РАНХиГС
РОССИЙСКАЯ АКАДЕМИЯ НАРОДНОГО ХОЗЯЙСТВА
И ГОСУДАРСТВЕННОЙ СЛУЖБЫ
ПРИ ПРЕЗИДЕНТЕ РОССИЙСКОЙ ФЕДЕРАЦИИ

ИГСУ
РОССИИСКАЯ АКАДЕМИЯ НАРОДНОГО ХОЗЯИСТВА
И ГОСУДАРСТВЕННОИ СЛУЖБЫ
ПРИ ПРЕЗИДЕНТЕ РОССИИСКОИ ФЕДЕРАЦИИ

Технологии анализа данных

Домрачев С.А., доцент,
кандидат технических наук

Цели анализа данных

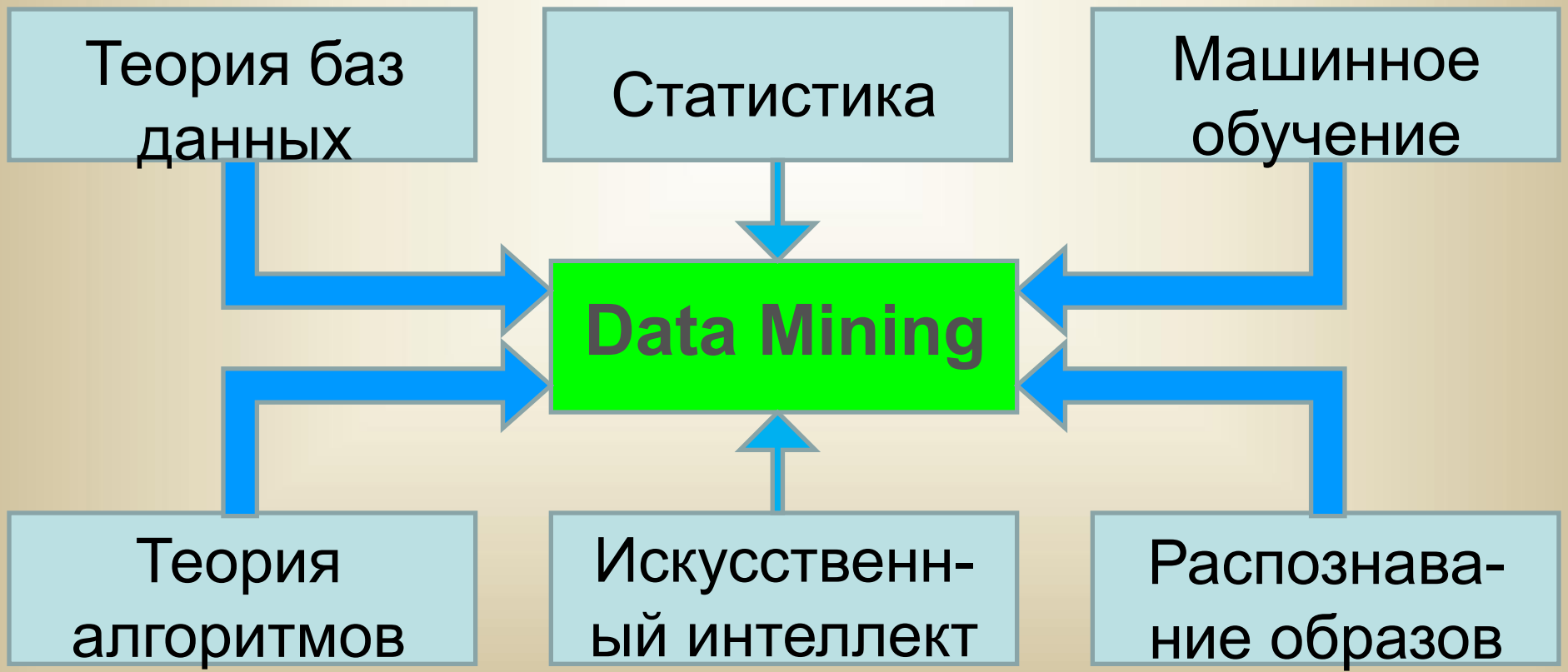
1. Выявление (подтверждение, корректировка) закономерности в поведении социального объекта (явления, процесса)
2. Объяснение на основе выявленной закономерности поведения социального объекта (явлении, процесса)
3. Предсказание его поведения в будущем

Интеллектуальный анализ данных

Процесс аналитического исследования больших массивов необработанных данных в целях выявления скрытых закономерностей и систематических взаимосвязей между ними, для применения к новым совокупностям данных

