Лабораторные работы по дисциплине «Анализ больших данных»
(бакалавры ИС)

СЕМЕСТР 1

Лабораторная работа (ЛР) 1. История и задачи анализа больших данных (АБД) (5 часов)

Цель лабораторной работы. Самостоятельное ознакомление с основами АБД, идеями и принципами, заложенными в его основу.

Задача. Собрать материал по АБД, систематизировать его и изложить в виде реферативного отчета; сдать отчет преподавателю, ответив на вопросы по выбранной теме.

Рекомендуемые направления тем рефератов:

1. Содержание АБД.
2. История АБД.
3. Задачи и методы АБД.
4. Статистические методы АБД.
5. Кластерный анализ в АБД.
6. Нейронные сети в АБД
7. Методы искусственного интеллекта в АБД
8. Методы хранения данных в АБД.
9. Программные системы АБД
10. Тема по выбору, относящаяся к изучаемой дисциплине

Объем реферата не более 10-15 стр.

Порядок выполнения.

1. Пользуясь поисковыми машинами интернет собрать материал по выбранной теме (не менее 3-4 источников).
2. Ознакомиться с источниками, систематизировать полученную информацию и изложить в форме реферата или эссэ (эссе – прозаическое сочинение небольшого объема и свободной композиции, выражающее индивидуальные впечатления и соображения по конкретному поводу или вопросу и не претендующее на определяющую или исчерпывающую трактовку предмета). Приветствуется выражение собственной позиции по рассмотренному вопросу. Достаточен объем изложения 7-10 страниц текста шрифтом 12пт через 1 интервал. В конце изложения указать список использованных источников.
3. Представить реферат или эссэ в виде отчета по ЛР.

Защитить отчет преподавателю, ответив на дополнительные вопросы:

- Что такое АБД?
- Какие задачи решает АБД?
- Какие методы используются в АБД?
- Откуда берется и как хранится информация для АБД?
- В чем особенности методов хранения информации для АБД?
- Вопросы, связанные с выполненной ЛР.

ЛР 2. Корреляционный и регрессионный анализы (6 часов)

Цель работы. Научиться пользоваться методами статистического анализа данных для решения прикладных задач.

Задача. Выполнить корреляционный анализ двух временных последовательностей, и оценить тенденцию их развития на ближайший период.

Порядок выполнения.

1. Выбрать предметную область
2. Сформировать две временный последовательности (рекомендуется взять реальные последовательности случайных величин из интернет)
3. Рассчитать их корреляцию
4. Построить уравнения регрессии для каждой
5. Оценить тенденцию развития на ближайший период (рост, убывание)
6. Оформить отчет по ЛР, включив туда логические модели для исходной БД и ХД и сопроводив их содержательными описаниями.

Защитить отчет преподавателю, ответив на дополнительные вопросы:

– Что такое случайная величина?
– Что такое корреляция случайных величин?
– Что такое регрессионная модель?
– Вопросы, связанные с выполненной ЛР.

Пример. Для облегчения расчетов можно проводить **корреляционный анализ** в Excel. В данной программе существует ряд инструментов, помогающих облегчить расчеты. Среди них функция «Корреляция», позволяющая сформировать матрицу из коэффициентов и разных параметров. Она изображается в форме таблицы. В качестве столбцов и строк используются корреляционные коэффициенты. На основе полученных данных таблицы необходимо будет провести корреляционный анализ. Пример последовательности проведения анализа:

1. В команде «Сервис» выбрать пункт «Анализ данных».
2. В качестве инструмента анализа выбрать пункт «Корреляция».
3. В появившемся окне в строке «Входной интервал» указать диапазон анализируемых данных, выбрать пункт «Группировка» в строке «Параметры вывода», ввести диапазон вывода результатов и нажать «ОК».

В результате получится корреляционная матрица, расположенная в диапазоне вывода. Внутри будет указан коэффициент линейной корреляции, оценивающий тесноту и форму связи между показателями

В MS-Excel используется функция «Корреляция» для того, чтобы провести корреляционно-регрессионный анализ. Пример расчета коэффициентов рассмотрим далее. Эта функция формирует матрицу с коэффициентами тесноты взаимосвязи между разными параметрами. В итоге формируется квадратная таблица, содержащая коэффициенты корреляции на пересечении строк и столбцов.

Для проведения **регрессионного анализа** необходимо будет выполнить ряд определенных действий:

1. Открыть команду «Сервис», а в ней пункт «Анализ данных».
2. В появившемся окне указать в перечне «Инструменты анализа» пункт «Корреляция».
3. В раскрывшемся окне «Корреляция» указать входной интервал в виде диапазона ячеек, содержащих анализируемую информацию (он должен быть не менее двух столбцов), поставить галочку в пункте «Группировка», а в поле «Параметры вывода» выбрать верхнюю левую ячейку, где будет начинаться корреляционная матрица.

