Кафедра Информатики

Спивакова Н.Я.

Интернет-курс по дисциплине

«Технологии обработки информации»

для специальности «Информационные системы и технологии»

Содержание

Аннотация

Программа дисциплины «Технологии обработки информации» разработана с учетом требований государственного образовательного стандарта для бакалавров по специальности «Информационные системы и технологии», утвержденной Министерством образования и науки Российской Федерации.

Дисциплина входит в состав цикла базовых дисциплин. Предметом изучения являются методы и алгоритмы сбора, хранения и анализа больших массивов данных. Объектом изучения выступают базы и хранилища данных, содержащиеся в них скрытые знания и возможности использования этих скрытых знаний лицом, принимающим решения.

Дисциплине «Технология обработки информации» предшествуют следующие предметы, необходимые при изучении данной дисциплины: «Информатика», «Дискретная математика», «Математическая логика и теория автоматов» из цикла математических и естественно-научных дисциплин, а также «Информационные технологии» и «Технологии программирования» из цикла профессиональных дисциплин.

Изучение дисциплины предусматривает лекционные и практические занятия.

Цель изучения дисциплины «Технологии обработки информации» состоит в формировании у студентов базовой системы знаний о процессах получения, хранения, очистки анализа информации, о методах преобразования информации в знания, а также в овладении навыками формализации задач и использовании программного инструментария для их решения.

Задачи изучения дисциплины:

· формирование систематизированного представления о концепциях, моделях и принципах технологий обработки информации;

· ознакомление с принципами организации информационного обмена и консолидации информации, ее поиска и извлечения;

· получение представления о трансформации данных и способах их визуализации;

· получение информации об интеллектуальном анализе данных, о превращении информации в знания.

<u>В результате изучения дисциплины обучаемый должен:</u>

знать:

· разные походы к хранению данных;
· организацию хранения данных в БД, OLAP-кубах, хранилищах данных;
· методы оценки качества данных;
· алгоритмы очистки данных;
· методы и задачи визуализации данных;
· алгоритмы анализа данных;
· методы Data Mining и Text Mining;

уметь:

· осуществлять математическую и информационную постановку задач по обработке информации;
· использовать алгоритмы обработки информации для различных приложений;
· анализировать качество данных, производить очистку данных;
· оценивать достоинства и недостатки создания хранилища данных для выбранных целей;

приобрести навыки:

· создания графиков и диаграмм разного вида для визуализации информации;
· использования интеллект-карт для мозгового штурма;
· оценки качества источников информации;
· оценки методов очистки грязных данных;
· работы с выбросами;
· владения инструментальными средствами обработки информации.

Тема 1. Предметная область дисциплины

Цели и задачи изучения данной темы – получение общетеоретических знаний об информации как основном продукте информационного общества. Изучение первой темы познакомит студентов с новыми возможностями использования информации, накопленной в обществе.

В результате успешного изучения темы Вы:

Узнаете:

· интерпретации понятия *информация* в разных предметных областях;
· откуда берутся данные и их особенности;
· о переходе данных в информацию и знания;
· о способах хранения данных;
· о современных методах анализа данных;
· о методах Data Mining и Text Mining.

Приобретете следующие профессиональные компетенции:

· умение рассчитывать количества информации, используя разные меры количества информации;
· умение описания свойств конкретной информации;
· способность дифференцировать информацию и данные;
· умение находить возможности извлекать знания из массивов данных;
· способность определять виды данных и преобразовывать данные из одного вида в другой.

В процессе освоения темы акцентируйте внимание на следующих ключевых понятиях:

Информация – любые сведения о каком-либо событии, сущности, процессе и т.п., являющиеся объектом некоторых операций, для которых существует содержательная интерпретация (для лица, принимающего решение).

Полнота информации – это свойство характеризует качество *информации* и определяет достаточность данных для принятия решений, т.е. *информация* должна содержать весь необходимый набор данных.

Достоверность информации. Информация может быть достоверной и недостоверной. В недостоверной *информации* присутствует *информационный шум*, и чем он выше, тем ниже достоверность *информации*.

Ценность информации. Ценность *информации* не может быть абстрактной. Информация должна быть полезной и ценной для определенной категории пользователей.

Адекватность информации. Это свойство характеризует степень соответствия *информации* реальному объективному состоянию. Адекватная *информация* - это полная и достоверная *информация*.

Актуальность информации. Информация должна быть актуальной, т.е. не устаревшей. Это свойство *информации* характеризует степень соответствия *информации* настоящему моменту времени.

Ясность информации. Информация должна быть понятна тому кругу лиц, для которого она предназначена.

Доступность информации. Доступность характеризует меру возможности получить определенную информацию. На это свойство *информации* влияют одновременно доступность данных и доступность адекватных методов.

Субъективность информации. *Информация* носит субъективный характер, она определяется степенью восприятия субъекта (получателя *информации*).

Данные – сведения об окружающем мире, записанные на материальном носители в любом формате (кодировке).

Знание – способность на основе информации и данных выбирать оптимальный путь достижения цели.

Формализованные знания представлены в виде научных законов, правил, моделей и записаны на каком-нибудь языке (естественном, формульном, алгоритмическом). Эти знания не привязаны к субъекту.

Скрытые (имплицированные) знания основаны на личном опыте, интуиции или хранятся внутри обученных нейронных сетей. Они не могут быть описаны и переданы другим с помощью того или иного языка, а в лучшем случае с помощью непосредственного обучения. Важной задачей является формализация скрытых знаний.

Тезаурус в теории информации обозначает совокупности всех сведений, которыми обладает субъект.

Вопросы темы:

1. Понятие информации, ее виды и свойства. Данные и знания.
2. Современные методы обработки информации.

Теоретический материал по теме

Вопрос 1. Понятие и свойства информации. Данные и знания.

***Информация*,** по своей сути, имеет многогранную природу. В разных предметных областях информацию определяют по-разному. Вот несколько определений информации:

Информация – любые сообщения о чем-либо (субъективная).

Информация – сведения, являющиеся объектом хранения, переработки и передачи (для любой системы).

Информация – количественная мера устранения неопределенности (в математике, кибернетике).

Информация – любые неизвестные ранее сведения о каком-либо событии, сущности, процессе и т.п., являющиеся объектом некоторых операций, для которых существует содержательная интерпретация (для лица, принимающего решение).

Под операциями здесь подразумевается восприятие, передача, преобразование, хранение и использование. Для восприятия *информации* необходима некоторая воспринимающая система, которая может интерпретировать ее, преобразовывать, определять соответствие определенным правилам и т.п. Таким образом, понятие *информации* следует рассматривать только при наличии источника и получателя *информации*, а также канала связи между ними.

Меры информации.

Количество информации по Шеннону. ***Получение информации рассматривается как снятие неопределенности. Пусть имеется N событий, каждое из которых может произойти с вероятностью*** $\mathbf{p_i}$ ***Тогда сообщение о том, что произошло*** **k-*е событие, снимает неопределенность для получателя информации. Количество информации, содержащееся в таком сообщении, равно.***

$$I_k = -\log_2 p_k$$

В частности, сообщение о безальтернативном событии (p=1) не содержит информации: $log_2 1 = 0$.

Обычно рассматривают **среднее** количество информации о системе. Чем менее вероятное событие, тем реже получатель будет получать информацию об этом.

$$I = -p_k \sum_1^N \log_2 p_k$$

Синтаксическая мера информации. Объем данных в сообщении измеряется количеством символов (разрядов) принятого алфавита в этом сообщении $I = V_Д$.

Семантическая мера информации. ***Для измерения смыслового содержания информации, то есть ее количества на семантическом уровне, наибольшее признание получила тезаурусная мера информации, которая связывает семантические свойства информации со способностью пользователя воспринимать поступившее сообщение.***

$$I = CV_Д,$$

где

С – коэффициент содержательности.

Прагматическая мера информации. ***Прагматическая мера информации — это полезность информации, ее ценность для пользователя (управления). Эта мера также является величиной относительной, обусловленной особенностями использования информации в той или иной системе управления. Ценность информации целесообразно измерять в тех же самых единицах (или близких к ним), в которых измеряется целевая функция.***

Информацию в зависимости от ее применения разделяют на *данные, (прагматическую) информацию и знания.*

В широком понимании *данные* ***представляют собой сведения об окружающем мире, записанные на материальном носители.***

Свойства данных:

· ***могут быть получены в на основе результате измерений, экспериментов, арифметических и логических операций;***

· ***должны быть представлены в форме, пригодной для хранения, передачи и обработки.***

Данные, по своей сути, являются *объективными*. Методы обработки являются субъективными. В основе методов обработки данных лежат алгоритмы, субъективно составленные и подготовленные. Таким образом, *информация* возникает и существует в момент диалектического взаимодействия объективных данных и субъективных методов.

Данные **– это необработанный материал, предоставляемый поставщиками данных и используемый потребителями для формирования на основе данных прагматической информации.**

Прагматическая информация – это сведения, используемые человеком для принятия решений.

В дальнейшем под **информацией** мы будем понимать именно прагматическую информацию и слово «прагматическая» будем опускать.

Свойства информации:

· **Полнота** *информации*.

Это свойство характеризует качество *информации* и определяет достаточность данных для принятия решений, т.е. *информация* должна содержать весь необходимый набор данных.

Пример. «Продажи товара А начнут сокращаться» Эта *информация* неполная, поскольку неизвестно, когда именно они начнут сокращаться.

Пример полной *информации*. «Начиная с первого квартала, продажи товара А начнут сокращаться.» Этой *информации* достаточно для принятия решений.

Достоверность *информации*.

Информация может быть достоверной и недостоверной. В недостоверной *информации* присутствует *информационный шум*, и чем он выше, тем ниже достоверность *информации*.

· **Ценность** *информации*.

Ценность *информации* не может быть абстрактной. Информация должна быть полезной и ценной для определенной категории пользователей.

· **Адекватность** *информации*.

· Это свойство характеризует степень соответствия *информации* реальному объективному состоянию. Адекватная *информация* - это полная и достоверная *информация*.

· **Актуальность** *информации*.

Информация должна быть актуальной, т.е. не устаревшей. Это свойство *информации* характеризует степень соответствия *информации* настоящему моменту времени.

· **Ясность** *информации*.

Информация должна быть понятна тому кругу лиц, для которого она предназначена.

· **Доступность** *информации*.

Доступность характеризует меру возможности получить определенную информацию. На это свойство *информации* влияют одновременно доступность данных и доступность адекватных методов.

Субъективность *информации*.

Информация носит субъективный характер, она определяется степенью восприятия субъекта (получателя *информации*).

Требования, предъявляемые к информации:

· **Динамический характер** информации.

Информация существует только в момент взаимодействия данных и методов, т.е. в момент информационного процесса. Остальное время она пребывает в состоянии данных.

· **Адекватность** используемых методов.

Информация извлекается из данных. Однако в результате использования одних и тех же данных может появляться разная *информация*. Это зависит от адекватности выбранных методов обработки исходных данных.

