеры визуализации в VosViewer.

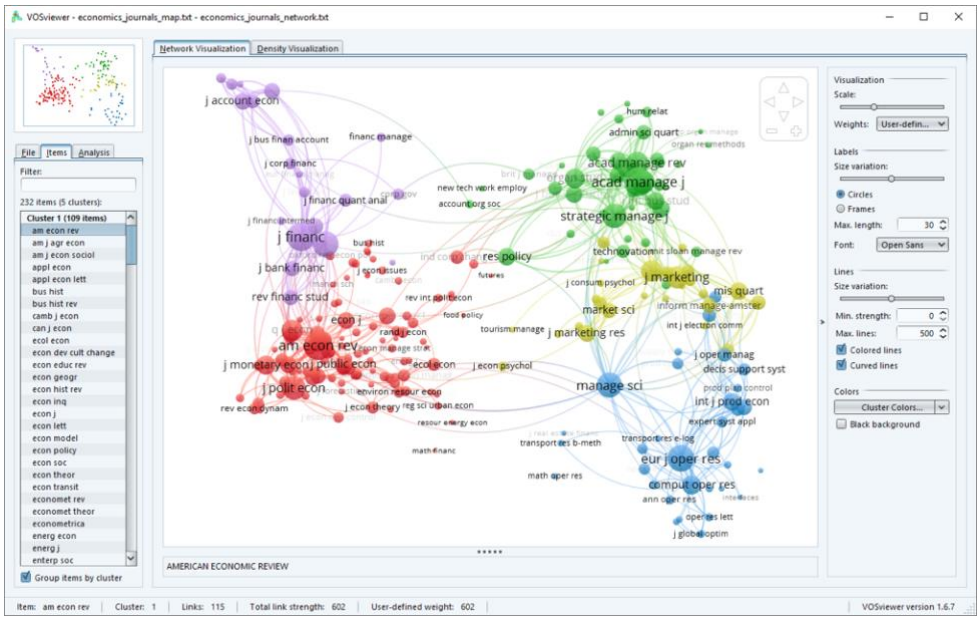



Рис. 3.10. Пример визуализации данных в виде сети (VosViewer)


Рис. 3.11. Пример визуализации данных в виде тепловой карты

## 3.2. Синтетический метод в анализе научного знания

### 3.2.1. Синтетический метод как комплексный подход к поиску, экспликации и анализу научного знания[46]

Для поиска и выделения контекстов в междисциплинарных исследованиях, анализа и экспликации контекстов, построения и интерпретации трендов предлагается использовать общий авторский подход, названный «синтетическим методом». Данный подход является комплексным и основывается на применении ИКТ на всех этапах проведения исследования по анализу научного знания – от выбора цифровых электронных ресурсов и поиска научных текстов до экспликации контекстного знания и его анализа.

Являясь по сути одной из разновидностей контент-анализа, синтетический метод выводит его на иной уровень. Контент-анализ – выявление и экспликация контекстов с целью анализа заключённых в контекстах смыслов. В традиционном контент-анализе первичными являются целевая функция и категории анализа, вторичными – получаемые обобщенно-текстовые единицы анализа. Для синтетического метода первичен обобщенный текст, вторичен получаемый контент – структурированное описание контекстуального знания. В контент-анализе поиск нужных единиц анализа делают люди (эксперты), между которыми не должно быть разногласий в толковании терминов. Использование синтетического метода для изучения содержательных и смысловых информационных контекстов ускоряет работу на данном этапе, но требует последующей экспертизы массива терминов с целью отбрасывания информационного шума (не соответствующих теме единиц анализа) и выделения необходимых смысловых групп. Обобщенный текст в этом случае является генератором эксплицируемых контекстов и структур контекстного знания. На основе выявления опорных термин-концептов и последующей итерации запросов происходит формирование релевантных рассматриваемой тематике текстовых массивов. Методика позволяет с использованием каскадного (результаты одного запроса автоматически входят в поисковый образ другого запроса) и тезаурусного (используется для автоматического расширения культурного контекста в ходе выполнения запроса) поиска, сочетания многослойных тематических абзацно-ориентированных (с вариацией используемых слоев) и частотно-ориентированных запросов, экспликации, кластеризации и статистической обработки научных текстов формировать коллекции тематически релевантных фрагментов (тематических контекстов), выявить контекст

[46] Кононова, О.В., Прокудин, Д.Е. Геймификация и социокультурные контексты цифрового урбанизма: общий подход к анализу и прогнозу научного дискурса [электронный текст] // Культура и технологии. 2019. Том 4. Вып. 3. Серьезные компьютерные игры: где, как и почему они работают. С. 83-105. URL: http://cat.ifmo.ru/ru/2019/v4-i3/193; Кононова, О., Ляпин, С., Прокудин, Д. Синтетический метод извлечения контекстного знания в русскоязычной социально-гуманитарной сфере: комплексный подход. Информационное общество: образование, наука, культура и технологии будущего. 2017. Вып. 1. С. 52-67. DOI: 10.17586/2587-8557-2017-1-52-67.

использования термин-концептов и ключевых слов, соотнести полученные контексты с областями знаний. Выявление тематических направлений исследуемых междисциплинарных областей позволяет распределить термин-концепты по семантическим группам.

Применение синтетического метода для анализа тематики и используемой терминологии научных публикаций и СМИ позволяет:

– выделить тематики научных исследований, составляющих междисциплинарные научные направления;

– сформировать терминологическую базу проводимого исследования или тезаурус междисциплинарного научного направления;

– сравнить локальные и мировые тенденции развития научных исследований;

– выявить научно значимые результаты, полученные в ходе проведения изучаемых исследований;

– сформировать тематические коллекции контекстов (абзацев), релевантных тематике проводимого исследования;

– выявить траектории развития перспективных научных направлений (построение предметно-тематических трендов).

Характерной особенностью синтетического метода является использование принципа доступности широкому кругу исследователей программных сред и систем, являющихся инструментами исследования:

– для их осваивания и использования не требуется специальных знаний и умений в области информатики и информационных технологий;

– их установка и эксплуатация не связана с дополнительными технологическими решениями;

– они не являются коммерческими продуктами (это либо свободно распространяемое ПО, либо открытые сетевые ресурсы), или используемый функционал коммерческих продуктов является открытым или условно бесплатным;

– коммерческие информационные базы (например, ScienceDirect, Web of Science, Scopus и другие) доступны подавляющему большинству отечественных исследователей, аффилированных с научно-образовательными учреждениями, имеющими на них подписку.

Такой подход позволяет применять предлагаемый метод максимально широким кругом исследователей по различным междисциплинарным научным направлениям.

Для демонстрации общего подхода к анализу междисциплинарных направлений научных исследований (МДННИ), применения синтетического метода и вложенных в него методик анализа и экспликации контекстов, построения и анализа трендов в качестве ресурса текстов взята Научная

электронная библиотека eLibrary (НЭБ Elibrary, http://elibrary.ru), а в качестве инструментальной среды – электронная библиотека T-Libra (рис. 3.12).

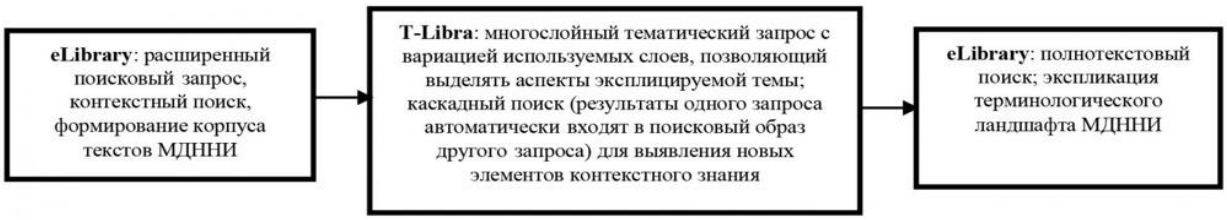



Рис. 3.12. Три фазы экспликации и анализа контекстного знания

Первая фаза общего подхода подразумевает формирование поисковых запросов, терминологического ядра и подборок текстовых материалов, релевантных междисциплинарному направлению научных исследований. Блок-схема, описывающая процедуру поиска литературы в НЭБ eLibrary, представлена на рис. 3.13.

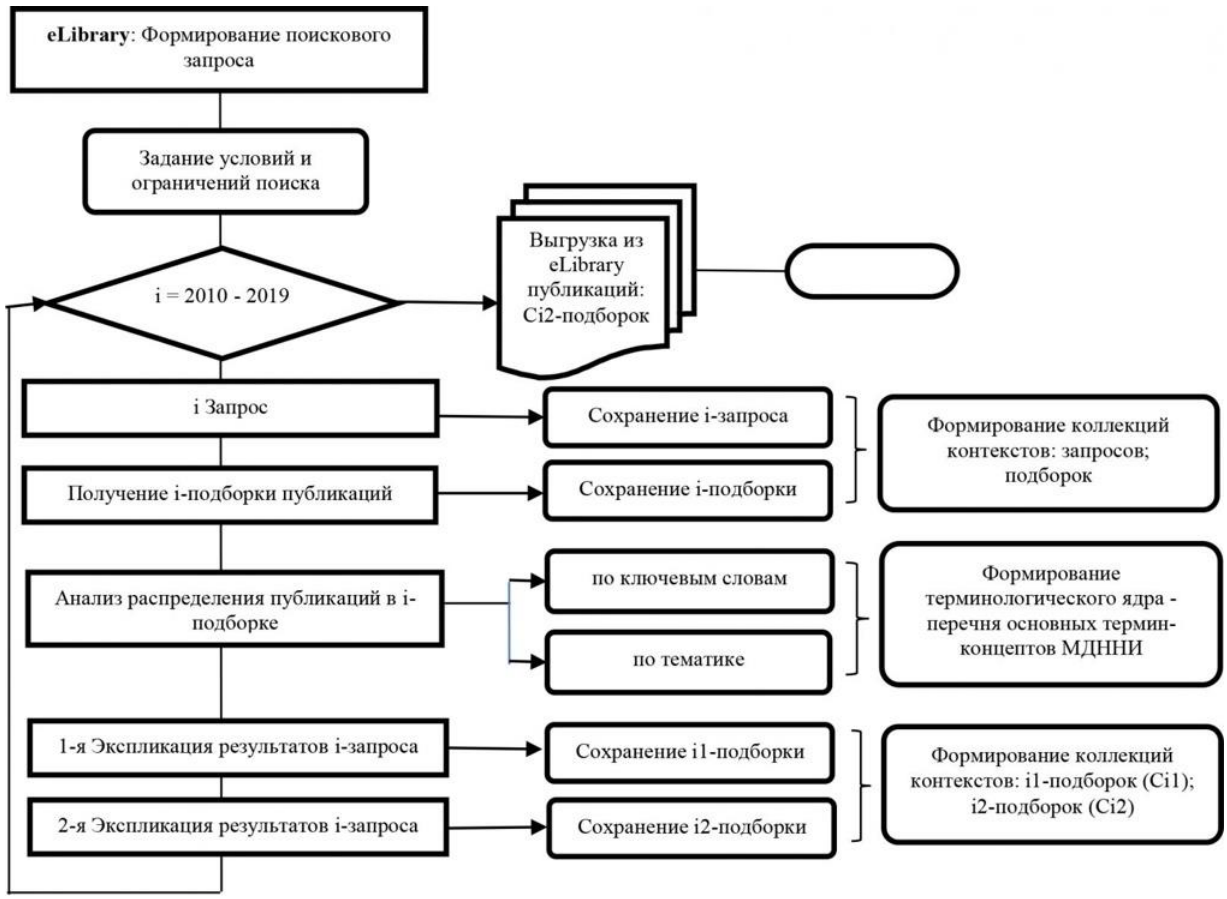



Рис. 3.13. Блок-схема, описывающая алгоритм поиска литературы в цифровом полнотекстовом ресурсе

Тематические коллекции формируются путём выполнения последовательной итерации поисковых запросов. При этом используются различные аналитические инструменты информационной системы НЭБ Elibrary.