4. Нажать на кнопку ОК.

В результате вычислений появится квадратная таблица с коэффициентами корреляции

Для того чтобы вычислить уравнение линейной регрессии, описывающие взаимосвязь между факторами и результатом, в MS Excel применяется статистическая функция «Линейн». Для того чтобы ее использовать, необходимо:

1. Выделить пустую область, в которую будут выведены результаты анализа.
2. Открыть «Мастер функций», в нем найти категорию «Статистические», а в ней функцию «Линейн» и нажать ОК.
3. В поле «Известные значения y» ввести диапазон анализируемых результатов, в поле «Известные значения x» – диапазон анализируемых факторов.
4. В поле «Константа» указывается присутствие свободного члена уравнения (1 – да, 0 – нет), а в поле «Статистика» – необходимость вывода дополнительных сведений (1 – появится дополнительная информация, 0 – появятся только оценки параметров). По умолчанию можно указывать в обоих полях 1.
5. Нажать кнопку ОК.

Вверху ранее выделенной области появится начальный элемент таблицы. Для того чтобы раскрыть все данные, необходимо нажать F2, а потом одновременно комбинацию клавиш Ctrl + Shift + Enter.

В итоге регрессионная информация будет изображаться в качестве таблицы из двух столбцов и пяти строк (табл. 1):

Таблица 1. Пример регрессионной информации

	Столбец 1	**Столбец 2**
Строка 1	Коэффициент b	Коэффициент a
Строка 2	Среднеквадратическое отклонение b	Среднеквадратическое отклонение a
Строка 3	Коэффициент детерминации	Среднеквадратическое отклонение y
Строка 4	F-статистика	Число степеней свободы
Строка 5	Регрессионная сумма квадратов	Остаточная сумма квадратов

Полученные результаты необходимо подставить в линейное уравнение регрессии, которое выглядит следующим образом: $y = a + bx$. В качестве коэффициента a подставляется значение из ячейки на пересечении строки 1 и столбца 2. В качестве коэффициента b – значение на пересечение строки 1 и столбца 1.

Коэффициент детерминации говорит о том, какая часть результата объясняется с помощью исследуемого фактора. Оставшаяся часть результатов определяется факторами, неучтенными в линейной модели.

ЛР 3. Многомерные базы данных. OLAP-анализ (6 часов)

Цель работы. Научиться проектировать и разрабатывать OLAP-кубы.

Задача. Разработать многомерную базу данных (OLAP-куб) по выбранному направлению (предметной области).

Порядок выполнения.

1. Выбрать предметную область
2. Спроектировать простую реляционную базу данных (БД) из нескольких таблиц в любой СУБД по выбору.
3. Заполнить БД 5-10 записями
4. Спроектировать на основе БД хранилище данных в виде многомерной БД (OLAP-куба, пример прилагается).
5. Оформить отчет по ЛР, включив туда логические модели для исходной БД и ХД и сопроводив их содержательными описаниями.

Защитить отчет преподавателю, ответив на дополнительные вопросы:

- Что такое ХД?
- Чем ХД отличаются от БД?
- Особенности хранения данных для АБД?
- Вопросы, связанные с выполненной ЛР.

Пример. Предметная область – торговля. БД – данные о продаже товаров. Состав БД: таблица продаж, включающая поля: «Код» (продажи), «Дата продажи», «Время продажи», «Код товара», «Количество товара». Справочник: «Код» (товара), «Наименование» (товара), «Стоимость» (товара), «Остаток» (на складе) (рис. 1).

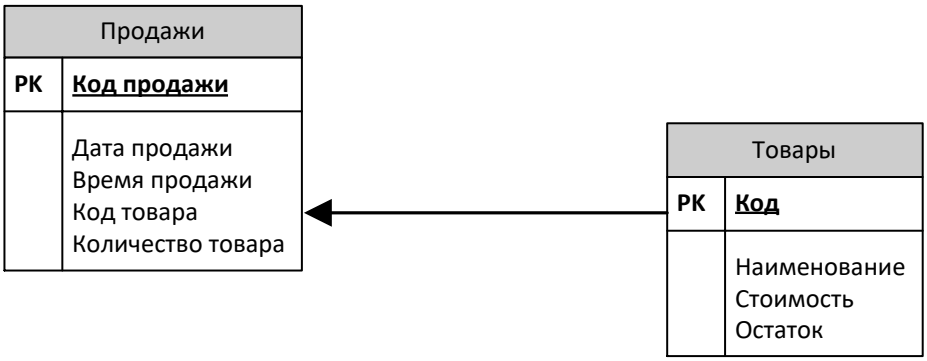

Рис. 1. БД продаж.