Для бизнеса *информация* является исходной составляющей принятия решений. Всю информацию, возникающую в процессе функционирования бизнеса и управления им, можно классифицировать определенным образом.

В зависимости от источника получения, информацию разделяют на **внутреннюю** и **внешнюю** (например, *информация*, описывающая факты, происходящие за пределами фирмы, но имеющие к ней непосредственное отношение).

Также *информация* может быть классифицирована на **фактическую** и **прогнозную**. К фактической *информации* о бизнесе относится *информация*, характеризующая свершившиеся факты; она является точной. Прогнозная *информация* является рассчитываемой или предполагаемой, поэтому ее нельзя считать точной, она может иметь определенную погрешность.

Обладая неполной информацией, нельзя точно предсказать последствия того или иного решения. Выбор решения определяется на основе научных гипотез, интуиции, накопленного опыта, то есть **знаний**.

Как и для понятия *информация*, однозначное определение понятия *знание* еще не сложилось.

Мы будем определять ***знание*** как способность на основе информации и данных выбирать оптимальный путь достижения цели в условиях неопределенности.

Неопределенность возникает из-за отсутствия достаточной информации или из-за ее противоречивости или недостоверности. Условие оптимальности подразумевает наличие альтернативных решений. Работодатель может требовать от работника специальных знаний для правильных действий или *вместо знаний* дать ему инструкции для всех случаев, лишив возможности выбора. Часто последнее оказывается дешевле.

С точки зрения машинного описания, знания можно считать **метаданными**, специально организованными и обладающими особыми свойствами.

Свойства знаний:

· **Структурированность**. *Знания* должны быть «разложены по полочкам».

· **Удобство доступа и усвоения**. Для человека – это способность быстро понять и запомнить или, наоборот, вспомнить; для компьютерных знаний – средства доступа к *знаниям*.

· **Лаконичность**. Лаконичность позволяет быстро осваивать и перерабатывать *знания* и повышает «коэффициент полезного содержания». В данный список лаконичность была добавлена из-за всем известной проблемы шума и мусорных документов, характерной именно для компьютерной *информации* – интернета и *электронного документооборота*.

· **Непротиворечивость**. *Знания* не должны противоречить друг другу.

· **Процедуры обработки**. *Знания* нужны для того, чтобы их использовать. Одно из главных свойств знаний – возможность их передачи другим и способность делать выводы на их основе. Для этого должны существовать процедуры обработки знаний.

Рассмотренные понятия являются составной частью так называемой *информационной пирамиды*, в основании которой находятся данные, следующий уровень – это информация, затем идет решение, завершает пирамиду уровень знания. По мере продвижения вверх по информационной пирамиде объемы данных переходят в ценность решений, т.е. ценность для бизнеса.

Знания могут быть:

· **Формализованные**, т.е. представленные в виде научных законов, правил, моделей и записаны на каком-нибудь языке (естественном, формульном, алгоритмическом). Эти знания не привязаны к субъекту.

· **Скрытые (имплицированные)**, т.е. основанные на личном опыте, интуиции или хранящиеся внутри обученных нейронных сетей. Они не могут быть описаны и переданы другим с помощью того или иного языка, а только с помощью непосредственного обучения.

Вопрос 2. Современные методы обработки информации.

Обработка данных может использоваться для:

· очистки и уточнения данных;
· упрощения их представления: создания репрезентативных выборок, снижение размерности;
· для более наглядного представления: визуализация;
· для извлечения знаний: Text Mining, Data Mining.

Рассмотрим подробнее процесс извлечения знаний из данных.

Технология глубинного анализа текста – **Text Mining** – позволяет анализировать большие объемы текстовой информации в поисках тенденций, шаблонов и взаимосвязей, способных помочь в принятии стратегических решений. Кроме того, Text Mining – это новый вид поиска, который в отличие традиционных подходов не только находит списки документов, формально релевантных запросам, но и помогает понять смысл, разобраться с проблематикой.

Клод Фогель, один из основателей и главный технолог компании Semio, поясняет: «Используя аналогию с библиотекой, технология Text Mining подобна открытию книги перед читателем с подчеркнутой необходимой информацией. Сравните это с выдачей читателю кипы документов и книг, в которых содержится информация, нужная читателю, однако найти ее будет непросто».

Задачи, решаемые технологией Text Mining:

1. **Классификация текста**, в которой используются статистические корреляции для построения правил размещения документов в *предопределенные категории*. Например, автоматическая рубрикация текстов – отнесение текста к той или иной известной рубрике.

Метод используют для группировки документов в intranet-сетях и на Web-сайтах, размещения документов в определенные папки, сортировки сообщений электронной почты, избирательного распространения новостей подписчикам.

2. **Кластеризация текста** – выделение компактных подгрупп объектов с близкими свойствами. Система должна *самостоятельно* найти признаки и разделить объекты по подгруппам.

Метод используют при реферировании больших документальных массивов, определение взаимосвязанных групп документов, упрощения процесса просмотра при поиске необходимой информации, нахождения уникальных документов из коллекции, выявления дубликатов или очень близких по содержанию документов.

3. **Нахождение исключений**, то есть поиск объектов, которые своими характеристиками сильно выделяются из общей массы. Для этого сначала выясняются средние параметры объектов, а потом исследуются те объекты, параметры которых наиболее сильно отличаются от средних значений.

Как известно, поиск исключений широко применяется, например, в работе спецслужб. Подобный анализ часто проводится после классификации, для того чтобы выяснить, насколько последняя была точна.

4. **Поиск связанных признаков** (полей, понятий) отдельных документов. От предсказания эта задача отличается тем, что заранее не известно, по каким именно признакам реализуется взаимосвязь; цель именно в том и состоит, чтобы найти связи признаков. Эта задача сходня с кластеризацией, но не по множеству документов, а по множеству присущих им признаков.

Например, путем анализа текста были выявления изменения стиля письма, которые могут возникать при попытке исказить или скрыть информацию. Тем самым анализировалась клиентская почта компании.

Задачи, решаемые технологией Data Mining, также основаны на анализе данных, но имеют другие применение:

1. **Классификация** позволяет, например, выделить косвенные признаки налогоплательщиков, применяющих незаконную минимизацию того или иного налога или платежа.
2. **Кластеризация,** например**,** дает представление о существовании нескольких обособленных групп населения, страдающих определенным набором болезней. Дальнейший анализ, возможно, свяжет эти группы с территорией, национальностью или образом жизни.
3. **Поиск ассоциативных правил** широко применяется маркетологами для предсказания последовательности покупок клиента.
4. *Прогнозирование* ***основано на статистической обработке данных и позволяет предсказывать численные значения фактов (демография, экономика и пр.) в будущем.***

Вопросы для самопроверки:

1. Что такое информация?
2. Какие свойства информации?
3. Как можно определить количество информации?
4. Почему используют разные меры количества информации?
5. Чем знания отличаются от информации?
6. Почему выделяют данные как особый вид информации?
7. Какие способы хранения данных используются в настоящее время?
8. Почему данные необходимо обрабатывать?
9. Какие задачи можно решать на основе имеющихся данных?
10. Какие задачи можно решить с помощью интеллектуального анализа данных?
11. Что такое Data Mining?
12. В чем особенности Text Mining?

Литература по теме:

Основная литература:

1. Чубукова И.А. Data mining. – БИНОМ. Лаборатория знаний, 2008 – Гл. 1 п.1–3.
2. Информатика. Базовый курс. Под ред. Алехиной Г.В. – М.: Маркет ДС, 2010. – Гл. 1–2.

Дополнительная литература:

1. Абдикеев Н.М., Киселев А.Д. Управление знаниями корпорации и реинжиниринг: Учебник для МБА. – М.: ИНФРА-МБ, 2011.
2. Дюк В. Data mining - интеллектуальный анализ данных. http://bizoffice.ru/search/?a=search&q=data+mining&sa.x=26&sa.y=1.
3. Амириди В. Business Intelligence и Business Performance Management: основные термины и концепции. – Управление в кредитной организации, 2011.

Напишите небольшое эссе (объемом в 1-2 страницы) по одному из перечисленных ниже вопросов:

1. Какой работник ценнее: носитель формализованных или имплицированных знаний предметной области?
2. Анализ данных: перспективы использования.
3. Персональные сведения в хранилищах данных: *за и против*.
4. Социальные сети глазами государства: источник информации или распространитель смуты?
5. Нужна ли информатику физика, *или* зачем знания, которые не могут пригодиться?

Тема 2. Визуализация данных

Цели и задачи: изучение методов и средств визуального представления информации, в частности, способов представления информации в более чем трех измерениях; разбор принципов качественной визуализации; знакомство с основными тенденциями в области визуализации.

В результате успешного изучения темы Вы:

Узнаете:

- о задачах визуализации;
- о дизайнерских подходах к инфографике;
- о разнообразии диаграмм и графиков;
- о параллельных координатах;
- о «лицах Чернова»;
- об основных тенденциях в области визуализации;
- о программных средствах визуализации.

Приобретете следующие профессиональные компетенции:

- умение строить графики и диаграммы;
- умение использовать визуализацию для анализа данных;
- умение работать с программными продуктами для визуализации;
- умение представлять пространственные объекты;
- способность оценить качество визуализации.

В процессе освоения темы акцентируйте внимание на следующих ключевых понятиях:

Инфографика – визуальное или графическое представления данных. К ифографике относят графики, диаграммы, таблицы, отчеты, списки, структурные схемы, карты и т.д.

Визуальная метафора выстраивается через соотнесение двух зрительных образов, выступающих в качестве иконических знаков. При их монтажной состыковке друг с другом возникающий смысл трактуется уже как символ такого явления, которое напрямую может быть и не связано с каждым из представленных образов.

Параллельные координаты – метод многомерной визуализации**.** Пусть имеется n-мерное пространство, для которого нужно сравнить несколько точек. Каждой координате будет соответствовать вертикальная линия, все эти линии располагаются на равных расстояниях друг от друга. Тогда каждой точке в пространстве будет соответствовать ломаная линия, состоящая из n отрезков.

Лица Чернова — это схема визуального представления мультивариативных данных в виде человеческого лица. Каждая часть лица: нос, глаза, рот — представляет собой значение определенной

переменной, назначенной для этой части (всего 18). В зависимости от значения переменной выбирается форма носа, бровей и прочее.

Интеллект-карты – это техника представления любого процесса или события, мысли или идеи в комплексной, систематизированной, визуальной форме.

OLAP (on-line analyze processing) – система, предоставляющая возможность анализировать большие объемы данных в режиме диалога.

Сравнение данных:

· ***Покомпонентное*** – прежде всего показывается *размер* каждого компонента *в процентах* от некоего целого.

· ***Позиционное*** – выявляется, как объекты соотносятся друг с другом.

· ***Временное*** – одно из наиболее распространенных. Анализируется не размер каждой доли в сравнении с целым, не соотношение долей, а то, как они изменяются во времени.

· ***Частотное*** – помогает определить, сколько объектов попадает в определенные последовательные области числовых значений.