Вторая (T-Libra) и третья (eLibrary) фазы общего подхода в целом отражают последовательность шагов от выделения и экспликации первичных контекстов до построения предметно-тематических трендов в результате проведения повторных тематических запросов в ранее эксплицированных подборках текстовых материалов (Рис. 3.14).

Параллельно основному процессу происходит выявление и формирование качественно нового вида контекстного знания (контекстов) – тематических коллекций полнотекстовых запросов, которые одновременно могут использоваться и как готовое тематизированное знание, расширяющее состав информационных ресурсов электронной библиотеки, и как пользовательский инструмент для создания и развития аналогичных коллекций.

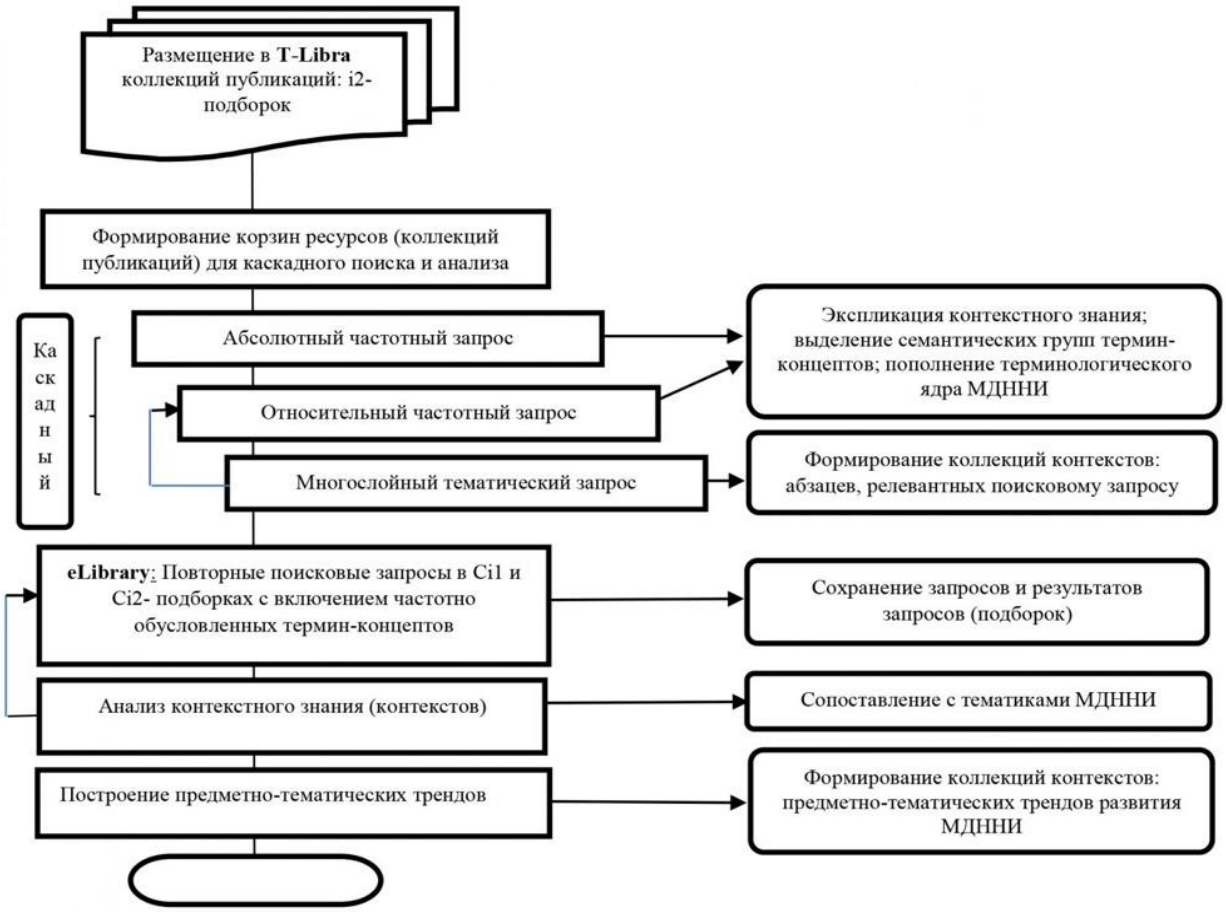



Рис. 3.14. Блок-схема, описывающая процедуры анализа литературы T-Libra и eLibrary

Синтетический метод как комплексный подход предполагает использование следующих методов и технологий:

– метод и технология сочетания абзацно-ориентированных полнотекстовых запросов, позволяющих эксплицировать «горизонтальный» контекст (микроконтекст) употребления поисковых терминов в составе авторского абзаца, и частотно-ранжированных запросов, позволяющих эксплицировать предметную

область документа или совокупности документов (вертикальный контекст, или макроконтекст);

– метод и технология гибридного поиска, то есть одновременного поиска по каталогу и по полным текстам; позволяет соотнести объекты каталога, в том числе нетекстовые (графические образы, аудио и видеоресурсы, найденные по описаниям в каталоге), и релевантные фрагмента текста, найденные по полнотекстовой библиотеке;

– метод и технология мультимодального поиска, объединяющая поиск по мультимодальной коллекции (например, музейной) и полнотекстовый поиск фрагментов текста, релевантных объектам коллекции;

– метод и технология кластеризации результатов абзацно-ориентированного поиска с управлением параметрами кластеризации. Позволяет выявить кластеры контекстного знания, соотнесенных с термин-концептами;

– метод и технология семантического картирования результатов запроса (визуализация кластеров результатов запроса).

Предлагаемые методы и технологии опираются также на следующие конкретные методики и инструменты поиска:

– методика сочетания абзацно-ориентированных и частотно-ориентированных запросов, позволяющий объединять в исследовательских целях эксплицируемые «горизонтальные» микроконтексты (в пределах авторского абзаца) и «вертикальные» макроконтексты (в пределах документа или их совокупности);

– инструмент автоматической кластеризации результатов абзацно-ориентированного запроса с обратной связью с поисковым запросом (позволяет осуществлять кластеризацию запроса и управлять ее параметрами);

– инструмент многослойного тематического запроса с вариацией используемых слоев, позволяющий выделять аспекты эксплицируемой темы (от 2 до 8 аспектов);

– инструмент фокусировки полнотекстового запроса (позволяющий задавать расстояние между поисковыми терминами в искомом абзаце и находить оптимальное соотношение между полнотой и точностью поиска);

– инструмент каскадного поиска (результаты одного запроса автоматически входят в поисковый образ другого запроса; позволяет осуществлять структурную модуляцию запроса для выявления новых элементов контекстного знания;

– инструмент гибридного квазисемантического поиска (одновременно по описаниям ресурса, взятым из каталога, и по полным текстам; используется для мультимодального поиска – например, по описаниям музейных артефактов и по полным текстам библиотеки;

– инструмент тезаурусного поиска (абзацный поиск с автоматическим включением в поисковый образ функциональной структуры тезауруса;

используется для автоматического расширения культурного контекста в ходе выполнения запроса).

В связи с тем, что современные междисциплинарные направления исследований подвержены постоянному изменению, требуется периодическое уточнение:

– терминологической базы;

– тематических направлений;

– трендов развития тематических направлений и терминологической базы (предметно-тематических трендов).

В соответствии с синтетическим методом это достигается после получения пар, состоящих из ключевых термин-концептов и выявленных на второй и третьей фазах зависимых термин-концептов. Эти выявленные термины необходимо использовать в формировании запросов в цифровых информационных ресурсах (например, НЭБ Elibrary) с соблюдением следующих основных принципов:

– в рассмотрение берутся документы начиная с года, которым ограничивался сверху этап первичной экспликации научных контекстов. Это объясняется тем, что наполнение цифровых ресурсов происходит постепенно и за прошедшее время были загружены документы текущего на момент первичной экспликации года;

– поиск необходимо производить по всему ресурсу, не ограничиваясь только релевантными тематическими направлениями. Это позволит как выявить новые научные тематики, входящие в исследуемое междисциплинарное направление, так и пополнить терминологическую базу новыми термин-концептами.

После проведения поиска из найденного массива документов экспертным образом отбираются тексты для пополнения исследуемого корпуса (например, загружаются в электронную библиотеку T-Libra) и дальнейшего контекстного анализа. На основе результатов анализа можно подтвердить или опровергнуть построенные до этого предметно-тематические тренды, а также построить новые.

### 3.2.2. Методика выбора источников, последующего контекстного поиска и отбора материалов[47]

Точность, ясность и воспроизводимость процесса поиска литературы, что важно для любого научного исследования, традиционно обеспечивается использованием одной базы данных вместо нескольких. Данную рекомендацию невозможно соблюсти, когда речь идет о текстах на разных языках или корпусе текстов из гетерогенных источников (научные базы и СМИ). Поэтому

[47] Кононова, О.В., Прокудин, Д.Е. Геймификация и социокультурные контексты цифрового урбанизма: общий подход к анализу и прогнозу научного дискурса [электронный текст] // Культура и технологии. 2019. Том 4. Вып. 3. Серьезные компьютерные игры: где, как и почему они работают. С. 83-105. URL: http://cat.ifmo.ru/ru/2019/v4-i3/193.

предлагается придерживаться данного метода, обеспечив независимость поиска, выделения, анализа и экспликации контекстного знания для каждого информационного ресурса отдельно. Так, для каждого из использованных информационных ресурсов при общем подходе – синтетическом методе – необходимо использовать поисковый запрос с уникальными условиями поиска. Воспроизводимость процесса поиска литературы затруднена постоянным обновлением информации о публикациях в некоторых базах данных за прошедшие периоды. Поэтому численные значения, полученные в один период времени, нельзя получить в последующий (это особенно характерно для НЭБ Elibrary, где на отмеченную проблему накладывается постоянный процесс внесения изменений в метаданные записей публикаций). Точность воспроизводимости экспериментов (результатов выполнения поисковых запросов) также зависит от точности повторения условий расширенного запроса. Эти факторы необходимо учитывать при повторе или проверке результатов представленных исследований. Обнадёживающим и экспериментально подтвержденным фактом является то, что при невозможности повторно получить те же численные значения относительная статистика по результатам запроса, сходимость используемого метода остается высокой и качественные результаты не претерпевают значимых изменений. Это говорит об устойчивости контекстов и трендов научного дискурса в пределах одного года при возрастающем общем объеме статейного материала от года к году.

Первоначальный выбор цифровых ресурсов базируется на учёте следующих критериев[48]:

– широта охвата по типам документов. При этом обязательным условием рассмотрения ресурсов должно быть наличие в них метаданных научных статей (в связи с этим, например, из рассмотрения можно сразу же исключить электронные каталоги библиотек);

– возможность выбора области поиска по метаданным, т.е. наличие выбора полей при формировании поискового запроса (например, название, аннотация, ключевые слова, годы публикаций и т.д.);

– достаточно глубокая временная ретроспектива содержащихся в ресурсе документов;

– возможность выбора тематики публикаций, что позволяет получить более релевантные выборки документов;

– политематический характер ресурса для максимального охвата документов и публикаций.