Хранилище данных – сведения обо всех продажах, систематизированные по товарам, дате и времени (рис. 2).

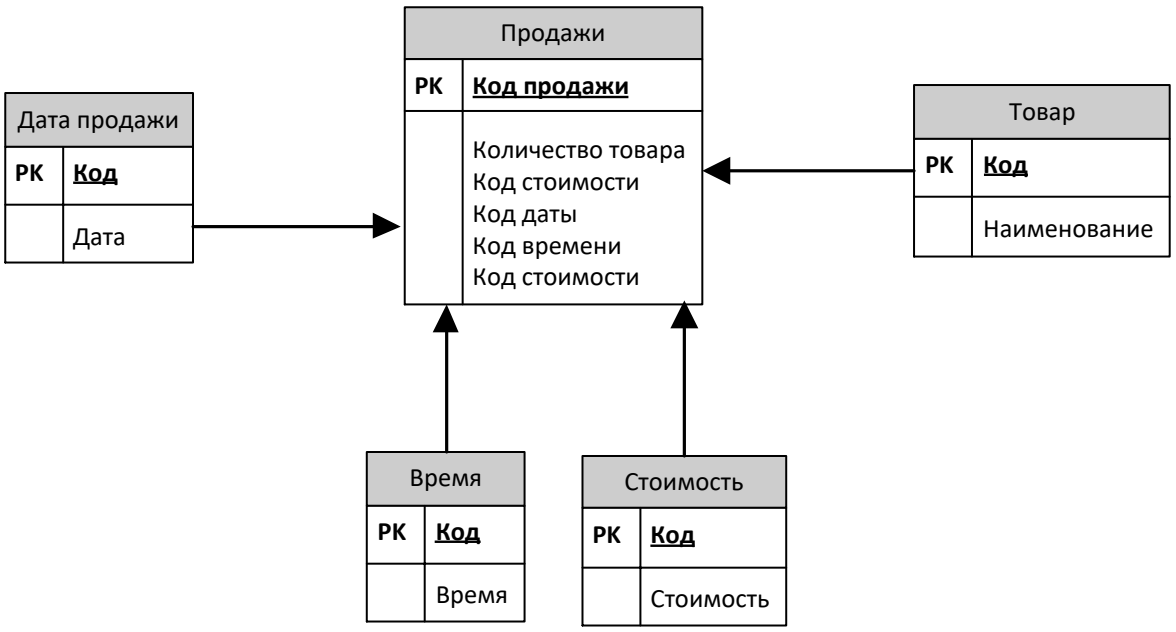

Рис. 2. ХД о продажах.

Хранилище включает основную таблицу, содержащую сведения о характере и количестве проданного товара, дате и времени продажи, стоимости покупки. Дата и время могут быть агрегированными величинами. Например, с точностью до недели или месяца для даты и с точностью до часа или времени суток для времени (хотя целесообразно дату хранить с точностью до дня, как и положено дате, а время – с точностью до часа). В результате такого представления каждая продажа привязана как точка в 4-хмерном пространстве к наименованию товара, дате его продажи, часу продажи, стоимости. Это позволяет строить запросы о проданных товарах в определенные дни, в определенный час, группировать информацию по видам товаров, стоимости, строить запросы о количестве проданного товара конкретного вида в определенный сезон или время суток. Причем анализ может носить агрегированный характер: продажи за сезон, продажи определенного товара или группы товаров в определенные часы определенного сезона, продажи в диапазоне стоимости и т.д.

Следует обратить внимание на менее экономную работу с памятью в ХД по сравнению с БД. За счет этого ускоряется и углубляется анализ.

В качестве еще одного примера рассмотрим 4-х мерный куб, содержащий информацию о количестве **патентов** или **торговых марок** зарегистрированных в определённой **стране** в заданной **предметной области**, за определенный **период времени**. При реализации многомерного хранилища данных в соответствующей реляционной СУБД используется следующая схема (рис. 3):