· ***Корреляционное*** – показывает наличие (или отсутствие) зависимости между двумя переменными.

Вопросы темы:

1. Общие визуализаторы.
2. Визуализация OLAP-анализа.
3. Качество визуализации.

Теоретический материал по теме

Вопрос 1. Общие визуализаторы.

С возрастанием количества накапливаемых данных, даже при использовании сколь угодно мощных и разносторонних алгоритмов *Data Mining*, становится все сложнее воспринимать и интерпретировать полученные результаты. А, как известно, одна из задач анализа – поиск практически полезных закономерностей. Закономерность может стать практически полезной, только если ее можно осмыслить и понять.

К **инфографике** (визуальное или графическое представления данных) относят графики, диаграммы, таблицы, отчеты, списки, структурные схемы, карты и т.д. На рис. 1 даны примеры инфографики.

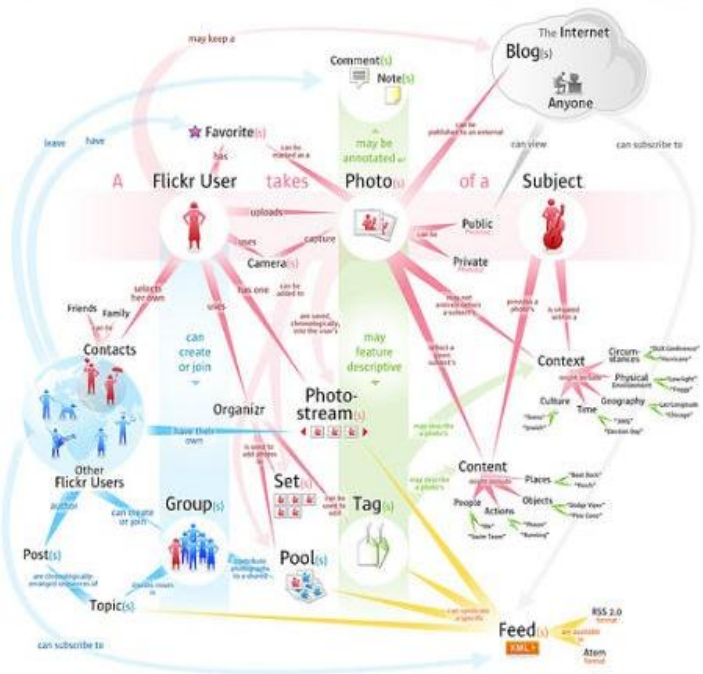

Рис. 1. Примеры инфографики

Инфографика как часть визуальных коммуникаций использует методы графического дизайна: законы композиции, образность. Одной из важнейших возможностей инфорграфики – использование визуальных метафор.

Визуальная метафора выстраивается через соотнесение двух зрительных образов, выступающих в качестве иконических знаков. При их монтажной состыковке друг с другом возникающий смысл трактуется уже как символ такого явления, которое напрямую может быть и не связано с каждым из представленных образов

Визуализация традиционно рассматривалась как вспомогательное средство при анализе данных, однако сейчас все больше исследований говорит о ее самостоятельной роли.

Традиционные методы визуализации могут находить следующее применение:

· представлять пользователю информацию в наглядном виде;

· компактно описывать закономерности, присущие исходному набору данных;

· снижать размерность или сжимать информацию;

· восстанавливать пробелы в наборе данных;
· находить шумы и выбросы в наборе данных.

Как визуализировать данные:

· **Таблицы**. Они демонстрируют подписи и значения в наиболее структурированном и организованном виде, раскрывая весь потенциал, а также позволяя сортировать и фильтровать данные. Кроме того удобно использовать условное форматирование для сравнения относительных значений (Рис. 2). **Недостаток** таблицы: наглядность теряется при увеличении размерности до 2-х.

№	Янв	Фев
123	-56,57	-133,573
124	-123,5	-81,8863
125	-77,01	46,519
126	41,579	-14,9893
127	123,52	-10,0492

Рис. 2. Таблица Excel с гистограммами условного форматирования

· Использование цвета (третье измерение) и плоскости (картографические данные) (Рис. 3).

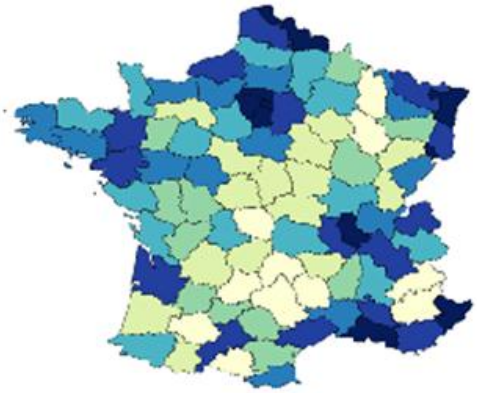

Рис. 3. Уровни преступности США

· **Лепестковые диаграммы** удобны для циклических процессов.
· **Гистограмму** применяют для сравнения значений в течение некоторого периода или же соотношения величин.
· **Круговые диаграммы** используют:
○ если необходимо отобразить соотношение частей и целого, т.е. для анализа состава или структуры явлений. Составные части целого изображаются секторами окружности. Секторы рекомендуют размещать по их величине: вверху – самый крупный, остальные – по движению часовой стрелки в порядке уменьшения их величины;
○ для отображения результатов факторного анализа, если действия всех факторов являются однонаправленными. При этом каждый фактор отображается в виде одного из секторов круга.

Все перечисленные методы относятся построению двумерных зависимостей: по оси x – факты, по оси y – измерения, соответствующие этим фактам. Но как сравнить графически ситуацию, в которой одному факту соответствует много измерений? Одним из методов является такая свертка многомерной информации к двумерной, при которой близко расположенные в многомерном пространстве поверхности будут давать близкие по форме и положению кривые.

Другой способ многомерной визуализации – использование **параллельных координат** (Рис. 4).

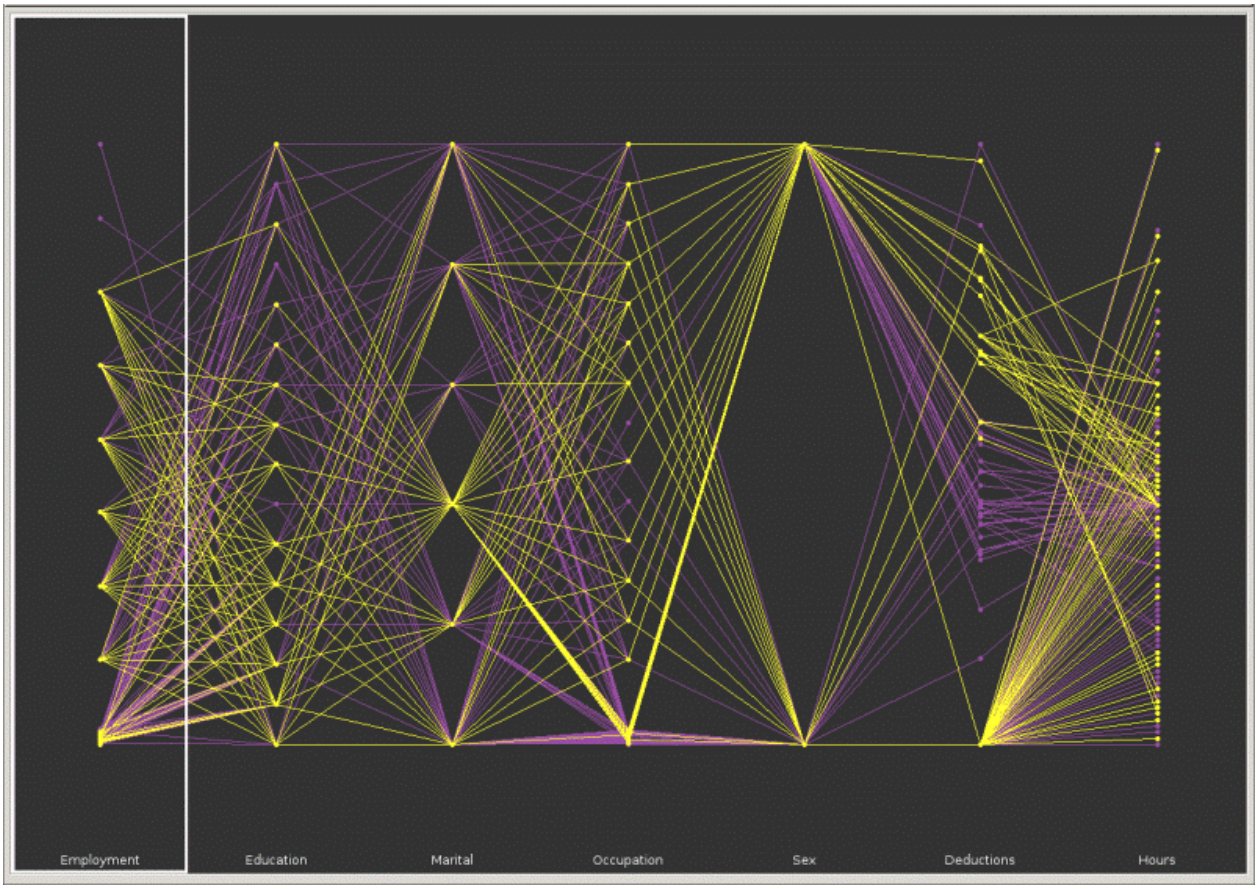

Рис. 4. Параллельные координаты

Источник: http://infographer.ru/parallels-2/

Пусть имеется n-мерное пространство, для которого нужно сравнить несколько точек. Каждой координате будет соответствовать вертикальная линия, все эти линии располагаются на равных расстояниях друг от друга. Тогда каждой точке в пространстве будет соответствовать ломаная линия, состоящая из n отрезков.

Лица Чернова — это схема визуального представления мультивариативных данных в виде человеческого лица. Каждая часть лица: нос, глаза, рот — представляет собой значение определенной переменной, назначенной для этой части (всего 18). В зависимости от значения переменной выбирается форма носа, бровей и прочее.

Человеческий глаз легко находит изменения и группирует лица по определенным признакам. Так, например, были проанализированы отличия фальшивых купюр от настоящих (Рис. 5).