Так как существует достаточно большое число цифровых научных ресурсов, то для оценки их применимости в рамках проводимого исследования

[48] Прокудин Д.Е., Левит Г.С. Методы отбора цифровых информационных ресурсов на примере исследовании влияния научных идей Г.Ф. Гаузе на развитие науки // Научный сервис в сети Интернет: труды XIX Всероссийской научной конференции (18-23 сентября 2017 г., г. Новороссийск). — М.: ИПМ им. М. В. Келдыша, 2017. — 480 с. doi:10.20948/abrau-2017-75 URL: http://keldysh.ru/abrau/2017/75.pdf. С. 389-399.

необходимо использовать контрольный поисковый запрос, который максимально полно отражает основную тематику исследования При этом поиск должен проходить достаточно широко – по всем полям (название, аннотация, ключевые слова), а также при возможности и по тексту публикации. После этого по числу полученных документов можно отобрать ресурсы для отбора русскоязычных и англоязычных (или представленных на других языках) документов.

Также необходимо учитывать основные возможности цифровых научных ресурсов по представлению информации: возможность выгрузки полных текстов, возможность экспорта библиографической информации (на уровне метаданных), получения списков пристатейной литературы и т.д. Основные возможности наиболее востребованных научных цифровых ресурсов представлены в таб. 3.2.

Таб. 3.2. Основные возможности некоторых научных цифровых ресурсов

| **Источник** | **Расширенный поиск** | **Наличие полных текстов, формат** | **Доступ к текстам** | **Возможность выгрузки библиографической информации** | **Наличие аналитических инструментов** |
|---|---|---|---|---|---|
| SCOPUS | да | нет | нет | да (формат csv) | да (по источникам) |
| Academic Search Complete (EBSCO) | да | да, pdf | по подписке | нет | нет |
| Web of Science Core Collection | да | нет | нет | да (формат csv) | да (по источникам) |
| JSTOR | да | Да, pdf | по подписке | Да (RefWorks, EasyBib, RIS, Text file) | нет |
| OAIster | да | Нет | свободный доступ (по ссылке) | нет | нет |
| PLOS | нет | | свободный доступ | нет | нет |
| Google Scholar | нет | Нет | свободный доступ (по ссылке) | нет | нет |
| PubMed Central | да | | свободный доступ | да (текст, XML) | нет |
| SpringerLink | да | Да, pdf | коммерческий/по подписке | да (формат CSV) | нет |
| НЭБ Elibrary | да | Да, pdf | свободный доступ/коммерческий | на коммерческой основе через API | да |
| Киберленинка | да | Да, pdf | Свободный доступ | нет | нет |

### 3.2.3. Методика построения и интерпретации трендов

Существует несколько основных общепринятых подходов интерпретации понятия «тренд», достаточно широко представленных в отечественных и зарубежных научных работах:

1) Тренд – это представление в графическом виде изменения некоторой величины или группы величин, отражающих рассматриваемые явления или процессы в динамике на временном интервале, часто с элементами прогноза на последующие временные интервалы. Такое представление тренда всегда основано на количественных измерениях.

2) Тренд – это представление информации о ситуации или явлениях в схематическом, табличном или текстовом (описательном) виде, отражающем «срез», «картинку» за определенный интервал времени или на определенную дату. Такое представление тренда отражает качественные характеристики, содержание (семантику) исследуемых явлений. Часто качественные характеристики являются следствием обработки и анализа числовых данных, например, частот встречаемости терминов в текстах.

При исследовании контекстного знания для оценки развития терминологической базы развивающихся междисциплинарных направлений научных исследований можно выделить несколько основных уровней, по которым строятся тренды:

– предметный уровень, на котором тренды строятся по распределению базовых термин-концептов по предметным областям. Могут быть представлены либо в процентном отношении, либо в виде круговой диаграммы в абсолютном представлении (на всём временном интервале) или за определённый временной интервал. Также для представления тренда в динамике могут быть построены графики по временным интервалам с вычислением самого тренда как функции, что определяет его прогностическое значение для оценки развития как самих предметных областей, так и их значимости для исследуемого междисциплинарного направления;

– концептуальный уровень, на котором представляются тренды развития основных тематических направлений, входящих в исследуемое междисциплинарное направление. Могут быть построены, например, по основным ключевым словам исследуемых научных публикаций. Также могут быть построены либо в статике (процентное отношение ключевых слов), либо в динамике (что отражает развитие самих тематических направлений и выявляет их значимость в дальнейшем развитии междисциплинарного направления);

– тематический уровень, на котором исследуются тенденции развития терминологической базы тематических направлений, составляющих междисциплинарное направление научных исследований. Термины выявляются на основе применения синтетического метода. Могут быть построены либо в графическом представлении на временном графике, либо представляются в виде

семантических групп, при котором семантические группы отражают тематические направления, а термины распределяются по ним в порядке убывания частоты их встречаемости (что определяет развитие терминологического ландшафта исследуемого междисциплинарного направления). На основе динамики развития терминологического ландшафта можно, например, прогнозировать выход из употребления одних терминов и вхождение в научный и общественно-политический дискурс новых терминов, а также выявлять устойчивое использование синонимических терминов.

Предметно-тематические тренды строятся на массивах отобранных релевантных научных публикаций. Для этого в одной из информационных систем производится загрузка этих массивов, а затем выполняются различные запросы по опорным термин-концептам для выявления связанных с ними термин-концептами. По выявленным зависимостям между термин-концептами производится их распределение по семантическим группам. Семантические группы содержат семантически близкие термин-концепты, позволяющие выделять из текстовых массивов искомый контекст и/или термин-концепты со сходными значениями (функциональные синонимы). Далее с использованием соответствующего ПО происходит построение трендов в виде графиков, схем, диаграмм.

### 3.2.4. Визуализация результатов анализа контекстного знания

Для наилучшего восприятия результатов исследования контекстного знания целесообразно представлять их в графической форме с использованием различных средств визуализации (рис. 3.15 – 3.17 и рис. 3.18 – 3.21).


Рис. 3.15. Визуальное представление результатов аналитического отчёта о распределении по тематическим направлениям исследований публикаций, полученных по запросу «digital economy» за 2015-2021 годы (ИС Scopus)

Зачастую наглядное представление результатов позволяет адекватно их интерпретировать и сделать более полные выводы. Кроме этого, именно за счёт визуализации полученные результаты найдут широкое понимание среди научной общественности при их обнародовании либо через публикацию, либо через доклад на научном мероприятии.

Визуализация результатов исследования может быть осуществлена следующими основными доступными способами:

1) использование средств визуализации, которыми обладают цифровые научные информационные ресурсы (рис. 3.15 – 3.17). Такого рода средства есть, например, в ИС Scopus.


Рис. 3.16. Визуальное представление результатов аналитического отчёта о распределении по организациям авторов публикаций, полученных по запросу «digital economy» за 2015-2021 годы (ИС Scopus)

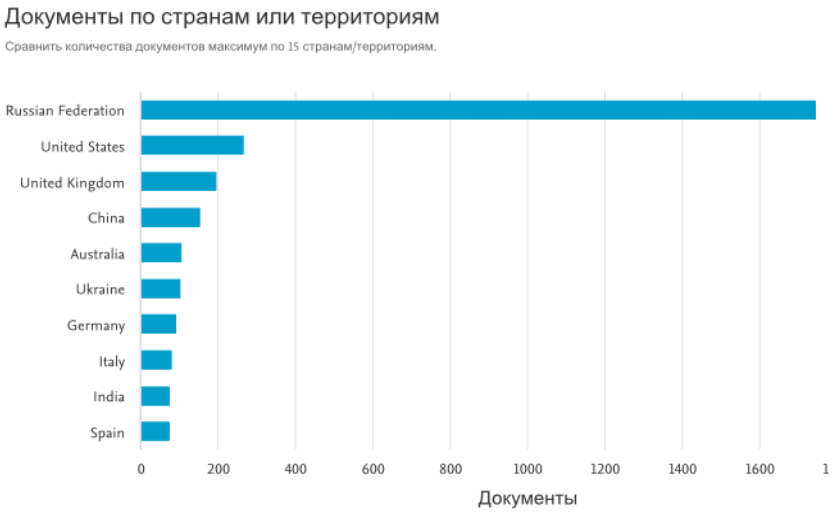



Рис. 3.17. Визуальное представление результатов аналитического отчёта о распределении по странам авторов публикаций, полученных по запросу «digital economy» за 2015-2021 годы (ИС Scopus)

2) использование инструментов визуализации результатов обработки контекстного знания в специализированных информационных системах. Эти системы обладают более развитыми средствами визуализации, а также возможности экспорта изображений. На рис. 3.18 – 3.21 представлены примеры визуализации в таких инструментах как Voyant-tools, Tropes и Sketch Engine.

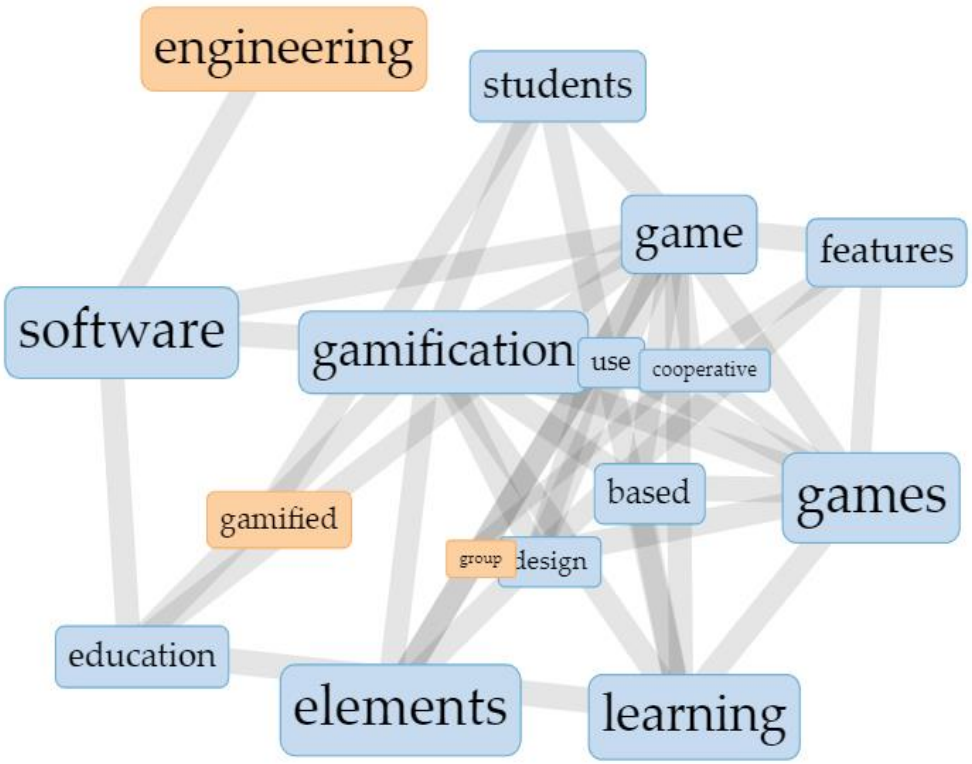



Рис. 3.18. Визуальное представление результатов анализа совместное использование терминов в результате анализа корпуса текстов (ИС Voyant-tools)