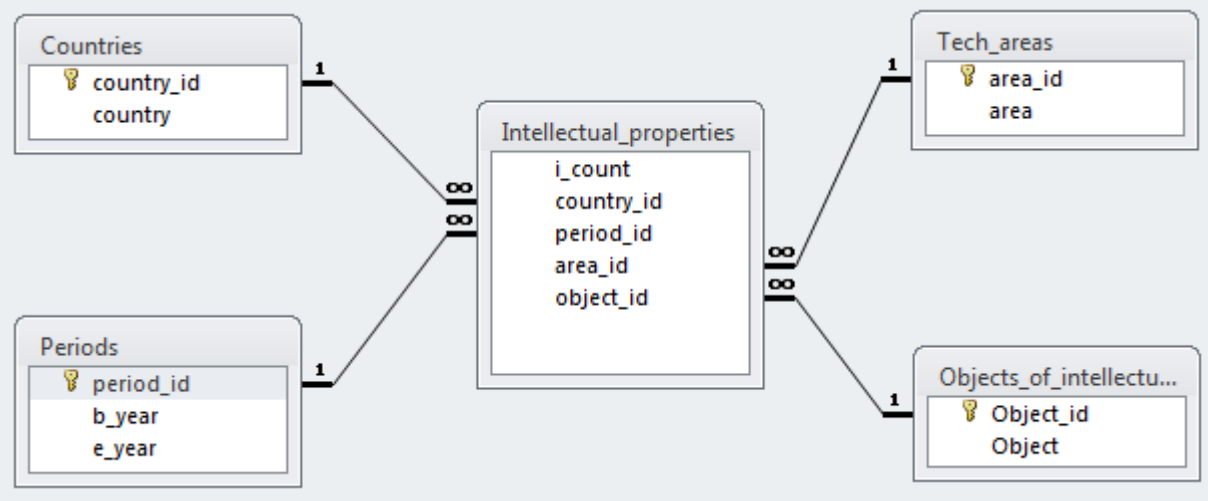

Рис. 3. ХД о патентах.

Данных куб также имеет 4 измерения, представленных таблицами: Tech_areas, Objects_of_intellectual_properties, Countries, Periods. Информация о количестве патентов хранится в центральной таблице intellectual_properties, в поле i_count. Измерения подключаются через внешние ключи к нужным таблицам.

СЕМЕСТР 2

ЛР 4. Классификация и кластеризация данных (6 часов)

Цель работы. Научиться пользоваться методами кластеризации данных для решения прикладных задач.

Задача. Выполнить кластеризацию множества данных, содержащего не менее 3-х кластеров, пользуясь выбранной мерой сходства.

Порядок выполнения

1. Сформировать множество данных в двумерном признаковом пространстве, содержащее не менее 3-х кластеров.
2. Выбрать одну из мер сходства (евклидово расстояния, расстояние городских кварталов, расстояние Чебышёва и т.п.).
3. Выбрать алгоритм кластеризации.
4. Выполнить кластеризацию с помощью выбранной меры и алгоритма.
5. Отобразить кластеры и результаты кластеризации графически (в виде диаграммы)
6. Оформить отчет по ЛР, включив туда исходное и кластеризованные множества, описание алгоритма кластеризации и графическое представление результатов кластеризации.

Защитить отчет преподавателю, ответив на дополнительные вопросы:

- Что такое признаковое пространство?
- Что такое кластер в признаком пространстве?
- Какие существуют меры сходства?
- Какие существуют алгоритмы кластеризации?
- Вопросы по выполненной ЛР.

Пример. Кластеризация на основе алгоритма минимального остовного дерева.

1. Строится полный граф, в котором каждая вершина связана со всеми другими вершинами по принципу «каждый с каждым». Каждая дуга имеет вес, равный расстоянию между соответствующими объектами. Сначала все объекты принадлежат одному кластеру (т.е. алгоритм дивизимный).

2. С помощью алгоритма Прима у полного графа удаляются «лишние» дуги и остаются кратчайшие дуги, входящие в минимальное остовное дерево.

3. В зависимости от выбранной стратегии кластеризации (по пороговому расстоянию или по заданному числу кластеров) у остовного дерева удаляются либо все дуги длиннее порогового значения, либо просто самые длинные дуги, но в таком количестве, чтобы обеспечить разбиение на заданное число кластеров. В результате общий кластер распадается (при соответствующем значении порога) на подкластеры. Пороговый размер дуги задается из условия:

$$R_{пор} = \sqrt[n]{\frac{V}{N}} ,$$

где V –величина n-мерного объема, занятого исходным кластером, N число кластеризуемых объектов, n- размерность признакового пространства. Иначе говоря, $R_{\text{пор}}$ – это среднее расстояние между объектами в признаковом пространстве.

4. Объекты с расстоянием между собой менее порогового составляют один кластер. Объекты удаленные более чем на величину порога – другие кластеры. Для управления работой

алгоритма вводится коэффициент умножения порога, который может принимать значения от 0 до 9,99.

5. Кластеризация выполняется до тех пор, пока все объекты не будут поделены на кластеры.

ЛР 5. Технология экспертных систем в больших данных. Деревья решений (6 часов)

Цель работы. Научиться строить классифицирующие правила на основе результатов сегментации образов.