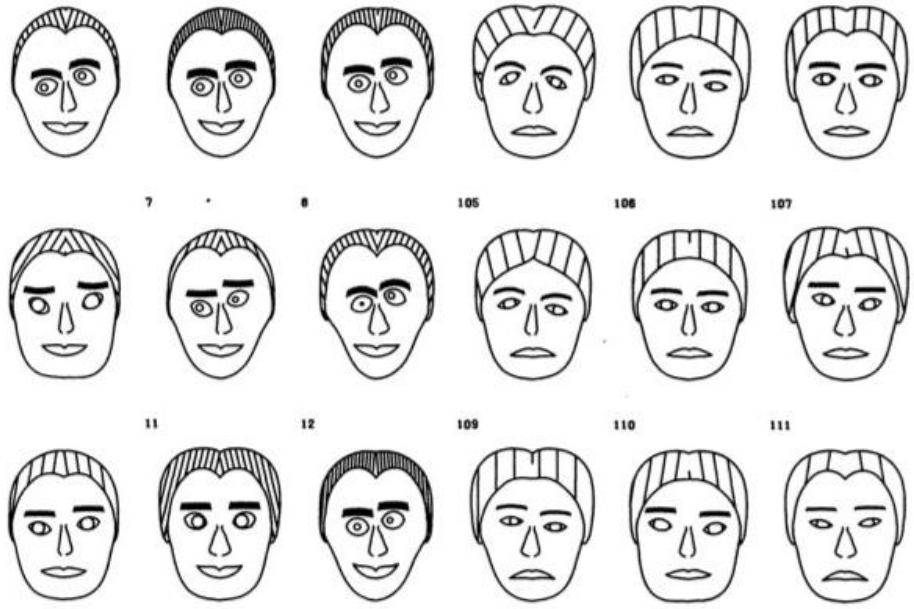

Рис. 5. Анализ фальшивых и подлинных купюр

Интеллект-карты – это техника представления любого процесса или события, мысли или идеи в комплексной, систематизированной, визуальной форме (Рис. 6).

Законы интеллект-карт разделяются на законы содержания и оформления и законы структуры.

Законы содержания и оформления:

1. Используйте эмфазу:

· Всегда используйте центральный образ.
· Как можно чаще используйте графические образы.
· Для центрального образа используйте три и более цветов.
· Чаще придавайте изображению объем, а также используйте выпуклые буквы.
· Пользуйтесь синестезией (комбинированием всех видов эмоционально-чувственного восприятия).
· Варьируйте размеры букв, толщину линий и масштаб графики.
· Стремитесь к оптимальному размещению элементов на интеллект-карте.
· Стремитесь к тому, чтобы расстояние между элементами интеллект-карты было соответствующим.

2. Ассоциируйте:

· Используйте стрелки, когда необходимо показать связи между элементами интеллект-карты.
· Используйте цвета.
· Используйте кодирование информации.

3. Стремитесь к ясности в выражении мыслей:

· Придерживайтесь принципа: по одному ключевому слову на каждую линию.
· Используйте печатные буквы.
· Размещайте ключевые слова над соответствующими линиями.
· Следите за тем, чтобы длина линии примерно равнялась длине соответствующего ключевого слова.
· Соединяйте линии с другими линиями и следите за тем, чтобы главные ветви карты соединялись с центральным образом.
· Делайте главные линии плавными и более жирными.
· Отграничивайте блоки важной информации с помощью линий.
· Следите за тем, чтобы ваши рисунки (образы) были предельно ясными.
· Держите бумагу горизонтально перед собой, предпочтительно в положении «ландшафт».
· Старайтесь располагать слова горизонтально.
· Вырабатывайте собственный стиль.

Законы структуры:

· Соблюдайте иерархию мыслей.
· Используйте номерную последовательность в изложении мыслей.

Рис. 6. Пример интеллект-карты

Вопрос 2. Визуализация OLAP-анализа.

OLAP (on-line analyze processing) – система, предоставляющая возможность анализировать большие объемы данных в режиме диалога. В следующей теме будет рассмотрена техническая сторона вопроса создания OLAP-машины. А здесь мы рассмотрим возможности визуализации этих данных.

В основе OLAP лежит идея многомерной модели данных. По измерениям в многомерной модели откладывают факторы, влияющие на деятельность предприятия (например: время, продукты, отделения компании, географию и т.п.). Таким образом получают гиперкуб (который затем наполняется показателями деятельности предприятия (цены, продажи, план, прибыли, убытки и т.п.).

Все OLAP-системы характеризуются общими принципами построения (см. рис 6):

1. В качестве внешнего интерфейса они предоставляют управляемую динамическую таблицу. На вход динамической таблицы подается многомерный массив. Массив состоит из данных двух типов: измерений и фактов. Измерения становятся колонками и строками динамической таблицы. В них отображаются члены измерений. На пересечении колонок и строк размещены факты.

2. Колонки и строки являются основными инструментами управления таблицей. С их помощью пользователь может манипулировать исходными данными: менять местами строки и колонки, устанавливать фильтры по измерениям, детализировать информацию или наоборот обобщать ее. При этом промежуточные и окончательные итоги по фактам автоматически пересчитываются. Выполнение этих операций обеспечивается OLAP-машиной (или машиной OLAP-вычислений). Сами манипуляции с данными носят название OLAP-операций.

3. Еще одной важной стороной OLAP-анализа является графическое отображение данных. График синхронизирован с динамической таблицей. После выполнения любой OLAP-операции данные пересчитываются, а график перерисовывается.

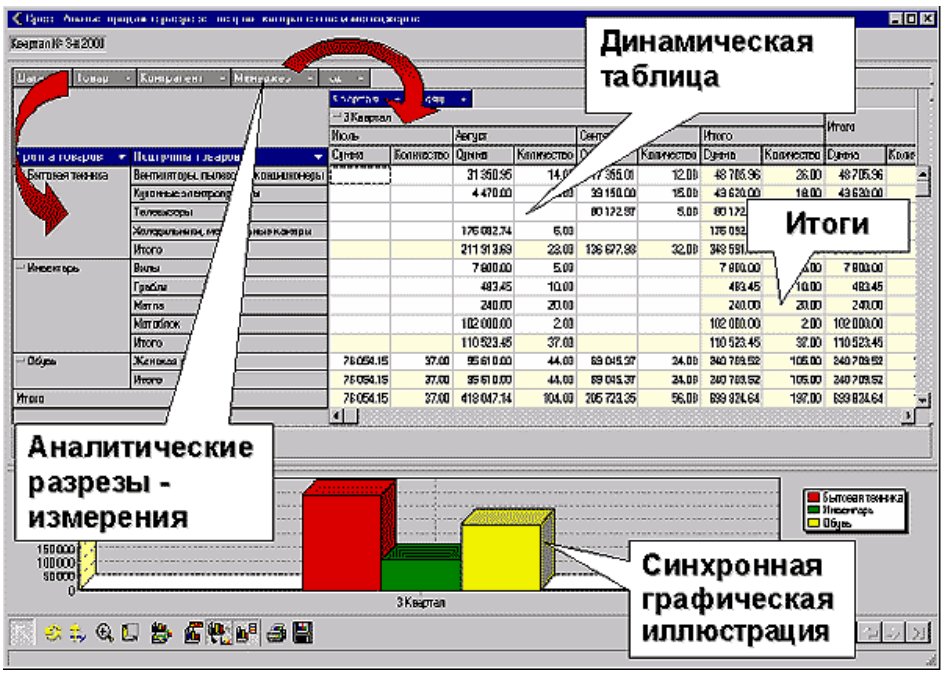

Рис. 7. Отображение результатов анализа в OLAP-системе

После настройки OLAP-системы на данные, пользователь получит возможность быстро получать ответы на ключевые вопросы путем простых манипуляций мышью над OLAP- таблицей. При этом будут доступны некоторые стандартные методы анализа, следующие из природы OLAP-технологии.

Как видно из рис. 7, диаграмма наглядно показывает долю каждой группы товаров в общей выручке 3-го квартала. Рассмотрев подобную информацию по всем кварталам, можно также оценить наличие (или отсутствие) сезонной цикличности.

Факторный (структурный) анализ. Анализ структуры продаж для выявления важнейших составляющих в интересующем разрезе. Для этого удобно использовать диаграмму типа «Пирог», в более сложных случаях, когда исследуется сразу 3 измерения – «Столбцы». Например, в магазине «Дары моря» за квартал продажи рыбы = $100000, пива = $1000, хлеба = $500. Вывод: оборот магазина зависит только от рыбы (на самом деле быть может пиво необходимо для продажи рыбы, но это уже анализ зависимостей).

Анализ динамики. Выявление тенденций, сезонных колебаний. Наглядно динамику отображает график типа «Линия». Например, объемы продаж мойвы в течение года падали, а объемы продаж форели росли. Возможно, улучшилось благосостояние среднего покупателя, или изменился имидж магазина, а с ним и состав покупателей. Требуется провести корректировку ассортимента. Другой пример, в течение 3 лет летом снижается объем продаж пива темных сортов.

Анализ зависимостей. Сравнение объемов продаж разных товаров во времени для выявления необходимого ассортимента - «корзины». Для этого также удобно использовать график типа «Линия». Например, при удалении из ассортимента пива в течение первых двух месяцев обнаружилось падение продаж воблы.

Сопоставление (сравнительный анализ). Сравнение результатов продаж во времени, или за заданный период, или для заданной группы товаров. В зависимости от количества анализируемых факторов (от 1 до 3-х) используется диаграмма типа «Пирог» или «Столбцы». Пример, сравнение результатов продаж однотипных магазинов для оценки качества работы менеджеров.

Этими видами анализа возможности OLAP не исчерпываются. Например, применяя в качестве алгоритма вычисления промежуточных и окончательных итогов среднее арифметическое, или функции статистического анализа – дисперсия, среднее отклонение и т.д. можно получить самые изощренные виды аналитических отчетов.

Вопрос 3. Качество визуализации.

Визуализация не всегда может помочь аналитику. Неправильно выбранная диаграмма может ввести в заблуждение или утомить его.

На рис. 8показаны распределение рынков компаний А и Б. Для сравнения их между собой потребуется взять карандаш и переписать цифры. Более наглядная гистограмма на рис. 9: по ней сразу видно, что на юге рынок захвачен А, а на севере Б; на востоке и западе компании активно конкурируют друг с другом.

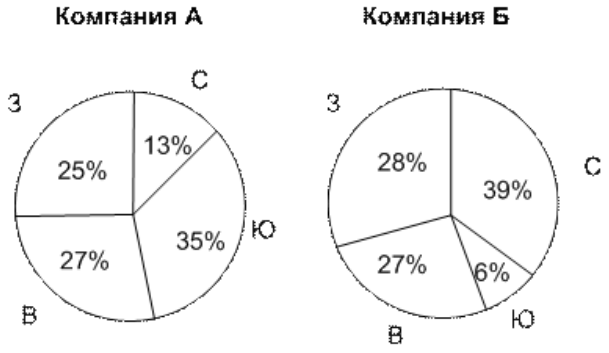

Рис. 9. Визуализация распределения рынков компаний А и Б

Рис. 10. Визуализация распределения рынков компаний А и Б показывает вытеснение одной компании другой

Правильный выбор метода визуализации состоит из трех шагов:

Шаг 1: Формулирование идеи

(от данных к идее)

Для того чтобы правильно выбрать тип диаграммы, вы в первую очередь должны четко сформулировать конкретную идею, которую вы хотите донести до аудитории при помощи диаграммы.