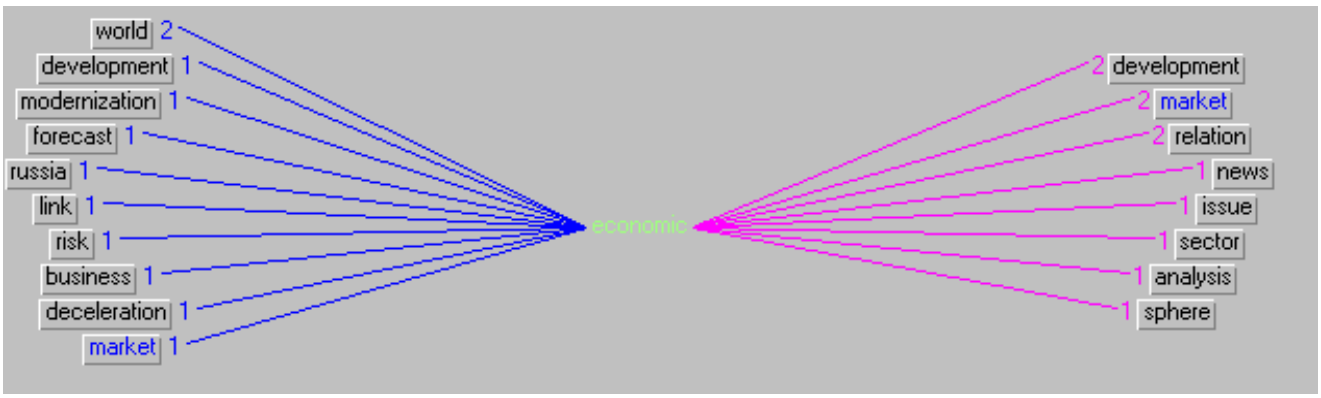



Рис. 3.19. Визуальное представление результатов анализа отношения между эквивалентными классами термин-концептов с опорным (ИС Tropes)

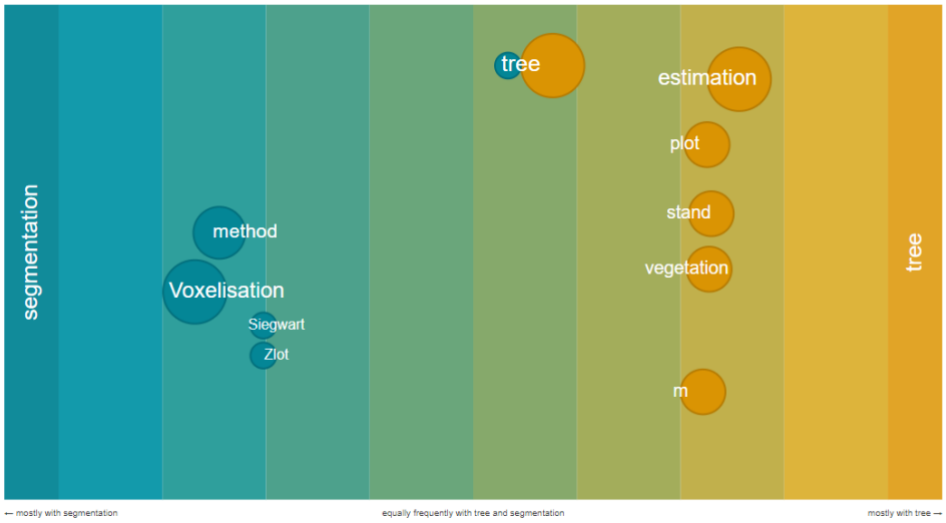



Рис. 3.20. Визуальное представление результатов анализа коллокаций для двух термин-концептов (ИС Sketch Engine)

3) использование для визуализации результатов анализа контекстного знания возможностей специализированного программного обеспечения, например электронных таблиц. При этом в такое программное обеспечение импортируются полученные результаты или вводятся вручную. Затем для визуализации используется мастер построения диаграмм или графиков.

## Контрольные вопросы к Главе 3

1. Дайте определение и перечислите основные задачи тематического моделирования.

2. Что в общем случае является результатом тематического моделирования? (выберите правильный ответ):

а) вероятности принадлежности слов/документов к теме, матрица распределения слов по темам и матрица распределения документов по темам

б) вероятности принадлежности слов/документов к теме, матрица распределения слов по темам, матрица распределения документов по темам

в) матрица распределения слов по темам, матрица распределения документов по темам

г) вероятности принадлежности слов/документов к теме

3. На каких из этапов тематического моделирования применяются библиотеки Rutermextract и BigARTM?

4. Сформулируйте возможности применения тематического моделирования при выполнении научно-исследовательской работы магистранта / аспиранта.

5. Что является сильной стороной методов тематического моделирования, а что слабой применительно к задачам определения тематики и выбора темы магистерской или научной диссертации?

6. Какие задачи решаются с использованием наукометрических и библиометрических методов?

7. Какие основные исследовательские задачи возможно решить с применением синтетического метода?

8. Для обработки каких видов контекстов в качестве источников используются реферативные базы научной информации?

9. Для чего строятся предметно-тематические тренды?

10. Какие средства используются для визуализации результатов анализа контекстного знания?

## Литература


1. Алексеева С.В., Кольцов С.Н., Кольцова О.Ю. Linis-crowd.org: лексический ресурс для анализа тональности социально-политических текстов на русском языке // Компьютерная лингвистика и вычислительные онтологии. Сборник научных статей XVIII Объединенной конференции «Интернет и современное общество» IMS-2015, Санкт-Петербург, 23-25 июня 2015 г. ИТМО, 2015. С. 25-34. URL: https://openbooks.itmo.ru/ru/file/2203/2203.pdf.
2. Будапештская Инициатива «Открытый Доступ» [Электронный текст] // Budapest Open Access Initiative. Russian Translation. URL: http://www.budapestopenaccessinitiative.org/translations/russian-translation
3. Вербицкий А.А. Контекст (в психологии) / Под ред. А.В. Петровского // Психологический лексикон. Энциклопедический словарь в 6 т. Т. 3. М.: ПЕР СЭ, 2005. С. 137-138.
4. Вербицкий А.А. Контекстное обучение в компетентностном формате (Компетентностный подход как новая образовательная парадигма) // Проблемы социально-экономического развития Сибири. 2011. № 4 (6). С. 67-73. URL: http://brstu.ru/static/unit/journal_2/docs/number6/67-73.pdf.
5. Воронцов К. В. Вероятностное тематическое моделирование. 2013 [Электронный ресурс].
   URL: http://www.machinelearning.ru/wiki/images/2/22/Voron-2013-ptm.pdf.
6. Глушков, Н. А. Анализ методов тематического моделирования текстов на естественном языке // Молодой ученый. 2018. № 19 (205). С. 101-103. URL: https://moluch.ru/archive/205/50247.
7. Дрокин И.С., Бухвалов О.Л., Сорокин С.Ю. Способ формирования математических моделей пациента с использованием технологий искусственного интеллекта // Патент RU 2 720 363 C2. 29.12.2017.
8. Дурай-Новикова К. М. Формирование профессиональной готовности студентов к педагогической деятельности: автореф. дис. … д-ра пед. наук. М., 1983. 18 с.
9. Елькина Е.Е., Кононова О.В., Прокудин Д.Е. Типология контекстов и принципы контекстного подхода в междисциплинарных научных исследованиях // Современные информационные технологии и ИТ-образование. 2019. Том 15, № 1. С. 141-153. DOI: 10.25559/SITITO.15.201901.141-153.
10. Земсков А.И., Шрайберг Я.Л. Системы открытого доступа к информации: причины и история возникновения [Электронный текст] // Науч. и техн. б-ки. 2008. №4.
    URL: http://ellib.gpntb.ru/subscribe/ntb/2008/4/ntb_4_2_2008.htm.

11. Иванников А.Д., Тихонов А.Н., Цветков В.Я. Критерии готовности к использованию информационных технологий // Международный журнал прикладных и фундаментальных исследований. 2009. №3. С. 84-85.
12. Кондаков Н. И. Логический словарь-справочник. М.: Книга по Требованию, 2013. 720 с.
13. Кононова О.В., Прокудин Д.Е. Подход к извлечению, экспликации и представлению контекстного знания при изучении развивающихся междисциплинарных направлений исследований // International Journal of Open Information Technologies. 2020. Том 8, № 1. С. 90-101. URL: http://injoit.org/index.php/j1/article/view/882/844.
14. Кононова О.В., Ляпин С.Х., Прокудин Д.Е. Синтетический метод извлечения контекстного знания в русскоязычной социально-гуманитарной сфере: комплексный подход. Информационное общество: образование, наука, культура и технологии будущего. 2017. Вып. 1. С. 52-67. DOI: 10.17586/2587-8557-2017-1-52-67.
15. Кононова О.В., Прокудин Д.Е. Геймификация и социокультурные контексты цифрового урбанизма: общий подход к анализу и прогнозу научного дискурса [электронный текст] // Культура и технологии. 2019. Том 4. Вып. 3. Серьезные компьютерные игры: где, как и почему они работают. С. 83-105. URL: http://cat.ifmo.ru/ru/2019/v4-i3/193.
16. Коршунов А., Гомзин А. Тематическое моделирование текстов на естественном языке // Труды Института системного программирования РАН. 2012. С. 215-242. https://elibrary.ru/item.asp?id=18361454.
17. Красавина В.Д., Мирзагитова А.Р. Оптимизация поиска в системе LeadScanner с помощью автоматического выделения ключевых слов и словосочетаний // Труды международной конференции «Корпусная лингвистика–2015». СПб. 2015. С. 296−306.
18. Кручинина Г.А. Готовность будущего учителя к использованию новых информационных технологий обучения (теоретические основы, экспериментальные исследования): Монография. Под ред. В.А. Сластенина. М., Изд-во «Прометей». 1996. 176 с.
19. Лагозе К. Связывая прошлое с будущим: Научные коммуникации в 21 веке [Электронный текст] // Электронные библиотеки. 2004. Т. 7, вып. 3. URL: http://www.elbib.ru/index.phtml?page=elbib/rus/journal/2004/part3/kl (дата обращения: 24.04.2015)
20. Манаев О.Т. Контент-анализ как метод исследования // Социология: энциклопедия. М., 2003. URL: http://psyfactor.org/lib/content-analysis3.htm.
21. Мбого И.А., Прокудин Д.Е., Чугунов В.А. Формирование информационного пространства междисциплинарного научного направления: подходы и решения // Межотраслевая информационная служба. 2015. №1. С. 36-44.

22. Микешина Л.А. Философия познания. Полемические главы. М.: Прогресс-Традиция, 2002. 624 с.
23. Можейко М. А. Контекст // Новейший философский словарь. Мн., 2001.
24. Москвина А.Д., Митрофанова О.А., Ерофеева А.Р., Харабет Я.К. Автоматическое выделение ключевых слов и словосочетаний из русскоязычных корпусов текстов с помощью алгоритма RAKE // Труды международной конференции «Корпусная лингвистика–2017». СПб. 2017. С. 268−277.
25. О базе данных системы Соционет [Электронный текст] // Соционет – научная информационная система. URL: http://socionet.ru/bd.htm.
26. Описание библиотеки rutermextract [Электронный ресурс]. URL: https://pypi.org/project/rutermextract/.
27. Паринов С.И. Развитие электронных библиотек – путь к Открытой Науке [Электронный текст] // Электронные библиотеки: перспективные методы и технологии, электронные коллекции: Труды XI Всероссийской научной конференции RCDL'2009. Петрозаводск: КарНЦ РАН, 2009. С. 225-234. URL: http://rcdl.ru/doc/2009/225_234_Invited-2.pdf.
28. Паринов С.И., Ляпунов В.М., Пузырев Р.Л. Система Соционет как платформа для разработки научных информационных ресурсов и онлайновых сервисов [Электронный текст] // Электронные библиотеки. 2003. Т. 6. № 1. С. 6-25. URL: http://elibrary.ru/item.asp?id=9121156.
29. Печерская С. А. Теоретико-методологические основы готовности студентов к использованию информационных технологий. Сочи: НОЦ РАО. 2007.
30. Прокудин Д.Е. Информатика в гуманитарных науках: Учебно-методическое пособие. СПб.: Фонд развития конфликтологии, 2016. 52 с.
31. Прокудин Д.Е. Проектирование и реализация комплексной информационной системы поддержки научных исследований // Технологии информационного общества в науке, образовании и культуре: сборник научных статей. Труды XVII Всероссийской объединенной конференции «Интернет и современное общество» (IMS-2014), Санкт-Петербург, 19 – 20 ноября 2014 г. С. 31-36.
32. Прокудин Д.Е. Тенденции научной коммуникации в информационном обществе // Информация – Коммуникация – Общество (ИКО–2014): Материалы XI Всероссийской научной конференции / Санкт-Петербург, 23 –24 января 2014 г. СПб., 2014. 168 с. С. 132-135.
33. Прокудин Д.Е., Левит Г.С. Методы отбора цифровых информационных ресурсов на примере исследовании влияния научных идей Г.Ф. Гаузе на развитие науки // Научный сервис в сети Интернет: труды XIX Всероссийской научной конференции (18-23 сентября 2017 г., г. Новороссийск). М.: ИПМ им. М.В.Келдыша, 2017. С. 389-499. DOI:10.20948/abrau-2017-75.