Понятие Data Mining

Data Mining - мультидисциплинарная область знаний, нацеленная на «раскопку» полезных данных в больших массивах необработанной информации



Методы и алгоритмы Data Mining

К методам и алгоритмам Data Mining можно отнести следующие:

- искусственные нейронные сети
- деревья решений
- кластерный анализ
- поиск ассоциативных правил
- эволюционное программирование (генетические алгоритмы)
- методы визуализации данных

и множество других…

Классификация стадий Data Mining

Состоит из трех стадий:

- Выявление закономерностей (свободный поиск)
- Использование выявленных закономерностей для предсказания неизвестных значений (прогностическое моделирование)
- Анализ исключений, для выявления и толкования аномалий в найденных закономерностях

Стадия свободного поиска

Осуществляется извлечение полезной информации из первичных данных и преобразование ее в некоторые формальные конструкции, обуславливающие имеющиеся закономерности

Состоит из следующих действий :

- выявление закономерностей условной логики

  применяются индукции правил условной логики для классификации и кластеризации (описание в компактной форме близких или схожих групп объектов)

- выявление закономерностей ассоциативной логики

  установление логических ассоциаций для последовательного извлечения при их помощи полезной информации

- выявление трендов и колебаний

  сбор исходных данных для задачи прогнозирования

Стадия прогностического моделирования

Использует результаты предыдущей стадии непосредственно для прогнозирования новых результатов, основанного на анализе прецедентов

Состоит из следующих действий :

- предсказание неизвестных значений
- прогнозирование развития процессов

Т.о. можно получить новое знание о некотором объекте или же группе объектов на основании:

1. знания класса, к которому принадлежат исследуемые объекты
2. знания общего правила, действующего в пределах данного класса объектов

Анализ исключений

Предназначен для выявления и формализации аномалий (отклонений), в найденных на предыдущих стадиях закономерностях

Пример:

Найдено правило - "Если возраст > 35 лет и желаемый уровень вознаграждения > 1200 условных единиц, то в 90 % случаев соискатель ищет руководящую работу"

Возникает вопрос - к чему отнести оставшиеся 10 % случаев?

Возможны два варианта:

1. существует некоторое логическое объяснение, которое также может быть оформлено в виде нового правила
2. оставшиеся 10% - это ошибки исходных данных, следует исправить (очистить) первичные данных

Разведочный анализ данных

Применяется:

- при отсутствии или недостаточности предварительной информации о природе связей;
- при необходимости учета и сравнения большого количества исходных данных;

Используется:

- корреляционный и регрессионный анализ;
- факторный и дискриминантный анализ;
- исчисление индексов и коэффициентов;
- анализ временных рядов и др.

Реализуется:

- программный пакет Statistica;
- программный пакет SyStat;
- программный пакет Stadia; и др.

Использование нейронных сетей

С методологической точки зрения:

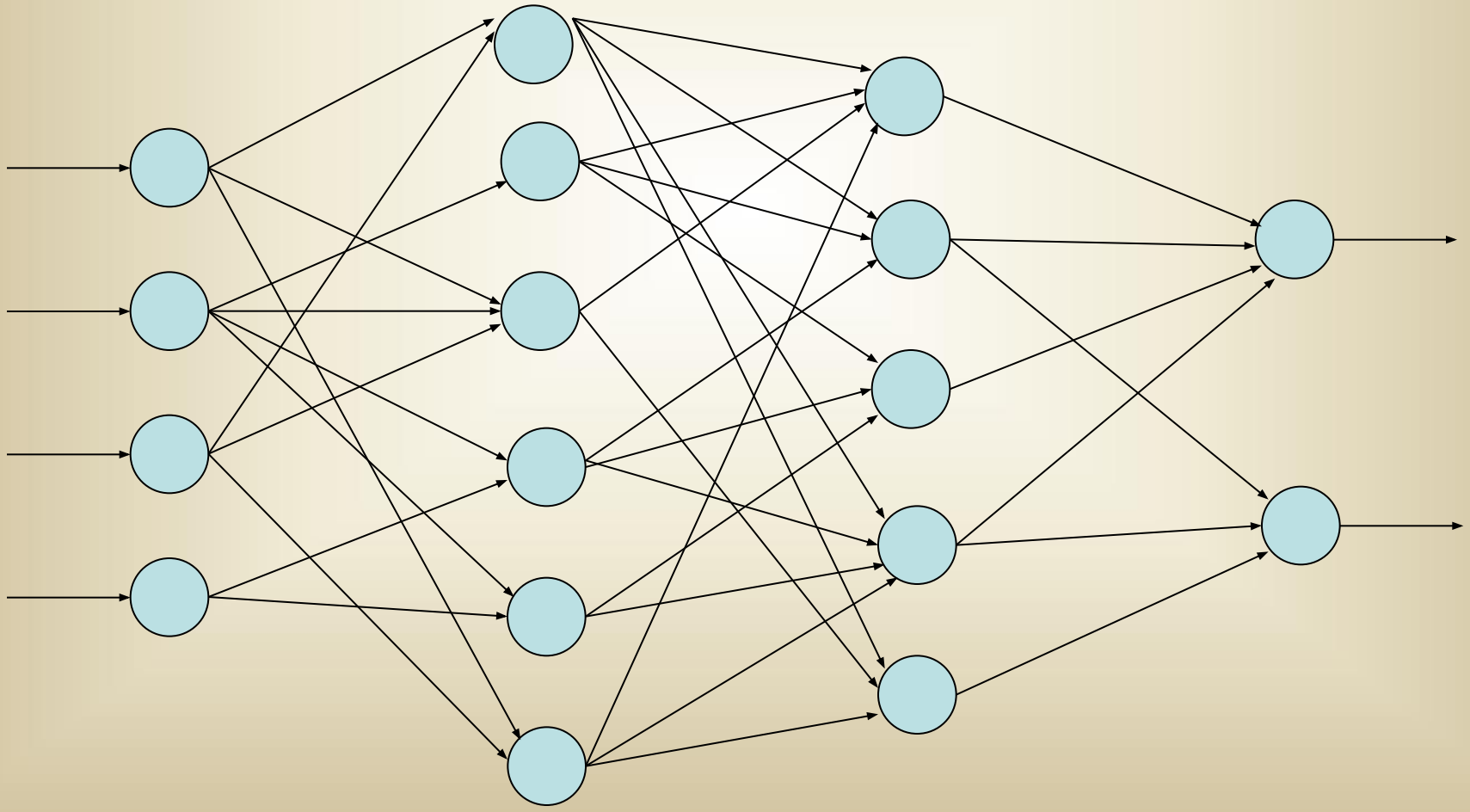
Класс аналитических методов, построенных на принципах обучения мыслящих существ и функционирования мозга, что позволяет прогнозировать значения некоторых переменных в новых ситуациях по данным имеющихся наблюдений

С точки зрения реализации:

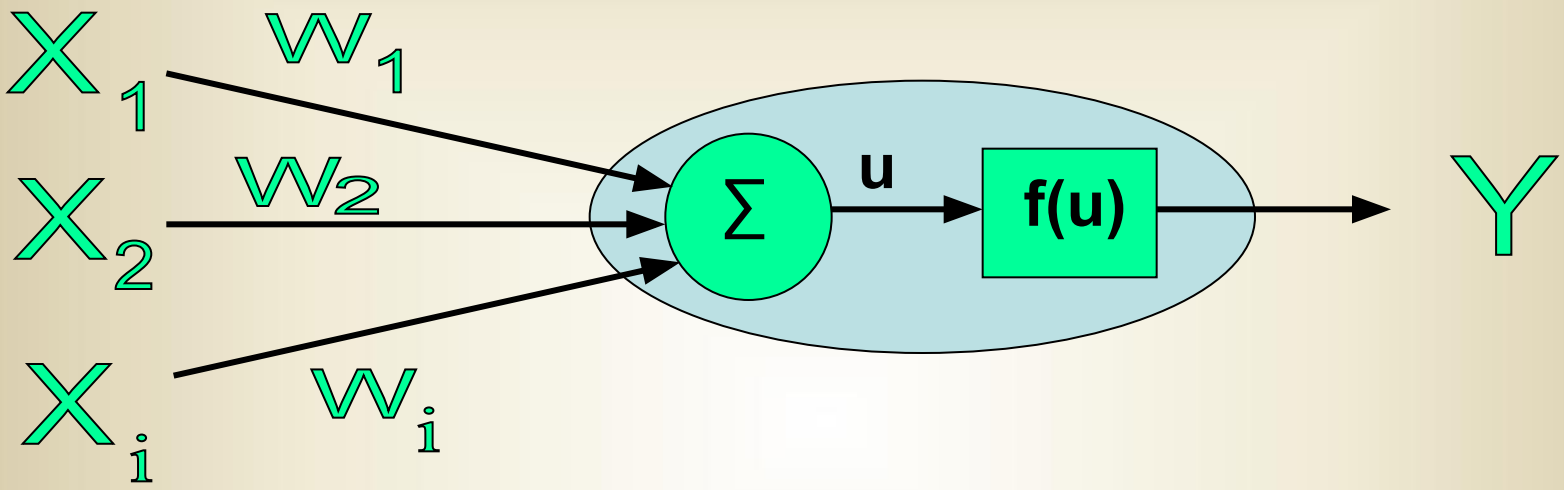
Компьютерная программа, результат работы которой зависит от результата функционирования большого количества однотипных элементов – нейронов (подпрограмм), обладающих некоторыми свойствами и признаками