Шаг 2: Определение типа сравнения данных

(от идеи к сравнению)

Сформулированная идея будет обязательно заключать в себе один из пяти основных типов сравнения данных:

· **покомпонентное** – прежде всего показывается *размер* каждого компонента *в процентах* от некоего целого;

· **позиционное** – выявляется, как объекты соотносятся друг с другом;

· **временное** – одно из наиболее распространенных. Анализируется не размер каждой доли в сравнении с целым, не соотношение долей, а то, как они изменяются во времени;

· **частотное** – помогает определить, сколько объектов попадает в определенные последовательные области числовых значений;

· **корреляционное** – показывает наличие (или отсутствие) зависимости между двумя переменными.

Шаг 3. Выбор типа диаграммы

(от сравнения к диаграмме)

Каждому типу сравнения соответствует один из пяти видов диаграмм или их модификации (рис. 11).

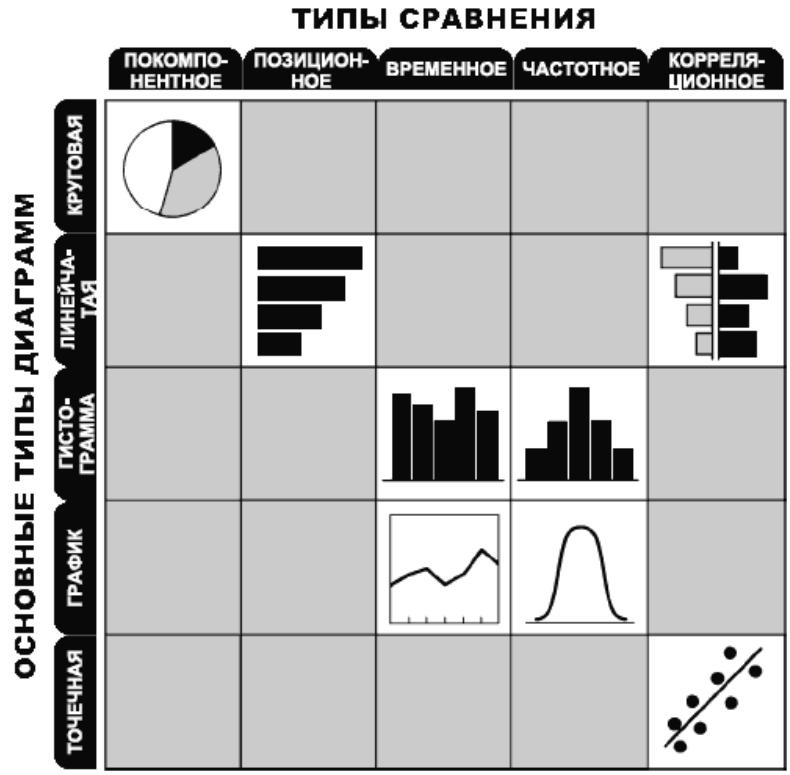

Рис. 11. Соответствие типов диаграмм задачам сравнения [2].

Вопросы для самопроверки:

1. Почему необходима визуализация?
2. Что такое инфографика и где она применяется?
3. Какие задачи решает визуализация?
4. Какие виды диаграмм вы используете?
5. Для чего используют параллельные координаты?
6. Как с помощью «лиц Чернова» визуализируют данные?
7. Что такое интеллект-карты?
8. Какие правила полезно знать при использовании интеллект-карт?
9. Какие задачи решает аналитик с помощью OLAP-машины?
10. К чему может привести некачественная визуализация?
11. Приведите правила выбора диаграмм.

Литература по теме:

Основная литература:

1. Чубукова И.А. Data mining. – БИНОМ. Лаборатория знаний, 2008 г. §16. Визуализация данных, §17 п.3: OLAP-системы.
2. Желязны Д. Говори на языке диаграмм: пособие по визуальным коммуникациям.

Дополнительная литература:

1. Бьюзен Т., Бьюзен Б. Интеллект-карты. Практическое руководство. – Попурри, 2010. – 368 стр.
2. Мамаев А. Н., Кудлай. А. Д. Визуализация данных в презентациях, отчетах и исследованиях. – Практическая Медицина, 2011. – 40 стр.
3. Бехтерев С., Архангельский Г. Майнд-менеджмент. Решение бизнес-задач с помощью интеллект-карт. – Альпина Паблишер, 2012. – 312 с.
4. Методы OLAP-анализа. http://www.rarus.nn.ru/products/olap/olap.htm

Тема 3. Информационный обмен и консолидация информации

Цели и задачи изучения данной темы – получение общетеоретических знаний о хранилищах данных как основном способе хранения данных в аналитических системах. Изучив эту тему, студенты познакомятся в основными методами обработки данных при создании хранилищ.

В результате успешного изучения темы Вы:

Узнаете:

- что такое хранилище данных;
- архитектуру хранилищ данных;
- особенности хранения данных для оперативной и аналитической обработки больших объемов;
- источники данных для аналитических систем;
- особенности трансформации данных для ускорения их последующего анализа.

Приобретете следующие профессиональные компетенции:

- умение подготовки данных для многомерными хранилищ;
- умение проводить элементы OLAP-анализа;
- способность к обработке временных рядов;
- умение группировать и разгруппировывать данные.

В процессе освоения темы акцентируйте внимание на следующих ключевых понятиях:

Консолидация – комплекс методов и процедур, направленных на извлечение данных из различных источников, обеспечение необходимого уровня их информативности и качества, преобразование в единый формат, в котором они могут быть загружены в хранилище данных или аналитическую систему.

Транзакция – некоторый набор операций над базой данных, который рассматривается как единое завершенное, с точки зрения пользователя, действие над некоторой информацией, обычно связанное с обращением к базе данных.

OLTP (On-Line Transaction Processing) – оперативная, то есть в режиме реального времени, обработка транзакций.

СППР – системы поддержки принятия решений.

Хранилище данных – разновидность систем хранения, ориентированная на поддержку процесса анализа данных, обеспечивающая целостность, непротиворечивость и хронологию данных, а также высокую скорость выполнения аналитических запросов.

Метаданные — высокоуровневые средства отражения информационной модели и описания структуры данных, используемой в ХД. Метаданные должны содержать описание структуры данных хранилища и структуры данных импортируемых источников. Метаданные хранятся отдельно от данных в так называемом *репозитарии* метаданных.

Измерения — это категориальные атрибуты, наименования и свойства объектов, участвующих в некотором бизнес-процессе. Значениями измерений являются наименования товаров, названия фирм-поставщиков и покупателей, ФИО людей, названия городов и т.д. Измерения могут быть и числовыми, если какой-либо категории (например, наименованию товара) соответствует числовой код, но в любом случае это данные дискретные, то есть принимающие значения из ограниченного набора. Измерения качественно описывают исследуемый бизнес-процесс.

Факты — это данные, количественно описывающие бизнес-процесс, непрерывные по своему характеру, то есть они могут принимать бесконечное множество значений. Примеры фактов — цена товара или изделия, их количество, сумма продаж или закупок, зарплата сотрудников, сумма кредита, страховое вознаграждение и т.д.

Сечение заключается в выделении подмножества ячеек гиперкуба при фиксировании значения одного или нескольких измерений. В результате сечения получается срез или несколько срезов, каждый из которых содержит информацию, связанную со значением измерения, по которому он был построен.

Транспонирование (вращение) обычно применяется к плоским таблицам, полученным, например, в результате среза, и позволяет изменить порядок представления измерений таким образом, что измерения, отображавшиеся в столбцах, будут отображаться в строках, и наоборот. В ряде случаев транспонирование позволяет сделать таблицу более наглядной.

Операции ***свертки*** (группировки) и ***детализации*** (декомпозиции) возможны только тогда, когда имеет место иерархическая подчиненность значений измерений. При свертке одно или несколько подчиненных значений измерений заменяются теми значениями, которым они подчинены. При этом уровень обобщения данных уменьшается. Так, если отдельные товары образуют группы, например Стройматериалы, то в результате свертки вместо отдельных наименований товаров будет указано наименование группы, а соответствующие им факты будут агрегированы. Проиллюстрируем результаты свертки: в табл. 2 представлена исходная таблица, а в табл. 3 — результат ее свертки по измерению Товар.

Детализация — это процедура, обратная свертке; уровень обобщения данных уменьшается. При этом значения измерений более высокого иерархического уровня заменяются одним или несколькими значениями более низкого уровня, то есть вместо наименований групп товаров отображаются наименования отдельных товаров.

Реляционная база данных (relational database) — совокупность отношений, содержащих всю информацию, которая должна храниться в базе. Физически это выражается в том, что информация хранится в виде двумерных таблиц, связанных с помощью ключевых полей.

Витрина данных – специализированное локальное тематическое хранилище, подключенное к централизованному ХД и обслуживающее отдельное подразделение организации или определенное направление ее деятельности.

ETL – комплекс методов, реализующих процесс переноса исходных данных из различных источников в аналитическое приложение или поддерживающее его хранилище данных.

Временным рядом называют серию величин, полученную через регулярные промежутки времени. Например, месячные показатели инфляции или значения температуры воздуха, измеряемые каждые 10 дней.

Квантование значений. При выполнении этой операции осуществляется разбиение диапазона числовых значений на указанное количество интервалов определенным методом и замена каждого обрабатываемого значения на число, связанное с интервалом, к которому оно относится, либо на метку интервала.

Табличная замена значений. В результате выполнения этой операции производится замена значений по таблице подстановки, которая содержит пары, состоящие из исходного и выходного значения. Например, 0 - «красный», 1 - «зеленый», 2 - «синий».

Скользящее окно. При решении некоторых задач, например, при прогнозировании временных рядов с помощью нейросети, требуется подавать на вход анализатора значения несколько смежных отсчетов из исходного набора данных. При этом эффективность реализации заметно повышается, если не выбирать данные каждый раз из нескольких последовательных записей, а последовательно расположить данные, относящиеся к конкретной позиции окна, в одной записи.

Преобразование даты. Разбиение даты необходимо для анализа всевозможных показателей за определенный период (день, неделя, месяц, квартал, год). Суть разбиения заключается в том, что на основе столбца с информацией о дате формируется другой столбец, в котором указывается, к какому заданному интервалу времени принадлежит строка данных. Тип интервала задается аналитиком, исходя из того, что он хочет получить, - данные за год, квартал, месяц, неделю, день или сразу по всем интервалам.

Группировка. Трудно делать какие-либо выводы по данным каждой записи в отдельности. Аналитику для принятия решения часто необходима сводная информация. Совокупные данные намного более информативны, тем более если их можно получить в разных разрезах. *Группировка* позволяет объединять записи по полям-измерениям, агрегируя данные в полях-фактах для дальнейшего анализа.

Разгруппировка. *Группировка* используется для объединения фактов по каким-либо измерениям. При этом под объединением понимается применение некоторой функции агрегации. Если в исходном наборе данных присутствовали какие-либо другие измерения, то теряется информация о значениях фактов в разрезе этих измерений. Алгоритм *разгруппировки* позволяет восстановить эти факты, но их значения восстанавливаются не точно, а пропорционально вкладу в сгруппированные значения.

Вопросы темы:

1. Создание хранилищ данных.
2. Многомерное представление данных.
3. Трансформация данных.