34. Прокудин Д.Е., Чугунов А.В. Сетевые электронные репозитории, системы агрегации и сохранения в концепции свободного доступа к научной информации // Сборники Президентской библиотеки / Серия «Электронная библиотека». Вып. 6: Интегрированные цифровые ресурсы: организационно-технологические и научно-методические основы развития. 2015. 271 с. С. 162-183.
35. Смелик Н.Д., Фильченков А.А. Мультимодальная тематическая модель текстов и изображений на основе использования их векторного представления // Машинное обучение и анализ данных, 2016. Т. 2. № 4. С. 421-441. DOI:10.21469/22233792.2.4.05.
36. Сорокина Ю. В. Понятие мультимодальности и вопросы анализа мультимодального лекционного дискурса // Филологические науки. Вопросы теории и практики. 2017. № 10. Ч. 1. С. 168-170.
37. Сучкова, Л.И. Win32 API: основы программирования: учебное пособие/ Л.И. Сучкова; АлтГТУ им. ИИ. Ползунова. -Барнаул, АлтГТУ, 2010. С. 138.
38. Черноскутов Ю.Ю. Контекст и логические теории пресуппозиции // Логико-философские штудии. 2005. № 3. С. 220-238. URL: http://ojs.philosophy.spbu.ru/index.php/lphs/article/view/135/136.
39. Электронные репозитории [Электронный ресурс]. Научная библиотека Дагестанского государственного университета. URL: http://elib.dgu.ru/?q=node/739.
40. Электронные репозитории: какие они и как в них зарегистрироваться? // Академическая среда. 2012. № 3(08). URL: http://www.hse.ru/data/2012/02/29/1265868424/Area_8.pdf.
41. Barrios F., LoÏpez F., Argerich L., Wachenchauzer, R. Variations of the Similarity Function of TextRank for Automated Summarization // Anales de las 44JAIIO. Jornadas Argentinas de InformaÏtica, Argentine Symposium on Artificial Intelligence, 2015.
42. Berlin 3 Open Access: Progress in Implementing the Berlin Declaration on Open Access to Knowledge in the Sciences and Humanities. Feb 28th - Mar 1st, 2005, University of Southampton, UK [Электронный текст] // http://www.eprints.org/events/berlin3/outcomes.html.
43. Bougiatiotis K., Giannakopoulos T. Enhanced movie content similarity based on textual, auditory and visual information // Expert Systems with Applications. 2018. Vol. 96. P. 86-102. DOI: 10.1016/j.eswa.2017.11.050.
44. Brezillon P., Pomerol J.-Ch. Contextual Knowledge and Proceduralized Context AAAI Technical Report WS-99-14. Compilation copyright © 1999, AAAI.
45. Chergui, W., et al. An approach to the acquisition of tacit knowledge based on an ontological model. Journal of King Saud University – Computer and Information Sciences. 2020. Vol. 32. Iss. 7. P. 818-828. DOI: 10.1016/j.jksuci.2018.09.012.

46. Cromley J. G., Kunze A. J., Parpucu Dane A. Multi-text multi-modal reading processes and comprehension // Learning and Instruction. 2021. Volume 71. 101413. DOI: 10.1016/j.learninstruc.2020.101413.
47. Dancygier B., Vandelanotte L. Image-schematic scaffolding in textual and visual artefacts // Journal of Pragmatics. 2017. Vol. 122. P. 91-106. DOI: 10.1016/j.pragma.2017.07.013.
48. Dimitrovski I., Kocev D., Kitanovski I., Loskovska S., Džeroski S.Improved medical image modality classification using a combination of visual and textual features // Computerized Medical Imaging and Graphics. 2015. Vol. 39. P. 14-26, DOI: 10.1016/j.compmedimag.2014.06.005.
49. Melucci M. Vector-Space Model. In: LIU L., ÖZSU M.T. (eds) // Encyclopedia of Database Systems. Springer, Boston, MA. 2009. DOI: 10.1007/978-0-387-39940-9_918.
50. Goczyła K., Waloszek A., Waloszek W. An Analysis of Contextual Aspects of Conceptualization: A Case Study and Prospects // Bembenik R., Skonieczny Ł., Rybiński H., Kryszkiewicz M., Niezgódka M. (eds) Intelligent Tools for Building a Scientific Information Platform: From Research to Implementation. Studies in Computational Intelligence. 2014. Vol. 541. P. 75-104. DOI: 10.1007/978-3-319-04714-0_6.
51. Jaglan G., Malik S.K. Blending Semantic Web with Recommender Systems // International Journal of Computer Sciences and Engineering. 2018. Vol. 6. Iss. 5. P. 523-531. DOI: 10.26438/ijcse/v6i5.523531.
52. Kanygin G., Kononova O. The Expression of Tacit Knowledge by Actors of Smart Technologies // CEUR Workshop Proceedings. 2020. Vol. 2784 /Proceedings of the 22nd Conference on Scientific Services & Internet (SSI-2020), Novorossiysk-Abrau (online), September 21-25, 2020. P. 98-112. http://ceur-ws.org/Vol-2784/rpaper09.pdf.
53. Kress G. What Is Mode? // A Handbook of Multimodal Analysis / ed. by C. Jewitt. L. N. Y.: Routledge, 2009. P. 54-67.
54. Krizhevsky A., Sutskever I., Hinton G. E. Imagenet classification with deep convolutional neural networks // Proceedings of the 25th International Conference on Neural Information Processing Systems. Vol. 1 (NIPS'12). Curran Associates Inc., Red Hook, NY, USA. 2012. P. 1097–1105.
55. Kumar A., Srinivasan K., Zomaya A.Y. Hybrid context enriched deep learning model for fine-grained sentiment analysis in textual and visual semiotic modality social data // Information Processing & Management. Vol. 57. Issue 1. 2020. 102141. DOI: 10.1016/j.ipm.2019.102141.
56. Lee C.K., Foo S., Goh D. On the Concept and Types of Knowledge // Journal of Information & Knowledge Management. 2006. Vol. 5. No. 2. P. 151–163.

57. OCLC: союз библиотек и удобство читателей [Электронный текст] // Российская государственная библиотека. URL: http://www.rsl.ru/ru/s7/s409/2013/20137642.
58. OCLC. OAIster Contributors [Электронный ресурс]. URL: http://www.oclc.org/oaister/contributors.en.html.
59. Ortiz R., Pinto D., Tovar M., Jimenez-Salazar H. An Unsupervised Approach to Automatic Keyphrase Extraction from Scientific Articles, 2010.
60. Perera S., Mendes P.N., Alex A., Sheth A.P., Thirunarayan K. Implicit Entity Linking in Tweets. http://www.internetlivestats.com/twitter-statistics.
61. Prabakaran N., Kannadasan R. Contexts enabled Decision Making using sensors to perceive pervasive environment // International Conference on Computational Intelligence and Data Science (ICCIDS 2018). Procedia Computer Science. 2018. Vol. 132. P. 477–485.
62. Wang K., Meng W., Li S., Yang S. Multi-Modal Mention Topic Model for mentionee recommendation // Neurocomputing. 2019. Volume 325. P. 190-199. DOI: 10.1016/j.neucom.2018.10.024
63. Xu J., Huang F., Zhang X., Wang S., Li C., Li Z., He Y. Visual-textual sentiment classification with bi-directional multi-level attention networks // Knowledge-Based Systems. 2019. Vol. 178. P. 61-73. DOI: 10.1016/j.knosys.2019.04.018.

## Словарь основных понятий

***Автор документа.*** Физическое или юридическое лицо, создавшее документ.

***Авторское право***. Форма защиты продукта интеллектуальной деятельности физического или юридического лица с помощью норм гражданского права.

***Авторское право на базу данных***. Авторское право на базу данных, состоящую из материалов, не являющихся объектами авторского права, принадлежит лицам, создавшим базу данных. Авторское право на базу данных признается при условии соблюдения авторского права на каждое из произведений, включенных в эту базу данных. Авторское право на каждое из произведений, включенных в базу данных, сохраняется. Эти произведения могут использоваться независимо от такой базы данных. Авторское право на базу данных не препятствует другим лицам осуществлять самостоятельный подбор и организацию произведений и материалов, входящих в эту базу данных.

***Агрегатор научной информации***. Сайт, на котором агрегируется (собирается) и классифицируется информация о различных научных сайтах и информационных системах.

***Анализ данных (наука о данных, англ. data science, DS)*** — объединение ряда научных областей (статистика, машинное обучение, большие данные, визуализация), занимающихся построением систем анализа и обработки данных.

***Анализ текста (text analysis)*** – это метод анализа, обычно неструктурированного текста, с целью извлечения информации. Результатом анализа текста, а также анализа содержания или аналитики текста (content analysis or text analytics) являются структурированные данные.

***Аннотация.*** Вторичный документ, содержащий краткую обобщенную в характеристику первичного документа с точки зрения его назначения, содержания, вида, формы и других особенностей.

***Аутентификация***. 1. Проверка принадлежности субъекту доступа предъявленного им идентификатора. 2. Подтверждение подлинности. 3. В вычислительных сетях – процесс установления подлинности сообщения, источника и приемника данных. 4. Установление подлинности информации независимо от источника этой информации. 5. Процесс проверки идентичности пользователя, терминала или провайдера.

***База данных***. Совокупность данных, относящихся к той или иной предметной области, организованных в соответствии с общими принципами описания, хранения и манипулирования данными.

***База данных автоматизированная***. Означает любой набор данных, к которым применяется автоматическая обработка.

***Библиографическая запись*** является элементом библиографической информации, фиксирующим в документальной форме сведения о документе, позволяющим его идентифицировать, раскрыть его составные части и содержание в целях его поиска. Библиографическая запись включает в себя заголовок, библиографическое описание, классификационные индексы, шифры хранения документа, сведения служебного характера.

***Библиографическая ссылка*** служит источником библиографической информации о документах – объектах ссылки. Объектами составления библиографической ссылки являются все виды опубликованных и неопубликованных документов на любых носителях (в том числе электронные ресурсы локального и удаленного доступа), а также составные части документов.

***Библиографическая эвристика.*** Дисциплина, занимающаяся теоретическими проблемами библиографического поиска.

***Библиографический поиск***. Информационный поиск, осуществляемый на основании библиографических данных.

***Библиотека.*** Информационное, культурное, образовательное учреждение, располагающее организованным фондом тиражированных документов и предоставляющее их во временное пользование физическим и юридическим лицам; библиотека может быть самостоятельным учреждение или структурным подразделением предприятия, учреждения, организации.