Задача. Построить набор решающих правил вида «Если *A*, то *B*», где *A* – граница сегмента в виде описанного прямоугольника, *B* – имя сегмента».

Порядок выполнения

1. Сформировать множество данных в двумерном признаковом пространстве, содержащее не менее 3-х кластеров.
2. Выбрать одну из мер сходства (евклидово расстояния, расстояние городских кварталов, расстояние Чебшёва и т.п.).
3. Выбрать алгоритм кластеризации.
4. Выполнить кластеризацию с помощью выбранной меры и алгоритма.
5. Определить границы кластеров как описанных прямоугольников.
6. Представить каждый кластер продукцией вида (координаты признакового пространства *X*, *Y*:

   Если $X^I_{\min} \le x \le X^I_{\max}$ & $Y^I_{\min} \le y \le Y^I_{\max}$, то «Образ принадлежит кластеру *I*»
7. Отобразить результаты графически
8. Пользуясь графическим представлением выполнить классификацию произвольного образа по полученным правилам.
9. Проанализировать достоинства и недостатки такого подхода
10. Оформить отчет по ЛР, включив туда исходное и кластеризованное множества, классифицирующие продукции, пример классификации и краткую характеристику достоинств и недостатков такого подхода. *По возможности дать предложения по его совершенствованию*.

Защитить отчет преподавателю, ответив на дополнительные вопросы:

- Что такое признаковое пространство?
- Что такое кластер в признаком пространстве?
- Как можно определять принадлежность к кластеру?
- Что такое продукция?
- Вопросы по выполненной ЛР.

ЛР 6. Классификация и кластеризация данных с помощью нейронных сетей (6 часов)

Цель работы. Изучить свободно-распространяемую систему АБД «Loginom Community» на примере задачи кластеризации.

Задача. Выполнить кластеризацию числовых данных различными методами и сравнить результаты.

Порядок выполнения

1. Сформировать множество данных в многомерном признаковом пространстве (предлагается кластеризовать множество ирисов Фишера – табл. 2).
2. Освоить систему **Loginom Community**
3. Выполнить кластеризацию тремя методами (описано ниже)
4. Оформить отчет по ЛР, включив таблицы исходного и кластеризованных множества.

Защитить отчет преподавателю, ответив на дополнительные вопросы:

– особенности работы с Loginom Community

– что такое самоорганизующиеся карты Кохонена?

– вопросы по выполненной ЛР.

Описание выполнения ЛР

1. Зайти на сайт https://basegroup.ru/
2. Скачать и установить у себя аналитическую систему Loginom Community (бесплатная версия, для загрузки указать «студент»)
3. Ознакомиться с описанием системы и особенностями работы с ней.
4. Создать в документах каталог(папку), назвав его, к примеру ЛР5
5. С помощью Exel создать в каталоге ЛР5 (или другом, самостоятельно названном) файл обрабатываемых объектов. Например ирисы Фишера (прилагается ниже)
6. Запустить программу Loginom Community и по справке ознакомиться с принципами её работы
7. Выполнить кластеризацию. Для этого:
   - 7.1. Создать пакет в выбранном каталоге
   - 7.2. В окне Компоненты захватить пункт меню Exel файл и перенести(скопировать) его в рабочее окно, указав в параметре Имя файла местоположение Вашего файла * .xls
   - 7.3. Выбрать диапазон числовых данных, сохранив заголовки столбцов данных (в примере B1:F151)
   - **7.4. Установить значение «Исключить» для команды «Пустые строки»**
   - 7.5. Указать количество строк заголовка - 1
   - 7.6. Перейти в след. окно
   - **7.7.** Указать количество строк для анализа согласно обрабатываемы данным (в примере 150). **Обновить данные!**
   - 7.8. (пропустить окно «Настройка соответствия между столбцами»)
   - 7.9. Сохранить результат и выполнить
   - 7.10. Перейти в раздел Data Mining и перенести(скопировать) в рабочее окно вкладку Самоорганизующиеся сети
   - 7.11. Соединить в рабочем окне блок данных с блоком Самоорганизующиеся сети
   - 7.12. Настроить работу сети, для чего:

7.12.1. В окне «Настройка входных столбцов» указать для столбцов с числами значение «Используемое», для столбца с текстом – виды ирисов – оставить «Не задано»)
7.12.2. (пропустить окно «Настройки нормализации»)
7.12.3. В окне «Самоорганизующаяся нейронная сеть» выбрать «Сеть Кохонена», остальное сохранить
7.12.4. Сохранить результат