Теоретический материал

Вопрос 1. Создание хранилищ данных.

Ценность и достоверность знаний, полученных в результате интеллектуального анализа бизнес-данных, зависит не только от эффективности используемых аналитических методов и алгоритмов, но и от того, насколько правильно подобраны и подготовлены исходные данные для анализа.

Поэтому, прежде чем приступать к анализу данных, необходимо выполнить ряд процедур, цель которых – доведение данных до приемлемого уровня качества и информативности, а также организовать их интегрированное хранение в структурах, обеспечивающих их целостность, непротиворечивость, высокую скорость и гибкость выполнения аналитических запросов.

Консолидация — комплекс методов и процедур, направленных на извлечение данных является начальным этапом реализации любой аналитической задачи или проекта.

В процессе консолидации данных решаются следующие задачи:

- **выбор источников данных;**
- **разработка стратегии консолидации;**
- оценка качества данных (вспомогательные);
- обогащение (вспомогательные);
- очистка (вспомогательные);
- **перенос в хранилище данных.**

Основные критерии оптимальности с точки зрения консолидации данных:

- обеспечение высокой скорости доступа к данным;
- компактность хранения;
- автоматическая поддержка целостности структуры данных;
- контроль непротиворечивости данных.

Источники данных могут быть разные, располагаться на удаленных компьютерах, принадлежащим разным сетям. Например, процесс сбора, хранения и оперативной обработки данных на типичном предприятии обычно содержит несколько уровней. На верхнем уровне располагаются реляционные SQL-ориентированные СУБД типа SQL Server, Oracle и т.д. На втором – файловые серверы с некоторой системой оперативной обработки или сетевые версии персональных СУБД типа R-Base, FoxPro, Access и т.д. И наконец, на самом нижнем уровне расположены локальные ПК отдельных пользователей с персональными источниками данных. Чаще всего информация на них собирается в виде файлов офисных приложений – Word, Excel, текстовых файлов и т.д.

Системы оперативной обработки информации OLTP существовали в организациях еще с 70-х годов прошлого века. Обобщенная структура системы OLTP представлена на рис. 12.

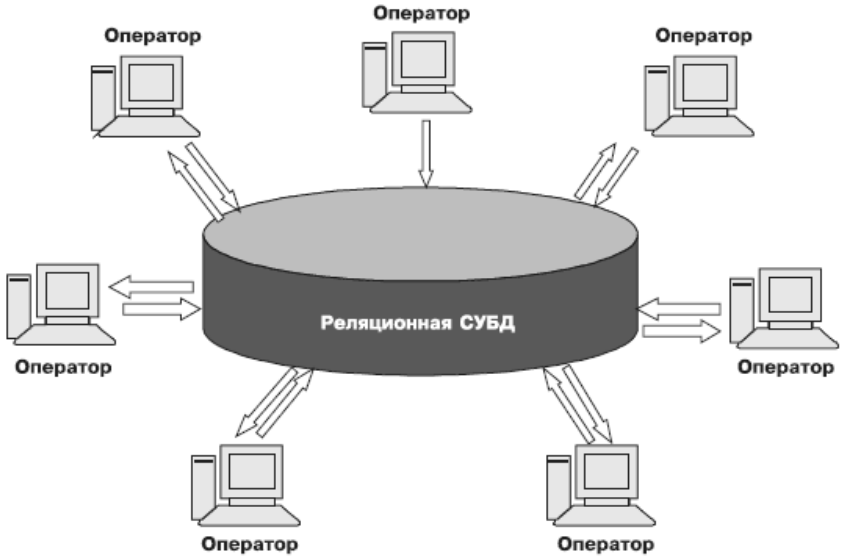

Рис. 12. OLTP-система

Рис. 13. структура информационной СППР

Позднее информационные системы стали применяться не только для оперативной обработки информации. Были разработаны интерактивные автоматизированные системы, которые помогают ЛПР использовать данные и модели, чтобы решать неструктурированные проблемы. Такие системы получили название СППР (системы поддержки принятия решений), или DSS (Decision Support Systems). В процессе разработки систем анализа информации и методологии их применения обнаружилось, что для эффективного функционирования такие системы должны быть организованы несколько иным способом, чем тот, который применяется в OLTP-системах. Это обусловлено следующими причинами:

· Для выполнения сложных аналитических запросов необходима обработка больших объемов данных из разнообразных источников.

· Для выполнения запросов, связанных с анализом тенденций, прогнозированием протяженных во времени процессов, необходимы исторические данные, накопленные за достаточно длительный период.

· Данные, используемые для целей анализа и обслуживания аналитических запросов, отличаются от используемых в обычных OLTP-системах. При аналитической обработке предпочтение отдается не детальным данным, а обобщенным (агрегированным).

В связи с этим можно выделить ряд принципиальных отличий СППР и OLTP-систем. Эти отличия представлены в табл. 1.

Таблица 1.

Отличия СППР и OLTP-систем

Свойство	OLTP-система	СППР
Цели использования данных	Быстрый поиск, простейшие алгоритмы обработки	Аналитическая обработка с целью поиска скрытых закономерностей, построения прогнозов и моделей и т.д.

Уровень обобщения (детализации) данных	Детализированные	Как детализированные, так и обобщенные (агрегированные)
Требования к качеству данных	Возможны некорректные данные (ошибки регистрации, ввода и т.д.)	Ошибки в данных не допускаются, поскольку могут привести к некорректной работе аналитических алгоритмов
Формат хранения данных	Данные могут храниться в различных форматах в зависимости от приложения, в котором они были созданы	Данные хранятся и обрабатываются в едином формате
Время хранения данных	Как правило, не более года (в пределах отчетного периода)	Годы, десятилетия
Изменение данных	Данные могут добавляться, изменяться и удаляться	Допускается только пополнение; ранее добавленные данные изменяться не должны, что позволяет обеспечить их хронологию
Периодичность обновления	Часто, но в небольших объемах	Редко, но в больших объемах
Доступ к данным	Должен быть обеспечен доступ ко всем текущим (оперативным) данным	Должен быть обеспечен доступ к историческим (то есть накопленным за достаточно длительный период времени) данным с соблюдением их хронологии
Характер выполняемых запросов	Стандартные, настроенные заранее	Нерегламентированные, формируемые аналитиком «на лету» в зависимости от требуемого анализа
Время выполнения запроса	Несколько секунд	До нескольких минут

Как видно из табл. 1, требования к СППР и OLTP-системам существенно отличаются, поэтому в СППР используются хранилища данных (ХД), ориентированные на аналитическую обработку данных.

Основные особенности концепции ХД.

В настоящее время однозначного определения ХД не существует, из-за того что разработано большое количество различных архитектур и технологий ХД, а сами хранилища используются для решения самых разнообразных задач. Каждый автор вкладывает в это понятие свое видение вопроса. Обобщая требования, предъявляемые к СППР, можно дать следующее определение ХД, которое не претендует на полноту и однозначность, но позволяет понять основную идею.

Хранилище данных – разновидность систем хранения, ориентированная на поддержку процесса анализа данных, обеспечивающая целостность, непротиворечивость и хронологию данных, а также высокую скорость выполнения аналитических запросов.

Важнейшим элементом ХД является семантический слой — механизм, позволяющий аналитику оперировать данными посредством бизнес-терминов предметной области. Семантический слой дает пользователю возможность сосредоточиться на анализе и не задумываться о механизмах получения данных.

Основные требования к ХД.

Чтобы ХД выполняло функции, соответствующие его основной задаче – поддержке процесса анализа данных – оно должно удовлетворять требованиям, сформулированным Р. Кимбаллом, одним из авторов концепции ХД:

- высокая скорость получения данных из хранилища;
- автоматическая поддержка внутренней непротиворечивости данных;
- возможность получения и сравнения срезов данных;
- наличие удобных средств для просмотра данных в хранилище;
- обеспечение целостности и достоверности хранящихся данных.

Чтобы соблюсти все перечисленные требования, для построения и работы ХД, как правило, используется не одно приложение, а система, в которую входит несколько программных продуктов. Одни из них представляют собой собственно систему хранения данных, другие — средства их просмотра, извлечения, загрузки и т.д.

В основе концепции ХД лежат следующие положения:

· интеграция и согласование данных из различных источников, таких как обычные системы оперативной обработки, базы данных, учетные системы, офисные документы, электронные архивы, расположенные как внутри предприятия, так и во внешнем окружении;

· разделение наборов данных, используемых системами выполнения транзакций и СППР.

Круг задач интеллектуального анализа данных весьма широк, а сами задачи существенно различаются по уровню сложности. Поэтому в зависимости от специфики решаемых задач и уровня их сложности архитектура ХД и модели данных, используемых для их построения, могут различаться. Однако в любом случае данные извлекаются из различных источников и загружаются в ХД, которое содержит как данные разного вида, представленные в соответствии с некоторой моделью, так и метаданные.

Виды данных в ХД:

· **Детализированные** данные поступают непосредственно из источников данных и соответствуют элементарным событиям, регистрируемым OLTP-системами.

· **Агрегированные** данные. Многие задачи анализа (например, прогнозирование) требуют использования данных определенной степени обобщения. Например, сумма, среднее, максимальное и минимальное значения за соответствующий период. Поскольку один и тот же набор детализированных данных может породить несколько наборов агрегированных данных с различной степенью обобщения, объем ХД возрастает, иногда существенно. Иногда это приводит к «взрывному», неконтролируемому росту ХД и вызывает серьезные технические проблемы: хранилище «распухает», из-за того что непрерывный поток входных данных автоматически агрегируется в соответствии с настройками ХД. Однако с этим приходится мириться: если бы агрегированные данные не содержались в ХД, а вычислялись в процессе выполнения запросов, время выполнения запроса увеличилось бы в несколько раз.

· **Метаданные** – высокоуровневые средства отражения информационной модели и описания структуры данных, используемой в ХД. Метаданные должны содержать описание структуры данных хранилища и структуры данных импортируемых источников. Метаданные хранятся отдельно от данных в так называемом репозитарии метаданных.

Можно выделить два уровня метаданных — технический (административный) и бизнес-уровень. Технический уровень содержит метаданные, необходимые для обеспечения функционирования хранилища (статистика загрузки данных и их использования, описание модели данных и т.д.). Бизнес-метаданные обеспечивают пользователю возможность концентрироваться на процессе анализа, а не на технических аспектах работы с хранилищем; они включают бизнес-термины и определения, которыми привык оперировать пользователь.

Вопрос 2. Многомерное представление данных.

Одним из способов представления данных в ХД являются многомерные модели, в которых. В основе многомерного представления данных лежит их разделение на две группы — измерения и факты.

Измерения — это категориальные атрибуты, наименования и свойства объектов, участвующих в некотором бизнес-процессе. Значениями измерений являются наименования товаров, названия фирм-поставщиков и покупателей, ФИО людей, названия городов и т.д. Измерения могут быть и числовыми, если какой-либо категории (например, наименованию товара) соответствует числовой код, но в любом случае это данные дискретные, то есть принимающие значения из ограниченного набора. Измерения качественно описывают исследуемый бизнес-процесс.