***Библиотечное дело***. Отрасль информационной, культурно-просветительской и образовательной деятельности, в задачи которой входят создание и развитие сети библиотек, формирование и обработка их фондов, организация библиотечного, информационного и справочно-библиографического обслуживания пользователей библиотек, подготовка кадров работников библиотек, научное и методическое обеспечение развития библиотек.

***Библиотечный поиск***. Поиск первичных документов в собрании этих документов.

***Браузер.*** Программа на клиентском компьютере, используемая для выборки/чтения гипертекстовых документов из веб-сайта (сервера), отображения их на экране и печати, для перехода к другим документам посредством гипертекстовых ссылок.

***Веб-сайт***. Страницы гипертекстовой информации в электронной форме, связанные друг с другом гипертекстовыми ссылками и расположенные на веб-сервере для доступа с любого клиентского компьютера в сети Интернет посредством браузера.

***Вертикальный контекст***. Крупный фрагмент текста, в котором его значение может подвергаться серьезным смысловым трансформациям.

***Владелец информационных ресурсов, информационных систем, технологий и средств их обеспечения***. Субъект, осуществляющий владение и пользование указанными объектами и реализующий полномочия распоряжения в пределах, установленных законом.

***Всемирная паутина (WWW)***. Совокупность информации, размещенной в электронной форме на серверах сети Интернет, доступ к которой можно получить при помощи браузера и других коммуникационных программ.

***Выпуск в свет (опубликование) программы для ЭВМ или базы данных.*** Предоставление экземпляров программы для ЭВМ или базы данных с согласия автора неопределенному кругу лиц, при условии, что количество таких экземпляров должно удовлетворять потребности этого круга лиц, принимая во внимание характер указанных произведений.

***Выходные данные периодического печатного издания***. Каждый выпуск периодического печатного издания должен содержать следующие сведения: 1) название издания; 2) учредитель (соучредители); 3) фамилия, инициалы главного редактора; 4) порядковый номер выпуска и дата его выхода в свет, а для газет – также время подписания в печать (установленное по графику и фактическое); 5) индекс – для изданий, распространяемых через предприятия связи; 6) тираж; 7) цена, либо пометка «Свободная цена», либо пометка «Бесплатно»; 8) адреса редакции, издателя, типографии. В настоящее время общепринятым является наличие у периодического печатного издания ISSN (см. ***Международный стандартный серийный номер***) – уникального номера, позволяющего идентифицировать издание.

***Генетические алгоритмы*** (англ. genetic algorithms). Семейство методов для поиска оптимального решения, в основе которых лежит симуляция эволюционного процесса.

***Гиперссылка.*** Указывающая ссылка с одной веб-страницы на любое место любой веб-страницы Всемирной паутины.

***Глубокая паутина*** (также известна как «невидимая сеть») — множество веб-страниц Всемирной паутины, не индексируемых поисковыми системами. Наиболее значительной частью глубокой паутины является «глубинный веб» (от англ. deep web, hidden web), состоящий из веб-страниц, динамически генерируемых по запросам к онлайн базам данных, а также веб-страницы, защищённые от доступа различными механизмами (аутентификация, «каптча» и т.д.).

***Горизонтальный контекст***. Ближайшее окружение слова в предложении.

***Готовность к использованию ИКТ в профессиональной деятельности.*** Это интегральное образование, включающее в себя высокую мотивацию к использованию ИКТ в своей профессиональной деятельности, знание теоретических аспектов использования ИКТ, проявление соответствующих эмоционально-волевых качеств и реализации определенного комплекса профессиональных умений.

***Дизайн исследования***. Комбинация требований относительно сбора и анализа данных, необходимых для достижения целей исследования.

***Дизайн кейс-стади***. Предназначен для подробного изучения одного или небольшого количества случаев. Акцент при этом делается не на распространении результатов на всю генеральную совокупность, а на качестве теоретического анализа и объяснении механизма функционирования того или иного явления.

***Документ.*** Общее имя, используемое для обозначения любого файла, документа или веб-страницы, из которых состоит корпус материалов/текстов. 1. Материальный объект с зафиксированной на нем информацией в виде текста, звукозаписи или изображения, предназначенный для передачи во времени и пространстве в целях хранения и общественного использования. 2. Носитель зафиксированной по установленным формам и правилам информации, необходимой для реализации информационных процессов. 3. Зафиксированная на материальном носителе информация с реквизитами, позволяющими ее идентифицировать.

***Документально-фактографический поиск***. Поиск фрагментов текста в документах.

***Доступ к информации и знаниям***. Всеобщая доступность необходимых методов, средств и навыков для эффективного использования знаний, т.е. доступность сетей, инфраструктуры и услуг, а также информационных ресурсов, необходимых для полноценной реализации политических и социокультурных прав личности в обществе; средство, позволяющее гражданам контактировать с релевантной внешней средой.

***Защита информации***. Совокупность методов и средств, обеспечивающих целостность, конфиденциальность и доступность информации в условиях воздействия на нее угроз естественного или искусственного характера, реализация которых может привести к нанесению ущерба владельцам или пользователям информации.

***Извлечение терминов*** (анг. term extraction). Процесс определения предметного словарного запаса в предметном тексте, обычно с использованием специализированного программного обеспечения. Поиск однословных и

многословных терминов может быть основан на сравнении с частотой этих слов и фраз в справочном корпусе.

***Индексирование документа***. Выражение содержания документов и запросов средствами информационно-поискового языка.

***Институциональный репозиторий***. Электронный архив для длительного хранения, накопления и обеспечения долговременного и надежного открытого доступа к результатам научных исследований, проводимых в учреждении.

***Интеллектуальный анализ текста*** (анг. text mining). Автоматический процесс извлечения информации из текста, например ключевых слов текста, контекстов или его источника(ов).

***Интернет***. Глобальная компьютерная сеть, которая возникла на базе американской сети ARPANET в конце 60-х годов XX века. Предлагает самые различные услуги (например, электронную почту, пересылку файлов, доступ к гипертекстовым страницам и т.д.), используя общий стандарт (протокол) передачи TCP/IP. Интернет не является однородной сетью, а представляет собой множество взаимосвязанных сетей, имеющих децентрализованную организацию.

***Интерфейс***. Набор правил непосредственного взаимодействия между компьютером и его пользователем или между двумя устройствами.

***Интрасеть***. Использование Интернет-технологий в пределах организации или предприятия, возможно, географически распределенных.

***ИнфограЇфика***. Графический способ подачи информации, данных и знаний, целью которого является быстро и чётко преподносить сложную информацию.

***Информатизация***. 1. Процесс интенсификации производства и распространение знаний и информации, основанный на использовании ИКТ. 2. Направленный процесс системной интеграции компьютерных средств, информационных и коммуникационных технологий с целью получения новых общесистемных свойств, позволяющих более эффективно организовать продуктивную деятельность человека, группы, социума. 3. Процесс целенаправленного внедрения ИКТ в определённую деятельность человека с целью повышения её эффективности.

***Информационная безопасность***. Имеет три основные составляющие: конфиденциальность, целостность и доступность. Конфиденциальность относится к защите чувствительной информации от несанкционированного доступа. Целостность означает защиту точности и полноты информации и программного обеспечения. Доступность – это обеспечение доступности информации и основных услуг для пользователя в нужное для него время.

***Информационное законодательство***. Совокупность законов, нормативных актов и других форм правового регулирования в сфере обращения информации, производства и применения информационных и коммуникационных технологий.

***Информационное общество***. Понятие «информационное общество» характеризует такую структуру экономики и организации общества, в которой информация является доминирующим ресурсом и играет центральную роль в развитии производительных сил. При этом эффективность её использования определяется представлением в электронной (цифровой) форме и применением информационно-коммуникационных технологий для её обработки.

***Информационное пространство***. Интегральное электронное информационное пространство, образуемое при использовании электронных сетей.

***Информационные ресурсы***. Документы и массивы документов в информационных системах (библиотеках, архивах, фондах, банках данных, депозитариях, музейных хранениях и др.).

***Информационно-коммуникационные технологии***. 1. Приемы, способы, методы применения средств вычислительной техники и телекоммуникаций при выполнении функций сбора, хранения, обработки, передачи и использования данных. 2. Система методов и способов сбора, накопления, хранения, поиска, обработки и передачи информации. 3. Совокупность программного, аппаратного обеспечения и методов их применения в определённой предметной деятельности человека.

***Информационно-поисковая система.*** Совокупность средств для хранения, поиска и выдачи по запросу нужной информации. Поиск (размещение) информации в информационно-поисковой системе осуществляется вручную или с помощью ЭВМ в соответствии с принятым информационным языком по определенным правилам (алгоритму). Пример простейшей информационно-поисковой системы - библиотечный каталог; автоматизированные информационно-поисковые системы применяют в автоматизированных системах управления.

***Информация***. Сведения о лицах, предметах, фактах, событиях, явлениях и процессах, независимо от формы их представления.

***Искусственный интеллект*** (ИИ, англ. artificial intelligence, AI). Научное направление, задачей которого является создание интеллектуальных систем, лежит на стыке информатики, статистики и анализа данных, а также занимается вопросами, связанными с философией и этичностью использования интеллектуальных систем.

***Качество информации***. Совокупность свойств, отражающих степень пригодности конкретной информации об объектах и их взаимосвязях для

достижения целей, стоящих перед пользователем, при реализации тех или иных видов деятельности. В состав наиболее важных параметров входят: достоверность, актуальность, полнота.

***Классификатор***. Официальный документ, представляющий систематизированный свод наименований и кодов классификационных группировок и (или) объектов классификации.

***Классификация***. Разделение множества объектов на подмножества по их сходству или различию в соответствии с принятыми методами. Классификация (классифицирование) – это процесс отнесения объектов заданного множества к группировкам (классам) какой-либо системы классификации, образованной на основе использования совокупности признаков, присущих этим объектам.

***Кластеризация***. Разбиение текстовых документов на определенное, заранее неизвестное количество подмножеств, помеченных какими-то семантическими описателями. В качестве семантических описателей могут быть взяты отдельные признаки, темы, семантические группы.

***Ключевые слова*** (анг. key words). Понятие, используемое в связи с извлечением ключевых слов и терминов. Ключевые слова — это слова (элементы с одним токеном), которые чаще встречаются в основном корпусе (focus corpus), чем в корпусе ссылок (reference corpus). Они могут использоваться для определения того, что является специфическим для одного корпуса (фокусного корпуса) или его субкорпуса по сравнению с другим корпусом (эталонным корпусом) или его субкорпусом.

***Коллекция*** или ***библиотека текстов.*** Подборка данных/материалов одной модальности или нескольких модальностей (мультимодальных) данных/материалов.

***Коллекция материалов***. Подборка текстовых фрагментов (контекстов размерностью в один или несколько абзацев), полученная как результат обработки и анализа одного или нескольких корпусов текста заданной тематики для дальнейшего использования в исследовательской работе, например цитирования или создания кейсов.

***Компиляция корпуса*** (анг. corpus compilation). Обработка массива текстовых данных с целью преобразования коллекции или подборки текста в корпус для автоматизированного анализа. Анализ текста, поиск и экспликация контекстов возможны только в скомпилированном корпусе.

***Конкорданс*** (анг. concordance). Список всех употреблений заданного языкового выражения (например, слова) в контексте, возможно, со ссылками на источник. Поиск в корпусе данных позволяет по любому слову построить конкорданс – список всех употреблений данного слова в контексте со ссылками на источник.