8. Вернуться в основное окно и выбрать «Переобучить узел» в контекстном меню узла «Самоорганизующиеся сети»
9. В левой части окна программы в разделе «Экспорт» выбрать «Exel файлы» и перенести в рабочее окно
10. Соединить выход модуля обработки с экспортным файлом
11. Задать имя экспортному файлу, отличное от имени исходного файла данных
12. Сохранить и выполнить
13. Ознакомиться с содержимым экспортного файла и сделать вывод о результатах кластеризации
14. Самостоятельно подключить, настроить и апробировать модули «Кластеризация» и «EM Кластеризация», сохранив результаты работы каждого в свой файл
15. Сравнить результаты, сделать выводы
16. СТОП

Процедура обработки представлена на рис. 4

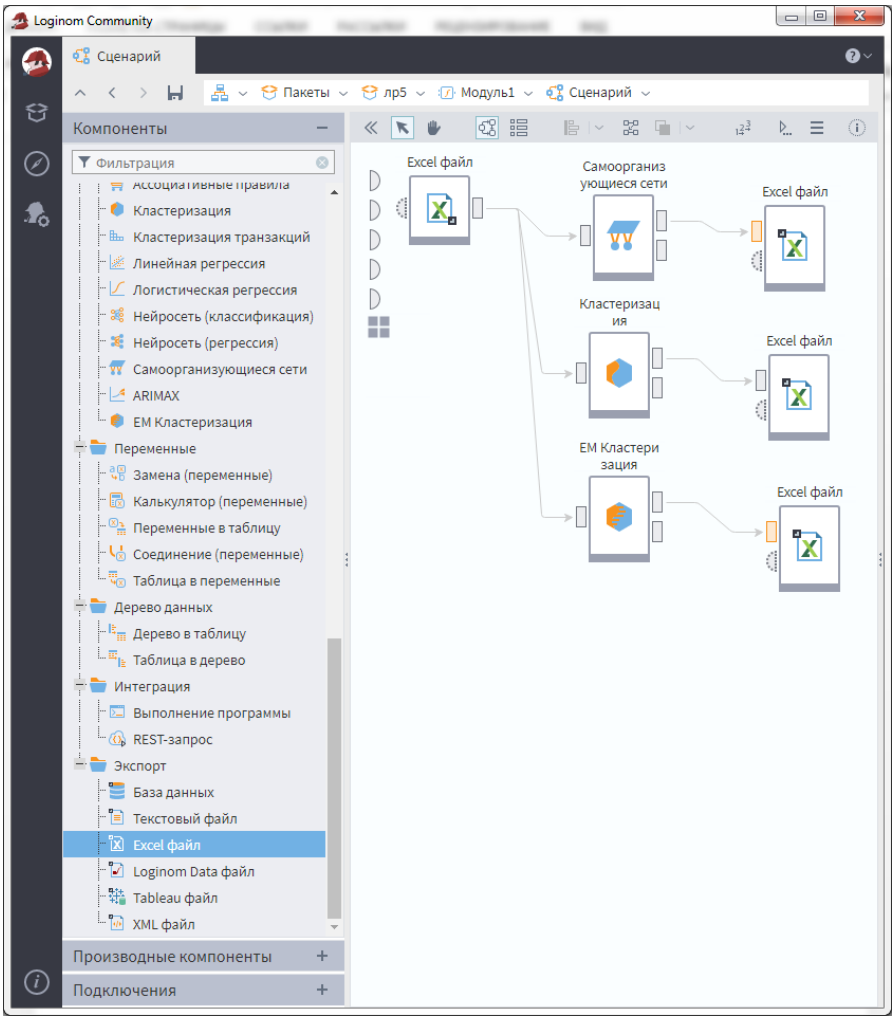

Рис. 4. Схема кластеризации в ЛР5

Таблица 2. Ирисы Фишера

№	Длина чашелистника	Ширина чашелистника	Длина лепестка	Ширина лепестка	Вид
1	5,1	3,5	1,4	0,2	setosa
2	4,9	3,0	1,4	0,2	setosa
3	4,7	3,2	1,3	0,2	setosa
4	4,6	3,1	1,5	0,2	setosa
5	5,0	3,6	1,4	0,2	setosa
6	5,4	3,9	1,7	0,4	setosa
7	4,6	3,4	1,4	0,3	setosa
8	5,0	3,4	1,5	0,2	setosa
9	4,4	2,9	1,4	0,2	setosa
10	4,9	3,1	1,5	0,1	setosa
11	5,4	3,7	1,5	0,2	setosa
12	4,8	3,4	1,6	0,2	setosa
13	4,8	3,0	1,4	0,1	setosa
14	4,3	3,0	1,1	0,1	setosa
15	5,8	4,0	1,2	0,2	setosa
16	5,7	4,4	1,5	0,4	setosa
17	5,4	3,9	1,3	0,4	setosa
18	5,1	3,5	1,4	0,3	setosa
19	5,7	3,8	1,7	0,3	setosa
20	5,1	3,8	1,5	0,3	setosa
21	5,4	3,4	1,7	0,2	setosa
22	5,1	3,7	1,5	0,4	setosa
23	4,6	3,6	1,0	0,2	setosa
24	5,1	3,3	1,7	0,5	setosa
25	4,8	3,4	1,9	0,2	setosa
26	5,0	3,0	1,6	0,2	setosa
27	5,0	3,4	1,6	0,4	setosa
28	5,2	3,5	1,5	0,2	setosa
29	5,2	3,4	1,4	0,2	setosa
30	4,7	3,2	1,6	0,2	setosa
31	4,8	3,1	1,6	0,2	setosa
32	5,4	3,4	1,5	0,4	setosa