Факты — это данные, количественно описывающие бизнес-процесс, непрерывные по своему характеру, то есть они могут принимать бесконечное множество значений. Примеры фактов — цена товара или изделия, их количество, сумма продаж или закупок, зарплата сотрудников, сумма кредита, страховое вознаграждение и т.д.

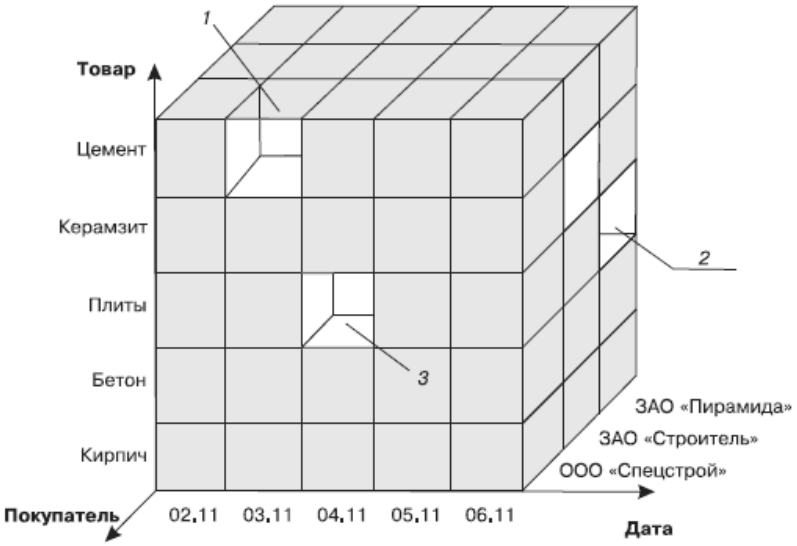

Рис. 14. Принцип организации многомерного куба

Многомерный взгляд на измерения Дата, Товар и Покупатель (см. рис. 14) Фактами в данном случае могут быть Цена, Количество, Сумма. Тогда выделенный сегмент будет содержать информацию о том, сколько плит, на какую сумму и по какой цене приобрела фирма ЗАО «Строитель» 3 ноября.

Таким образом, информация в многомерном хранилище данных является логически целостной. Это уже не просто наборы строковых и числовых значений, которые в случае реляционной модели нужно получать из различных таблиц, а целостные структуры типа «кому, что и в каком количестве было продано на данный момент времени». Преимущества многомерного подхода очевидны:

· Представление данных в виде многомерных кубов более наглядно, чем совокупность нормализованных таблиц реляционной модели, структуру которой представляет только администратор БД.

· Возможности построения аналитических запросов к системе, использующей МХД, более широки.

· В некоторых случаях использование многомерной модели позволяет значительно уменьшить продолжительность поиска в МХД, обеспечивая выполнение аналитических запросов практически в режиме реального времени. Это связано с тем, что агрегированные данные вычисляются предварительно и хранятся в многомерных кубах вместе с детализированными, поэтому тратить время на вычисление агрегатов при выполнении запроса уже не нужно.

В принципе, OLAP-куб может быть реализован и с помощью обычной реляционной модели. В этом случае имеет место эмуляция многомерного представления совокупностью плоских таблиц. Такие системы получили название ROLAP – Relational OLAP.

Использование многомерной модели данных сопряжено с определенными трудностями. Так, для ее реализации требуется больший объем памяти. Это связано с тем, что при реализации физической многомерности используется большое количество технической информации, поэтому объем данных, который может поддерживаться МХД, обычно не превышает нескольких десятков гигабайт. Кроме того, многомерная структура труднее поддается модификации; при необходимости встроить еще одно измерение требуется выполнить физическую перестройку всего многомерного куба. На основании этого можно сделать вывод, что применение систем хранения, в основе которых лежит многомерное представление данных, целесообразно только в тех случаях, когда объем используемых данных сравнительно невелик, а сама многомерная модель имеет стабильный набор измерений.

Работа с измерениями.

В процессе поиска и извлечения из гиперкуба нужной информации над его измерениями производится ряд действий, наиболее типичными из которых являются:

· сечение (срез);
· транспонирование;
· свертка;
· детализация.

Сечение заключается в выделении подмножества ячеек гиперкуба при фиксировании значения одного или нескольких измерений. В результате сечения получается срез или несколько срезов, каждый из которых содержит информацию, связанную со значением измерения, по которому он был построен.

Транспонирование (вращение) обычно применяется к плоским таблицам, полученным, например, в результате среза, и позволяет изменить порядок представления измерений таким образом, что измерения, отображавшиеся в столбцах, будут отображаться в строках, и наоборот. В ряде случаев транспонирование позволяет сделать таблицу более наглядной.

Операции **свертки** (группировки) и **детализации** (декомпозиции) возможны только тогда, когда имеет место иерархическая подчиненность значений измерений. При свертке одно или несколько подчиненных значений измерений заменяются теми значениями, которым они подчинены. При этом уровень обобщения данных уменьшается.

Детализация — это процедура, обратная свертке; уровень обобщения данных уменьшается. При этом значения измерений более высокого иерархического уровня заменяются одним или несколькими значениями более низкого уровня, то есть вместо наименований групп товаров отображаются наименования отдельных товаров.

В начале 1970-х гг. англо-американский ученый Э. Кодд разработал реляционную модель организации хранимых данных, которая положила начало новому этапу эволюции СУБД. Благодаря простоте и гибкости реляционная модель стала доминирующей, а реляционные СУБД стали промышленным стандартом де-факто.

Витрина данных – специализированное локальное тематическое хранилище, подключенное к централизованному ХД и обслуживающее отдельное подразделение организации или определенное направление ее деятельности.

Использование витрин данных имеет следующие преимущества, делающие их ближе и доступнее конечному пользователю:

· содержание данных, тематически ориентированных на конкретного пользователя;

· относительно небольшой объем хранимых данных, на организацию и поддержку которых не требуется значительных затрат;

· улучшенные возможности в разграничении прав доступа пользователей, так как каждый из них работает только со своей витриной и имеет доступ только к информации, относящейся к определенному направлению деятельности.

Вопрос 3. Трансформация данных.

Для переноса данных из разных источников требуются специальные средства.

ETL – комплекс методов, реализующих процесс переноса исходных данных из различных источников в аналитическое приложение или поддерживающее его хранилище данных.

Основные цели и задачи процесса ETL:

· Исходные данные расположены в источниках самых разнообразных типов и форматов, созданных в различных приложениях, и, кроме того, могут использовать различную кодировку, в то время как для решения задач анализа данные должны быть преобразованы в единый универсальный формат, который поддерживается ХД и аналитическим приложением.

· Данные в источниках обычно излишне детализированы, тогда как для решения задач анализа в большинстве случаев требуются обобщенные данные.

· Исходные данные, как правило, являются «грязными», то есть содержат различные факторы, которые мешают их корректному анализу.

Трансформация данных – комплекс методов и алгоритмов, направленных на оптимизацию представления и форматов данных с точки зрения решаемых задач и целей анализа. Трансформация не ставит целью изменить информационное содержание данных. Её задача – представить эту информацию в таком виде, чтобы она могла быть использована наиболее эффективно.

Вообще, трансформация данных – очень широкое понятие, не имеющее четких границ. Однако в контексте аналитических технологий трансформация данных имеет вполне конкретные цели и задачи, а также использует достаточно стабильный набор методов. К основным из них относятся:

· нормализация;

· преобразование типов и форматов;

· сортировка;

· группировка;

· слияние;

· обработка временных рядов и пр.

На каждом этапе аналитического процесса имеются свои приоритетные цели трансформации, например, в системах оперативной обработки (OLTP) это обеспечение поддержки корректности форматов и типов данных, оптимизация процессов доступа и выгрузки данных. На этапе ETL-процесса трансформация производится с целью приведения данных в соответствие с моделью, которая используется в хранилище, а также обеспечения процесса консолидации данных и их загрузки в хранилище. И наконец, в аналитическом приложении производится непосредственная подготовка данных к анализу, объединение и выделение наиболее ценной информации, обеспечение корректной работы аналитических алгоритмов, методов и моделей.

Квантование значений. При выполнении этой операции осуществляется разбиение диапазона числовых значений на указанное количество интервалов определенным методом и замена каждого обрабатываемого значения на число, связанное с интервалом, к которому оно относится, либо на метку интервала. Интервалы разбиения включают в себя нижнюю границу, но не включают верхнюю, кроме последнего интервала, который включает в себя обе границы. Результатом преобразования может быть: номер интервала (от нуля до значения, на единицу меньшего количества интервалов), значение нижней или верхней границы интервала разбиения, среднее значение интервала разбиения, метка интервала. *Квантование* может быть осуществлено интервальным или квантильным методом.

Интервальное *квантование* подразумевает разбиение диапазона значений на указанное количество значений равной длины. Например, если значения в поле попадают в диапазон от 0 до 10, то при интервальном *квантовании* на 10 интервалов мы получим отрезки от 0 до 1, от 1 до 2 и т.д. При этом 0 будет относиться к первому интервалу, 1 - ко второму, а 9 и 10 - к десятому.

Квантильное *квантование* подразумевает разбиение диапазона значений на равновероятные интервалы, то есть на интервалы, содержащие равное (или, по крайней мере, примерно равное) количество значений. Нарушение равенства возможно только тогда, когда значения, попадающие на границу интервала, встречаются в набореданных несколько раз. В этом случае все они относятся к одному определенному интервалу и могут вызвать «перевес» в его сторону.

Табличная замена значений. В результате выполнения этой операции производится замена значений по таблице подстановки, которая содержит пары, состоящие из исходного и выходного значения. Например, 0 - «красный», 1 - «зеленый», 2 - «синий». Или «зима» - «январь», «весна» - «апрель», «лето» - «июль», «осень» - «октябрь». Для каждого значения исходного набора данных ищется соответствие среди исходных значений таблицы подстановки. Если соответствие найдено, то значение меняется на соответствующее выходное значение из таблицы подстановки. Если значение не найдено в таблице, оно может быть либо заменено значением, указанным для замены «по умолчанию», либо оставлено без изменений (если такое значение не указано).

Обработка временных рядов. Временным рядом называют серию величин, полученную через регулярные промежутки времени. Например, месячные показатели инфляции или значения температуры воздуха, измеряемые каждые 10 дней.

Скользящее окно. При решении некоторых задач, например, при прогнозировании временных рядов с помощью нейросети, требуется подавать на вход анализатора значения несколько смежных отсчетов из исходного набора данных. При этом эффективность реализации заметно повышается, если не выбирать данные каждый раз из нескольких последовательных записей, а последовательно расположить данные, относящиеся к конкретной позиции окна, в одной записи.