***Компьютер***. См. электронно-вычислительная машина.

***Компьютерная грамотность***. Овладение минимальным набором знаний и навыков работы на персональном компьютере.

***Контекст***. 1. Фрагмент семиотической системы и динамический когнитивный механизм, порождающий смыслы и значения в структуре слова, текста, гипертекста, ситуации, объекта в процессе взаимодействия внутренних (психических, когнитивных) процессов и внешних (текст, объект, событие, картина мира) структурных уровней семантической сети. 2. Связное словесное целое по отношению к входящему в него определенному слову или фразе, в данном определении характеризуется как целостность, определяющая смысл фразы. 3. Некоторая локализованная в пространстве и времени совокупность высказываний и терминов, в которую входит исследуемый термин. 4. Независимая понятийная единица категориального аппарата, которая может быть положена в основу классификации научных текстов, визуализации иерархических и ассоциативных отношений между терминами. Контексты бывают текстовой и нетекстовой модальности, типы контекстов разных модальностей представлены в типологии контекстов. Практическое использование тех или иных типов контекстов определяется выбором подходов, технологий и инструментов исследования.

***Контекстное знание.*** Умение правильно «читать» контекст, извлекать и интерпретировать профессионально значимую информацию из любых источников.

***Контекстный поиск.*** Метод последовательного поиска фрагментов текстовых записей релевантных пользовательскому запросу в соответствии с требуемым контекстом.

***Контекстуальная модель*** (модель контекста). Система управления контекстным знанием (дискурсом), ответственная за конструирование и объединение нескольких ситуационных моделей в структуре связного текста, определяющая его жанр (дискурс), цели, интересы и социально-ролевой статус участников дискурса

***Контент***. Любое информационно значимое наполнение сервера, информационного комплекса, информационной системы – тексты, графика, мультимедиа. Существенными параметрами контента является его объем, актуальность и релевантность.

***Конфиденциальность***. 1. Право на защиту данных, принадлежащих отдельным лицам или организациям. 2. Правовой режим информации, не подлежащий огласке. 3. Содержание критичной информации в секрете, доступ к ней ограничен узким кругом пользователей (отдельных лиц или организаций).

***Концепты***. Интерпретаторы смыслов, форма обработки субъективного опыта путем подведения его под определенные категории и классы.

***Корпус*** (анг. corpus, corpora). Подобранная и обработанная по определённым правилам совокупность текстов на некотором языке в электронной форме, используемых в качестве базы для исследования. Корпус содержит разметку, или аннотацию – особую характеристику корпуса, несущую дополнительную информацию о свойствах входящих в него текстов. Наличие разметки – основное отличие корпуса от коллекций или библиотек текстов. С исследовательской позиции различают несколько типов корпусов: основной или эталонный (focus corpus), ссылочный (reference corpus), обучающий корпус (learner corpus). Могут быть введены и другие типы корпусов в зависимости от задач, используемых подходов, технологий и инструментов, предметной области исследования.

***Кросс-секционный дизайн.*** Предполагает сбор данных относительно большого числа единиц наблюдения. Как правило, предполагает использование выборочного метода с целью репрезентации генеральной совокупности. Данные собираются один раз, носят количественный характер. Далее рассчитываются описательные и корреляционные характеристики, делаются статистические выводы.

***Лемматизация***. Процесс приведения слова к нормальной̆ форме. Применительно к русскому языку процесс лемматизации приводит существительные в любой̆ форме к форме именительного падежа единственного числа, прилагательные в любой̆ форме — к форме именительного падежа единственного числа мужского рода, а глаголы, причастия и деепричастия в любой̆ форме — к форме глагола в инфинитиве. Лемматизация требует больших временных затрат, а также заставляет хранить большие словари. Это оправдано для языков, относящихся к агглютинативным или синтетическим.

***Лонгитюдный дизайн.*** Состоит из повторяемых кросс-секционных опросов для установления изменений во времени. Делится на панельные исследования (в повторяемых опросах принимают участие одни и те же люди) и когортные исследования (в повторяемых опросах принимают участие разные группы людей, которые представляют одну и ту же генеральную совокупность).

***Макроконтекст***. Языковое окружение данной единицы, выходящее за рамки предложения. Точные границы макроконтекста указать нельзя — это может быть контекст абзаца, главы или даже всего произведения в целом.

***Машинное обучение*** (англ. machine learning, ML). Математическая область, связанная с построением предсказательных алгоритмов (как правило представленных статистическими моделями) на основе данных.

***Международный стандартный сериальный номер*** (англ. *International Standard Serial Number — ISSN*) — уникальный номер, позволяющий идентифицировать

любое периодическое или сериальное издание независимо от того, где оно издано, на каком языке, на каком носителе. Состоит из 8 цифр. Восьмая цифра — контрольное число, рассчитываемое по предыдущим 7 и модулю 11. Для транслитерирования кириллических букв в латинские используется международный стандарт ISO 9 1995 года. Этот стандартный серийный номер широко используется во всём мире: он необходим библиотекам, подписным агентствам, исследователям и учёным, работающим в области информации, новостным агентствам и т.д.

***Метаданные*** (анг. metadata). Иинформация о другой информации, или данные, относящиеся к дополнительной информации о содержимом или объекте. Метаданные раскрывают сведения о признаках и свойствах, характеризующих какие-либо сущности, которые позволяют автоматически искать и управлять ими в больших информационных потоках. Метаданные могут являться информацией о текстах, документах в корпусе или других структурных элементах корпуса: например, год публикации, имя автора, издательство, среда (письменная, устная), реестр (формальный, неофициальный) и т. д. Метаданные могут рассматриваться как контекст текстовой или нетекстовой модальности и быть объектом анализа.

***Метод контекстного поиска***. Совокупность моделей и алгоритмов реализации отдельных технологических этапов: построения поискового образа запроса (ПОЗ), отбора документов (сопоставление поисковых образов запросов и документов), расширения и реформулирования запроса, локализации и оценки выдачи.

***Микроконтекст***. Непосредственное лингвистическое окружение слова в словосочетании или предложении.

***Модальность данных***. Принадлежность данных к некоторому источнику данных, определяющему структуру упомянутых данных, формат, структурно-функциональные взаимосвязи и/или процедуры.

***Модальность изображения***. Фундаментальная визуальная характеристика изображения, которую можно использовать для повышения эффективности поиска. Однако аннотации или подписи, связанные с изображениями, часто не содержат информации о модальности.

***Мониторинг***. Наблюдение за состоянием информационного пространства, охватывающего все компоненты ИТ-отрасли. Целесообразно различать первичный мониторинг (технология оперативного сбора и классификации данных) и вторичный мониторинг (технология прогнозирования и стратегических маркетинговых исследований).

***Мониторинг электронных баз данных***. Непрерывный контроль за состоянием электронных баз данных (ЭБД), принятие решений по развитию ЭБД, принятие решений по изменению данных в ЭБД.

***Мультимедиа***. Интерактивная технология интеграции различных видов информации – текстовой, графической, звуковой, анимационной и пр.

***Научная деятельность.*** 1. Творческая деятельность, направленная на получение новых знаний о природе, человеке и обществе и на использование научных знаний и новых способов их применения в интересах научно-технического прогресса, экономического благосостояния, гуманитарного сотрудничества, культурного и нравственного развития, обеспечения здоровья людей, безопасности их жизнедеятельности и сохранения окружающей среды. 2. Специфический вид когнитивной активности, предметом которой является множество любых возможных объектов (эмпирических и теоретических), целью — производство научного знания о свойствах, отношениях и закономерностях этих объектов, средствами — различные методы и процедуры эмпирического и теоретического исследования.

***Научное исследование***. Процесс выработки новых научных знаний. Основными компонентами исследования являются: постановка задачи, предварительный анализ информации, условий и методов решения задач данного класса; формулировка исходных гипотез; теоретический анализ гипотез; планирование и организация эксперимента; анализ и обобщение полученных результатов; проверка исходных гипотез на основе полученных фактов; окончательная формулировка новых фактов и законов; получение объяснений или научных предсказаний; внедрение полученных результатов в производство.

***Научная коммуникация***. Совокупность видов профессионального общения в научном сообществе, один из главных механизмов развития науки, способа осуществления взаимодействия исследователей и экспертизы полученных результатов.

***Неслово*** (анг. nonword). Токен, который нельзя отнести к словам и выражениям.

***Онтология***. Всеобъемлющая и детальная формализация некоторой области знаний с помощью концептуальной схемы. Обычно такая схема состоит из иерархической структуры данных, содержащей все релевантные классы объектов, их связи и правила (теоремы, ограничения), принятые в этой области.

***Основной (эталонный) корпус*** (анг. focus corpus). Корпус, из которого извлекаются ключевые слова и термины.

***Открытая наука***. 1. Зонтичный термин для движения, цель которого — сделать научные исследования, данные и их распространение доступными для всех уровней заинтересованного общества, будь то любители или профессионалы. 2. Парадигма, движение, целью которого является изменение исследовательской культуры во всем мире как в ценностном, так в и техническом отношении. Парадигма открытой науки подразумевает полную прозрачность всего

исследовательского процесса, а не только возможность свободно знакомиться с его результатами в форме статей в журналах или книг.

***Открытый доступ*** (англ. Open access (OA)). Бесплатный, быстрый, постоянный, полнотекстовый доступ в режиме реального времени к научным и учебным материалам, реализуемый для любого пользователя в глобальной информационной сети без каких-либо ограничений как по инструментам доступа, так и по дальнейшему использованию (например, лицензионных). Режим открытого доступа может применяться ко всем формам опубликованных результатов исследований, в том числе к статьям рецензируемых и не рецензируемых научных журналов, материалам конференций, текстам научных диссертаций и квалификационным работам, главам книг и монографиям.

***Относительная частота, частота на миллион*** (анг. relative frequency, frequency per million, freq/mill). Количество вхождений (совпадений) элемента на миллион, также называемое i.p.m. (экземпляров на миллион). Относительная частота используется для сравнения частот в корпусах разного размера. Относительная частота равна количеству совпадений, приходящихся на корпус размером в миллион токенов. Частота на миллион всегда связана со всем корпусом или подкорпусом текстов, а не с типом текста.

***Относительная частота типа текста*** (анг. relative text type frequency). Показатель, соотносящий частоту в конкретном типе текста с частотой во всем корпусе. Он показывает, насколько типично слово (слова) определенного типа текста, например устной части корпуса или определенного веб-сайта, с которого были загружены тексты. Частоту можно интерпретировать как то, насколько чаще (реже) возникает результат запроса в этом типе текста по сравнению со всем корпусом. Менее 100% означает, что в данном типе текста термин встречается реже, чем во всем корпусе, не является типичным или специфическим для данного типа текста.

***Перечень выходных сведений, размещаемых в непериодических печатных изданиях***. Сведения об авторах и других лицах, участвовавших в создании издания. Заглавие (название) издания. Надзаголовочные данные. Подзаголовочные данные. Выходные данные (место выпуска издания, название издательства или издающей организации, год выпуска издания). Авторский знак. Классификационные индексы универсальной десятичной классификации (УДК) и библиотечно-библиографической классификации (ББК). Аннотация на книгу. Знак охраны авторского права. Международный стандартный номер книги (ISBN). Выпускные данные (номер лицензии на издательскую деятельность и дата ее выдачи; дата подписания в печать, сдачи в набор, вид, номер, формат и доля листа бумаги; гарнитура шрифта основного текста; вид печати; объем издания в условных печатных листах; объем издания в учетно-издательских листах; тираж; номер заказа полиграфического предприятия; название и полный

почтовый адрес издательства или издающей организации; название и полный почтовый адрес полиграфического предприятия).

***Периодическое печатное издание***. Газета, журнал, альманах, бюллетень, иное издание, имеющее постоянное название, текущий номер и выходящее в свет не реже одного раза в год.

***Перплексия***. Мера качества моделей, показывает, насколько хорошо вероятностная модель предсказывает тему, может быть использована для сравнения вероятностных моделей. Низкая перплексия указывает на то, что распределение вероятностей хорошо подходит для прогнозирования выборки.