33	5,2	4,1	1,5	0,1	setosa
34	5,5	4,2	1,4	0,2	setosa
35	4,9	3,1	1,5	0,1	setosa
36	5,0	3,2	1,2	0,2	setosa
37	5,5	3,5	1,3	0,2	setosa
38	4,9	3,1	1,5	0,1	setosa
39	4,4	3,0	1,3	0,2	setosa
40	5,1	3,4	1,5	0,2	setosa
41	5,0	3,5	1,3	0,3	setosa
42	4,5	2,3	1,3	0,3	setosa
43	4,4	3,2	1,3	0,2	setosa

44	5,0	3,5	1,6	0,6	setosa
45	5,1	3,8	1,9	0,4	setosa
46	4,8	3,0	1,4	0,3	setosa
47	5,1	3,8	1,6	0,2	setosa
48	4,6	3,2	1,4	0,2	setosa
49	5,3	3,7	1,5	0,2	setosa
50	5,0	3,3	1,4	0,2	setosa
51	7,0	3,2	4,7	1,4	versicolor
52	6,4	3,2	4,5	1,5	versicolor
53	6,9	3,1	4,9	1,5	versicolor
54	5,5	2,3	4,0	1,3	versicolor
55	6,5	2,8	4,6	1,5	versicolor
56	5,7	2,8	4,5	1,3	versicolor
57	6,3	3,3	4,7	1,6	versicolor
58	4,9	2,4	3,3	1,0	versicolor
59	6,6	2,9	4,6	1,3	versicolor
60	5,2	2,7	3,9	1,4	versicolor
61	5,0	2,0	3,5	1,0	versicolor
62	5,9	3,0	4,2	1,5	versicolor
63	6,0	2,2	4,0	1,0	versicolor
64	6,1	2,9	4,7	1,4	versicolor
65	5,6	2,9	3,6	1,3	versicolor
66	6,7	3,1	4,4	1,4	versicolor
67	5,6	3,0	4,5	1,5	versicolor
68	5,8	2,7	4,1	1,0	versicolor
69	6,2	2,2	4,5	1,5	versicolor
70	5,6	2,5	3,9	1,1	versicolor
71	5,9	3,2	4,8	1,8	versicolor
72	6,1	2,8	4,0	1,3	versicolor
73	6,3	2,5	4,9	1,5	versicolor
74	6,1	2,8	4,7	1,2	versicolor
75	6,4	2,9	4,3	1,3	versicolor
76	6,6	3,0	4,4	1,4	versicolor
77	6,8	2,8	4,8	1,4	versicolor
78	6,7	3,0	5,0	1,7	versicolor
79	6,0	2,9	4,5	1,5	versicolor
80	5,7	2,6	3,5	1,0	versicolor
81	5,5	2,4	3,8	1,1	versicolor
82	5,5	2,4	3,7	1,0	versicolor
83	5,8	2,7	3,9	1,2	versicolor
84	6,0	2,7	5,1	1,6	versicolor
85	5,4	3,0	4,5	1,5	versicolor
86	6,0	3,4	4,5	1,6	versicolor
87	6,7	3,1	4,7	1,5	versicolor
88	6,3	2,3	4,4	1,3	versicolor
89	5,6	3,0	4,1	1,3	versicolor
90	5,5	2,5	4,0	1,3	versicolor
91	5,5	2,6	4,4	1,2	versicolor