Построение нейронных сетей

Входной слой **Скрытые слои** **Выходной слой**



Принцип функционирования нейронов



Таким образом, передаточная функция имеет вид:

$$Y = f\left(\sum Wi * Xi\right)$$

где,

Xi – значение входного признака;

Y – значение выходного признака;

Wi – вес входного признака, отражающий степень его влияния на выходной

Инструментальные средства

Для разработки и применения нейронных сетей используются:

- программный пакет NeurOn-line GENSYM
- NeuralWorks Professional II/Plus NeuralWare
- FOREX-94 Уралвнешторгбанк

и др.

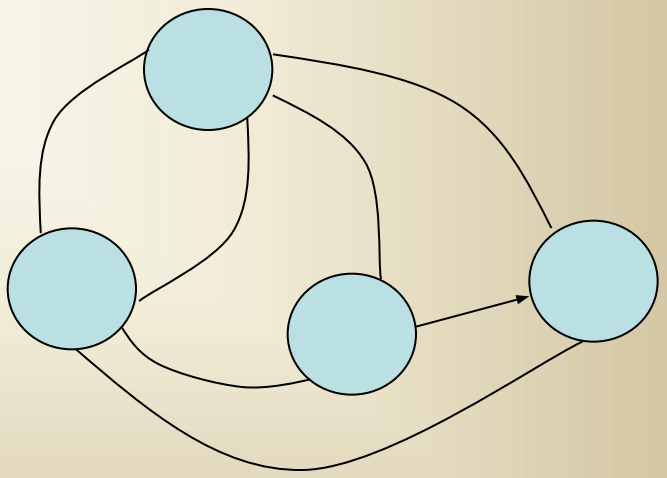
Когнитивное моделирование

Представляет собой структурно-параметрическую формализацию социально-экономических и политических процессов

Выражается в виде ориентированного графа

Вершины графа – существенные факторы, определяющие динамику развития исследуемого процесса

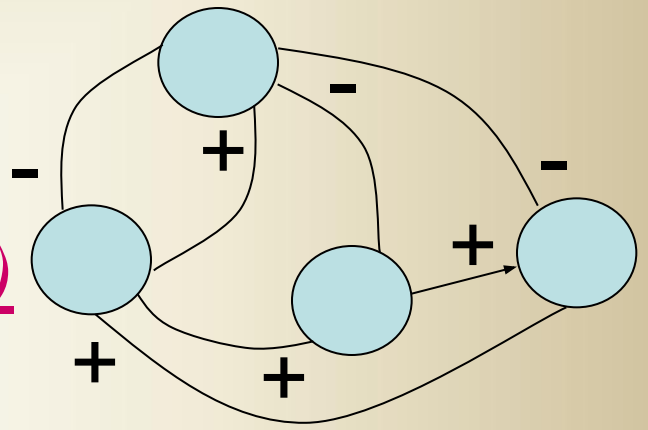
Дуги графа – непосредственные причинно-следственные отношения между факторами



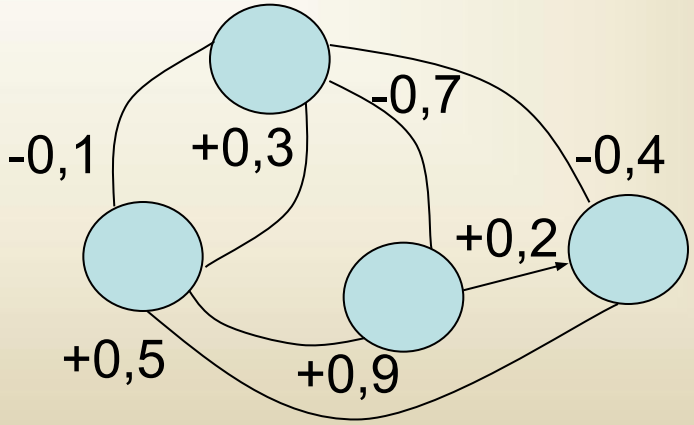
Особенности структурного представления

Для повышения адекватности когнитивных моделей изменяют качество орграфа:

Знаковый граф (когнитивная карта)



Взвешенный граф



Функциональный граф

Методы анализа текстовых документов

РАНХиГС
РОССИЙСКАЯ АКАДЕМИЯ НАРОДНОГО ХОЗЯЙСТВА И ГОСУДАРСТВЕННОЙ СЛУЖБЫ ПРИ ПРЕЗИДЕНТЕ РОССИЙСКОЙ ФЕДЕРАЦИИ

ИГСУ
РОССИЙСКАЯ АКАДЕМИЯ НАРОДНОГО ХОЗЯЙСТВА И ГОСУДАРСТВЕННОЙ СЛУЖБЫ ПРИ ПРЕЗИДЕНТЕ РОССИЙСКОЙ ФЕДЕРАЦИИ

Анализ текстовых документов

Анализ документов позволяет выявить определенные особенности, свойства и взаимосвязи тех или иных явлений и процессов, специфику включения в них различных субъектов социально-экономической и политической жизни, проследить динамику их развития

Анализ символьных данных представляет собой творческий процесс, зависящий от:

- содержания и сложности построения документа
- условий, целей и задач проводимого исследования
- научной квалификации, богатства опыта и творческой интуиции исследователя

Оценка надежности документальной информации

При оценке надежности учитывают следующие факторы:

- является ли документ официальным
- является ли документ личным или безличным
- подвергался ли документ контролю (юридический, финансовый и т.п.)
- тенденциозный характер документа (биографии, мемуары и т.п.)

РАНХиГС
РОССИЙСКАЯ АКАДЕМИЯ НАРОДНОГО ХОЗЯЙСТВА И ГОСУДАРСТВЕННОЙ СЛУЖБЫ ПРИ ПРЕЗИДЕНТЕ РОССИЙСКОЙ ФЕДЕРАЦИИ

ИГСУ
РОССИЙСКАЯ АКАДЕМИЯ НАРОДНОГО ХОЗЯЙСТВА И ГОСУДАРСТВЕННОЙ СЛУЖБЫ ПРИ ПРЕЗИДЕНТЕ РОССИЙСКОЙ ФЕДЕРАЦИИ

Информационно-аналитическая обработка текстов

Технологии автоматического извлечения знаний могут быть сведены к следующим направлениям:

1. классификация
2. кластерный анализ
3. семантическое сжатие текста
4. построение семантических сетей

Классификация текстовых документов

Представляет собой систему рубрицирования текстовых документов, базирующуюся на разделении понятий «тема» и «проблема»

Тема более простая и устойчивая в лексическом плане конструкция, допускающая возможность автоматического распознавания

Проблема более сложная, меняющаяся со временем и обстоятельствами лексическая конструкция, синтезируемая из тематических категорий

Система рубрицирования

обеспечивает:

1. интеграцию разнородной информации
2. профилирование пользователей и проблем
3. проблемно-тематическую навигацию по информационным фондам
4. интерпретацию содержания документов на модели предметной области

обладает свойствами:

1. тематическая полнота, обеспечивающая соотнесение документа соответствующим рубрикам
2. временная устойчивость, дающая возможность ретроспективного сопоставительного анализа текстов
3. компактность представления

Кластерный анализ подборок текстовых документов

Применяется при реферировании больших документальных массивов и выделении компактных подгрупп документов с близкими свойствами

Различают два основных типа кластеризации:

1. иерархический
   построение дендритной структуры, выраженной деревом кластеров, содержащих близкие по смыслу группы документов
2. бинарный
   группировка и просмотр документальных кластеров по ссылкам подобия, основанных на весах и определяемых ключевых словах

Семантическое сжатие текста

Заключается в использовании технологических процедур:

1. индексирование ключевыми словами

анализ смыслового содержания текста для выделения сведений об известных объектах, их свойствах и отношениях между собой с целью создания терминологического портрета документа

2. автоматическое реферирование текстов

квазирефераты – последовательность извлеченных фрагментов текста, наиболее репрезентативно представляющих содержание документа

рефераты-клише – набор извлеченных из текста наиболее информативных слов, которые вставляются в заготовленные шаблоны

3. построение гипертекстовых структур

Построение семантических сетей

Реализует функцию выявления и идентификации ассоциативных и причинно-следственных связей между существенными темами и информационными объектами целевой подборки документов или потока входящих документов

Позволяет автоматизировать решение задач:

- исследование тематического состава подборки документов
- поиск новой, неожиданной информации (фактов) связанной с исследуемым объектом
- выявление в документах подтверждений связей между исследуемыми объектами