***Печатное издание***. Печатным изданием признается газета, книга, журнал, брошюра, альбом, плакат, буклет, открытка, иное изделие полиграфического производства независимо от тиража и способа изготовления, предназначенное для передачи содержащейся в нем информации.

***Подкорпус*** (анг. subcorpus). Корпус можно разделить на неограниченное количество частей, называемых подкорпусами. Подкорпус может использоваться для разделения корпуса по типу (художественная литература, газета), средствам массовой информации (устная, письменная) или времени (например, по годам) или по любым другим критериям.

***Пользователь (потребитель) информации***. Субъект, обращающийся к информационной системе или посреднику за получением необходимой ему информации и пользующийся ею.

***Портал***. Комплексный веб-сайт, предназначенный для предоставления интегрированной информации. Обычно содержит ссылки на другие сайты, содержание которых отвечает интересам посетителя портала.

***Предметно-тематический тренд***. Понятие, используемое для обозначения динамики развития терминологической базы междисциплинарных научных направлений в результате изменения контекста использования терминов. Предметно-тематический тренд характеризует тенденции и связи контекстного знания в междисциплинарной среде. Предметно-тематические тренды в совокупности с тезаурусом предметной области могут служить для выявления формирующихся перспективных тематик и научных областей исследования.

***Регуляризация***. В статистике, машинном обучении, теории обратных задач – метод добавления некоторой дополнительной информации к условию с целью решить некорректно поставленную задачу или предотвратить переобучение.

***Результат исследования***. Совокупность теоретических положений и практических рекомендации, полученных в научно-исследовательской работе. Этот результат должен быть представлен таким образом, чтобы он мог быть использован в научной и практической деятельности, раскрыт с содержательной

и внутренне связанной с ней ценностной стороны. Только при таком условии новые знания могут быть включены в общенаучный фонд, взяты на вооружение научными и практическими работниками.

***Релевантность.*** 1. Соответствие между желаемой и действительно получаемой информацией. 2. Релевантность можно еще представить как меру близости между реально полученными документами и тем, что следовало бы получить из системы. 3. Степень соответствия полученной пользователем информации его ожиданиям.

***Репозиторий***. См. Хранилище.

***Реферат***. Вторичный документ, результат аналитико-синтетической переработки информации, представляющий собой краткое изложение содержания первичного документа, включая его основные практические выводы и сведения.

***Сводные каталоги***. Библиографические картотеки, издания или базы данных, отражающие фонды нескольких организаций и указывающие на место нахождения этих документов.

***Сеть.*** Комбинация компьютеров и других устройств, связанных таким образом, чтобы пользователи могли обмениваться данными и программами, совместно пользоваться ресурсами и общаться между собой.

***Слово*** (анг. word). Тип токена, начинающегося с буквы алфавита.

***Сообщество***. Сообщество представляет собой интерактивное объединение людей, организаций или обществ, которые имеют общие интересы, идеи и цели и которые связаны друг с другом через информационно-коммуникационные технологии.

***Сплошной поиск***. Библиографический поиск, при котором обследуется все источники, относящиеся к теме поиска без исключения. В современных условиях автоматизированный и сплошной поиск часто отождествляются.

***Справочный корпус*** (анг. reference corpus). Корпус, который используется при извлечении ключевых слов и терминов. Справочный корпус – это корпус, который сравнивается с эталонным корпусом.

***Стемминг***. Процесс нахождения основы слова путем отбрасывания окончания или применением других, более сложных вариантов. На практике чаще всего применяется стемминг.

***Стоп-слова***. Слова, встречающиеся почти во всех документах, так как они, скорее всего, не будут характеризовать тему. Это так называемые слова общей лексики. К ним относятся предлоги, союзы, местоимения, числительные, прилагательные, некоторые глаголы и наречия.

***Структура***. Под структурой корпуса понимаются сегменты или части, на которые можно разделить корпус с целью проведения анализа. Структура определяется возможностями, задаваемыми теми или иными аналитическими инструментами. Фактически структура корпуса отражает типологию контекстов, принятую в качестве стандарта в рамках конкретного инструмента – программы или среды. Обычно корпус делится на предложения, абзацы и документы, но в отдельных инструментах автор корпуса может вводить различные другие под структуры, чтобы позволить анализу сосредоточиться на меньших или больших частях корпуса.

***Текстовые данные***, также ***текстовый формат***. Представление информации строкового типа (то есть, последовательности печатных символов) в вычислительной системе.

***Телеконференция***. Метод проведения дискуссий между удаленными группами пользователей. Она осуществляется в режиме реального времени или просмотра документов.

***Тема***. Совокупность слов, которые имеют тенденцию встречаться совместно в одних и тех же текстах (компьютерная лингвистика). Такая интерпретация темы позволяет сформулировать лингвистическую модель генерации контента документов коллекции и на основании модели разработать алгоритм вычисления распределения документов и слов по темам. Синонимично «теме», как правило для других предметных областей или методов может быть использовано понятие «семантическая группа», которое в целом имеет то же назначение.

***Тематическое моделирование***. Способ построения модели коллекции текстовых документов с целью определить к каким темам относится тот или иной документ.

***Термин***. Многословное выражение (состоящее из нескольких токенов), которое чаще встречается в одном корпусе (корпусе фокуса) по сравнению с другим корпусом (корпусом ссылок), и в то же время выражение имеет формат термина в язык. Формат определяется грамматикой терминов, специфичной для каждого языка. Термин грамматика обычно фокусируется на определении словосочетаний с существительными. Извлеченные термины типичны для содержания корпуса и могут использоваться для определения темы корпуса.

***Термин-концепт.*** Ключевые слова или термины, отнесённые к терминологическому ядру предметной области или направления исследований.

***Тезаурус.*** В общем смысле — специальная терминология, более строго и предметно — словарь, собрание сведений, корпус или свод, полномерно охватывающие понятия, определения и термины специальной области знаний или сферы деятельности, что должно способствовать правильной лексической, корпоративной коммуникации (проще говоря — пониманию в общении и взаимодействии лиц, связанных одной дисциплиной или профессией); в

современной лингвистике — особая разновидность словарей общей или специальной лексики, в которых указаны семантические отношения (синонимы, антонимы, паронимы, гипонимы, гиперонимы и т. п.) между лексическими единицами. Таким образом, тезаурусы, особенно в электронном формате, являются одним из действенных инструментов для описания отдельных предметных областей. В отличие от толкового словаря, тезаурус позволяет выявить смысл не только с помощью определения, но и посредством соотнесения слова с другими понятиями и их группами, благодаря чему может использоваться для наполнения баз знаний систем искусственного интеллекта. 1. Полный систематизированный набор данных о какой-нибудь области знаний, который разрешает человеку или вычислительной машине в ней ориентироваться. Тезаурус — совокупность терминов, описывающих данную предметную область, с указанием семантических отношений (связей) между ними. Заметим, что кроме классификации других ресурсов, тезаурус может быть создан и использоваться как самостоятельная база знаний, показывая место тех или иных понятий в предметной области. 2. Словарь с дополнительной информацией о связях терминов, таких как синонимы, омонимы, родовидовые отношения, часть/целое, т. е. чаще всего используются некоторые виды семантических отношений. 3. Упорядоченный управляемый словарь, структурированный таким образом, что в нем между терминами четко определены и идентифицированы отношения эквивалентности, гомографии, иерархии и ассоциации с использованием стандартизированных индикаторов этих отношений… Первичная задача тезаурусов заключается в том, чтобы облегчить поиск документов и достигнуть согласованности в выполнении индексации письменных или другим способом полученных документов.

***Терминологический ландшафт***. Многообразие термин-концептов и синонимичных им фраз и терминов, относящихся к определенной предметной области или направлению исследований.

***Терминологическое ядро***. Совокупность базовых термин-концептов, относящихся к определенной предметной области или направлению исследований.

***Токен*** (анг. token). Наименьшая единица, из которой состоит корпус. Есть два типа токенов: слова и неслова. В корпусах больше токенов, чем слов. Пробелы — это не токены. К токенам обычно относят: словоформы – собираюсь, деревья, Дмитрий, двадцать пять ...; знаки пунктуации – запятая, точка, вопросительный знак, кавычки …; цифры: 50 000…; сокращения, наименования продуктов: 3M, i600, XP, FB … .

***Транзакция***. В коллекции текстов – это вхождение слова в документ, взаимодействие/взаимосвязь двух или более слов (объектов) в документе.

***Тренд*** (анг. trend). Функция для обнаружения слов, частота использования которых меняется во времени (диахронический анализ). Тренды определяются словами, употребление которых со временем увеличивается или уменьшается. Тренды могут использоваться для выявления новых слов (неологизмы), для определения момента времени, когда слово начало использоваться, перестало использоваться или когда его употребление стало необычно увеличиваться или уменьшаться. Тренды будут работать только с корпусом, имеющим временные метки, то есть с корпусом, документы которого снабжены датой (публикации).

***Фактографический поиск.*** Синтезирование фактографических описаний.

***Хранилище*** (репозиторий, анг. repository). Место, где хранятся и поддерживаются какие-либо данные. Чаще всего данные в репозитории хранятся в виде файлов, доступных для дальнейшего распространения по сети.

***Эвристики*** (англ. heuristics). Правила, придуманные экспертами для решения научных, инженерных и прикладных задач. Примером эвристики может служить использование простых правил вместо сложных научных решений или построения сложных моделей машинного обучения.

***Экспертные системы*** (англ. expert systems). Прикладное направление ИИ, развивающее методы построения баз знаний и правил, явным образом описывающих знания экспертов.

***Экспликация***. Уточнение понятий и утверждений естественного и/или научного языка с целью устранения выявленных в них неясностей и неточностей применительно к объекту научного исследования. Экспликация может пониматься как замена или только как уточнение исходных понятий через уточнение их места в понятийной системе выбранной предметной области.

***Электронная библиотека***. Распределенная информационная система, позволяющая надежно сохранять и эффективно использовать разнородные коллекции электронных документов (текст, графика, аудио, видео и т.д.) через глобальные сети передачи данных в удобном для конечного пользователя виде.

***Электронно-вычислительная машина*** (ЭВМ). Электронное, программно-управляемое устройство, предназначенное для вычислений. В настоящее время назначение ЭВМ понимается шире – для обработки информации.

***Электронный документ***. 1. Совокупность данных, объединенная смысловым содержанием, представленная в электронной форме и предназначенная для восприятия с помощью соответствующих программных и аппаратных средств. 2. Носитель информации и записанные на нем данные, предназначенные для использования человеком. Документ характеризуется жестким позиционным закреплением различных элементов данных.

***Электронный каталог.*** Машиночитаемый библиотечный каталог, работающий в реальном времени и предоставленный в пользование читателям. Электронный каталог обеспечивает одновременный многоаспектный оперативный поиск.

***Эпизодический (выборочный) поиск***. Поиск, при котором, отталкиваясь от главных источников или известных фактов, постепенно расширяется круг поиска путем привлечения все новых источников.

Кононова Ольга Витальевна
Прокудин Дмитрий Евгеньевич

# ТЕХНОЛОГИИ ИЗВЛЕЧЕНИЯ И ИНТЕЛЛЕКТУАЛЬНОГО АНАЛИЗА ДАННЫХ В НАУЧНЫХ ИССЛЕДОВАНИЯХ

**Учебное пособие**

В авторской редакции
Редакционно-издательский отдел Университета ИТМО
Зав. РИО Н.Ф. Гусарова
Подписано к печати
Заказ №
Тираж
Отпечатано на ризографе

**Редакционно-издательский отдел**
**Университета ИТМО**
197101, Санкт-Петербург, Кронверкский пр., 49, литер А