92	6,1	3,0	4,6	1,4	versicolor
93	5,8	2,6	4,0	1,2	versicolor
94	5,0	2,3	3,3	1,0	versicolor
95	5,6	2,7	4,2	1,3	versicolor
96	5,7	3,0	4,2	1,2	versicolor
97	5,7	2,9	4,2	1,3	versicolor
98	6,2	2,9	4,3	1,3	versicolor
99	5,1	2,5	3,0	1,1	versicolor
100	5,7	2,8	4,1	1,3	versicolor
101	6,3	3,3	6,0	2,5	virginica
102	5,8	2,7	5,1	1,9	virginica
103	7,1	3,0	5,9	2,1	virginica
104	6,3	2,9	5,6	1,8	virginica
105	6,5	3,0	5,8	2,2	virginica
106	7,6	3,0	6,6	2,1	virginica
107	4,9	2,5	4,5	1,7	virginica
108	7,3	2,9	6,3	1,8	virginica
109	6,7	2,5	5,8	1,8	virginica
110	7,2	3,6	6,1	2,5	virginica
111	6,5	3,2	5,1	2,0	virginica
112	6,4	2,7	5,3	1,9	virginica
113	6,8	3,0	5,5	2,1	virginica
114	5,7	2,5	5,0	2,0	virginica
115	5,8	2,8	5,1	2,4	virginica
116	6,4	3,2	5,3	2,3	virginica
117	6,5	3,0	5,5	1,8	virginica
118	7,7	3,8	6,7	2,2	virginica
119	7,7	2,6	6,9	2,3	virginica
120	6,0	2,2	5,0	1,5	virginica
121	6,9	3,2	5,7	2,3	virginica
122	5,6	2,8	4,9	2,0	virginica
123	7,7	2,8	6,7	2,0	virginica
124	6,3	2,7	4,9	1,8	virginica
125	6,7	3,3	5,7	2,1	virginica
126	7,2	3,2	6,0	1,8	virginica
127	6,2	2,8	4,8	1,8	virginica
128	6,1	3,0	4,9	1,8	virginica
129	6,4	2,8	5,6	2,1	virginica
130	7,2	3,0	5,8	1,6	virginica
131	7,4	2,8	6,1	1,9	virginica
132	7,9	3,8	6,4	2,0	virginica
133	6,4	2,8	5,6	2,2	virginica
134	6,3	2,8	5,1	1,5	virginica
135	6,1	2,6	5,6	1,4	virginica
136	7,7	3,0	6,1	2,3	virginica
137	6,3	3,4	5,6	2,4	virginica
138	6,4	3,1	5,5	1,8	virginica
139	6,0	3,0	4,8	1,8	virginica

140	6,9	3,1	5,4	2,1	virginica
141	6,7	3,1	5,6	2,4	virginica
142	6,9	3,1	5,1	2,3	virginica
143	5,8	2,7	5,1	1,9	virginica
144	6,8	3,2	5,9	2,3	virginica
145	6,7	3,3	5,7	2,5	virginica
146	6,7	3,0	5,2	2,3	virginica
147	6,3	2,5	5,0	1,9	virginica
148	6,5	3,0	5,2	2,0	virginica
149	6,2	3,4	5,4	2,3	virginica
150	5,9	3,0	5,1	1,8	virginica

ЛР 7 Категоризация и кластеризация текстов (6 часов)

Цель работы. Изучить методы кластеризации и категоризации текстов.

Задача. Выполнить кластеризацию/категоризацию текста

Порядок выполнения

1. Выбрать в Сети произвольный текст объемом не менее 10 тыс. знаков (считается, что одна страница формата А4 имеет около 2-2.5 тыс. знаков).
2. Установить ПО, предложенное преподавателем.
3. Выполнить его кластеризацию/кластеризацию.
4. Оформить отчет по ЛР, включив исходный и обработанный текста и объяснив результаты кластеризации/категоризации

Защитить отчет преподавателю, ответив на дополнительные вопросы:

– Что такое кластеризация текста?

– Что такое категоризация текста?

– Вопросы по выполненной ЛР.

ЛР 8. Инструментальные средства АБД (4 часа)

Цель работы. Познакомиться с возможностями инструментальных средств АБД

Задача. Найти в Сети описания инструментальных средств АБД и ознакомиться с их возможностями.

Порядок выполнения

1. Запустить в Сети поиск по имеющимся там описаниям ПО по АБД.
2. Ознакомиться с возможностями соответствующего ПО.
3. Выбрать одно из них, описав его назначение и функциональные возможности
4. Провести сравнительный анализ с другими программными средствами
5. Сделать вывод о преимуществах/недостатках выбранного ПО с точки зреиня решения тех или иных задач АБД
6. Оформить отчет по ЛР, включив описание выбранного ПО, перечен систем, с которыми мы проводился сравнительный анализ и результаты анализа.

Защитить отчет преподавателю, ответив на дополнительные вопросы:

- Какие программные системы по АБД Вам известны?
- Их сравнительные достоинства и недостатки?
- Вопросы по выполненной ЛР.