Преобразование даты. Разбиение даты необходимо для анализа всевозможных показателей за определенный период (день, неделя, месяц, квартал, год). Суть разбиения заключается в том, что на основе столбца с информацией о дате формируется другой столбец, в котором указывается, к какому заданному интервалу времени принадлежит строка данных. Тип интервала задается аналитиком, исходя из того, что он хочет получить, - данные за год, квартал, месяц, неделю, день или сразу по всем интервалам.

Группировка. Трудно делать какие-либо выводы по данным каждой записи в отдельности. Аналитику для принятия решения часто необходима сводная информация. Совокупные данные намного более информативны, тем более если их можно получить в разных разрезах. *Группировка* позволяет объединять записи по полям-измерениям, агрегируя данные в полях-фактах для дальнейшего анализа.

Разгруппировка. *Группировка* используется для объединения фактов по каким-либо измерениям. При этом под объединением понимается применение некоторой функции агрегации. Если в исходном наборе данных присутствовали какие-либо другие измерения, то теряется информация о значениях фактов в разрезе этих измерений. Алгоритм *разгруппировки* позволяет восстановить эти факты, но их значения восстанавливаются не точно, а пропорционально вкладу в сгруппированные значения.

Вопросы для самопроверки:

1. Для чего создаются хранилища данных?
2. Что служит источником данных при создании хранилищ?

3. Почему данные приходится консолидировать?
4. Каковы задачи консолидации данных?
5. В каких случаях применяется OLTP-технология?
6. Что такое СППР?
7. Какие основные отличия OLTP и СППР?
8. Какие основные требования к ХД?
9. Что такое многомерный куб?
10. Как проводится анализ многомерных кубов?
11. Чем в концепции OLAP отличаются факты от измерений?
12. Что включает в себя процесс трансформации данных?
13. Для чего используют временные ряды?
14. В чем состоят основные задачи ELT?

Литература по теме:

Основная литература:

1. Чубукова И.А. Data mining. – БИНОМ. Лаборатория знаний, 2008 г. §17 п.1–2, 26: OLAP-системы.

Дополнительная литература:

1. Орешков В.И., Паклин Н.Б. Бизнес-аналитика – от данных к знаниям. – Питер. гл. 3. Консолидация данных – ключевые понятия. http://www.cfin.ru/itm/olap/cons.shtml.

2. Степанов В.Г. Эконометрика. http://www.e-college.ru/xbooks/xbook019/book/index/index.html?go=part-011*page.htm.

3. Можно ли увидеть 88-е измерение? / Computerworld, 2002, №46. - http://www.osp.ru/cw/2002/46/59084/.

Напишите эссе на 1-2 страницы на тему:

Скачайте демо-версию программы Контур-стандарт http://www.rarus.nn.ru/products/olap/down.htm

1. Интерфейс *Контур-стандарт:* достоинства и недостатки.
2. Сравнение сводных таблиц MS Excel и динамических таблиц *Контур-стандарт*.

Тема 4. Очистка и предобработка информации

Цели и задачи изучения данной темы – получение общетеоретических знаний о процессах оценки качества данных, их очистки и предобработке. Серьезное и целенаправленное изучение четвертой темы познакомит студентов с алгоритмами очистки и восстановления данных.

В результате успешного изучения темы Вы:

Узнаете:

· о понятии «грязные данные»;
· о причинах возникновения грязных данных;
· о способах очистки данных;
· классификацию инструментов очистки и редактирования данных;
· о программном обеспечении очистки данных;
· основные функции инструментов очистки данных;
· классификацию ошибок в данных, которые возникают в результате использования средств очистки данных;
· об особенностях предобработки данных;
· о сокращении размерности.

Приобретете следующие профессиональные компетенции:

· способность оценивать причины возникновения грязных данных;
· способность к классификации инструментов очистки и редактирования данных;
· способность к оценке качества данных;
· умение использовать алгоритмы восстановления данных;
· применения средств снижения размерности.

В процессе освоения темы акцентируйте внимание на следующих ключевых понятиях:

Предобработка – процедура подготовки данных к анализу, в процессе которой они приводятся в соответствие с требованиями, определяемыми спецификой решаемой задачи.

Предобработка данных включает два направления:

· очистку;

· оптимизацию.

Предобработка данных является важнейшим этапом аналитического процесса, и ее элементы выполняются на всех его шагах, начиная от OLTP-систем и заканчивая аналитическим приложением.

Очистка данных производится с целью исключения факторов, снижающих качество данных и мешающих работе аналитических алгоритмов. Она включает обработку дубликатов, противоречий данных является важнейшим этапом, от качества выполнения которого зависит возможность получения качественных результатов всего процесса анализа. Кроме того, следует помнить, что на этап подготовки данных, по некоторым оценкам, может быть потрачено до 80% всего времени, отведенного на проект.

Качество данных – это критерий, определяющий полноту, точность, своевременность и возможность интерпретации данных. Данные могут быть высокого качества и низкого качества, последние – это так называемые грязные или «плохие» данные.

Данные высокого качества – это полные, точные, своевременные данные, которые поддаются интерпретации. Такие данные обеспечивают получение качественного результата: знаний, которые смогут поддерживать процесс принятия решений.

Пропущенные данные могут возникнуть:

· в случаях, когда они вообще не были собраны (например, при анкетировании скрыт возраст);

· некоторые атрибуты могут быть неприменимы для некоторых объектов (например, атрибут «годовой доход» не применим к ребенку).

Дубликатами называются записи с одинаковыми значениями всех атрибутов. Наличие дубликатов в наборе данных может являться способом повышения значимости некоторых записей. Такая необходимость иногда возникает для особого выделения определенных записей из набора данных. Однако в большинстве случаев, продублированные данные являются результатом ошибок при подготовке данных.

Выбросы – резко отличающиеся объекты или наблюдения в наборе данных.

Шумы и выбросы являются достаточно общей проблемой в анализе данных. Выбросы могут как представлять собой отдельные наблюдения, так и быть объединенными в некие группы. Задача аналитика – не только их обнаружить, но и оценить степень их влияния на результаты дальнейшего анализа. Если выбросы являются информативной частью анализируемого набора данных, используют робастные методы и процедуры.

Оптимизация данных, как элемент предобработки, включает снижение размерности входных данных, выявление и исключение незначащих признаков. Основное отличие оптимизации от очистки в том, что факторы, устраняемые в процессе очистки, существенно снижают точность решения задачи или делают работу аналитических алгоритмов невозможной. Проблемы, решаемые при оптимизации, адаптируют данные к конкретной задаче и повышают эффективность их анализа.

Факторный анализ представляет собой группу методов, используемых для сокращения числа переменных и их обобщения.

Фактор – латентная переменная, конструируемая таким образом, чтобы можно было объяснить корреляцию между набором имеющихся переменных.

Латентность – ключевое понятие анализа; оно означает не явность характеристик, раскрываемых при помощи методов факторного анализа.

Вопросы темы:

1. Очистка данных.
2. Оптимизация данных.

Теоретический материал по теме

Вопрос 1. Очистка данных.

Для того чтобы результат анализа данных был достоверным, необходимо иметь качественные данные. При накоплении данных по технологии OLTP нет времени и ресурсов для серьезной проверки и исправления ошибок.

Предобработка – процедура подготовки данных к анализу, в процессе которой они приводятся в соответствие с требованиями, определяемыми спецификой решаемой задачи.

Предобработка данных включает два направления:

· очистку;

· оптимизацию.

Предобработка данных является важнейшим этапом аналитического процесса, и ее элементы выполняются на всех его шагах, начиная от OLTP-систем и заканчивая аналитическим приложением

Очистка данных производится с целью исключения факторов, снижающих качество данных и мешающих работе аналитических алгоритмов. Она включает обработку дубликатов, противоречий данных является важнейшим этапом, от качества выполнения которого зависит возможность получения качественных результатов всего процесса анализа. Кроме того, следует помнить, что на этап подготовки данных, по некоторым оценкам, может быть потрачено до 80% всего времени, отведенного на проект.

Рассмотрим подробно, что же представляет собой этот процесс.

Определение и анализ требований к данным.

На этом этапе осуществляется так называемое моделирование данных, т.е. определение и анализ требований к данным, которые необходимы для осуществления анализа. При этом изучаются вопросы распределения пользователей (географическое, организационное, функциональное); вопросы доступа к данным, которые необходимы для анализа, необходимость во внешних и/или внутренних источниках данных; а также аналитические характеристики системы (измерения данных, основные виды выходных документов, последовательность преобразования информации и др.).

Сбор данных.

· наличие в организации хранилища данных делает анализ проще и эффективней, его использование, с точки зрения вложений, обходится дешевле, чем использование отдельных баз данных или витрин данных. Однако далеко не все предприятия оснащены хранилищами данных. В этом случае используются оперативные, справочные и архивные БД, т.е. данные из существующих информационных систем;

· информация из информационных систем руководителей, внешних источников, бумажных носителей, а также знания экспертов или результаты опросов;

· определение необходимого количества данных;

· количество записей (примеров) в наборе данных должно быть значительно больше количества факторов (переменных);

· набор данных должен быть репрезентативным и представлять как можно больше возможных ситуаций. Пропорции представления различных примеров в наборе данных должны соответствовать реальной ситуации.

Предварительная обработка данных.

Анализировать можно как качественные, так и некачественные данные. Результат будет достигнут и в том, и в другом случае. Для обеспечения качественного анализа необходимо проведение предварительной обработки данных, которая является необходимым этапом процесса Data Mining.

Оценивание качества данных. Данные, полученные в результате сбора, должны соответствовать определенным критериям качества. Таким образом, можно выделить важный подэтап процесса Data Mining – оценивание качества данных.

Качество данных – это критерий, определяющий полноту, точность, своевременность и возможность интерпретации данных. Данные могут быть высокого качества и низкого качества, последние – это так называемые грязные или «плохие» данные.

Данные высокого качества – это полные, точные, своевременные данные, которые поддаются интерпретации. Такие данные обеспечивают получение качественного результата: знаний, которые смогут поддерживать процесс принятия решений.

Грязные данные могут появиться по разным причинам:

· ошибка при вводе данных;

· использование иных форматов представления или единиц измерения;

· несоответствие стандартам;

· отсутствие своевременного обновления;

· неудачное обновление всех копий данных;

· неудачное удаление записей-дубликатов и т.д.

Необходимо оценить стоимость наличия грязных данных; другими словами, наличие грязных данных может действительно привести к финансовым потерям и юридической ответственности, если их присутствие не предотвращается или они не обнаруживаются и не очищаются

Типы грязных данных:

· грязные данные, которые могут быть автоматически обнаружены и очищены;
· данные, появление которых может быть предотвращено;
· данные, которые непригодны для автоматического обнаружения и очистки;
· данные, появление которых невозможно предотвратить.

Поэтому важно понимать, что специальные средства очистки могут справиться *не со всеми видами грязных данных*.

Рассмотрим наиболее распространенные виды грязных данных:

· пропущенные значения;
· дубликаты данных;
· шумы и выбросы.

Некоторые значения данных могут быть **пропущены** в связи с тем, что:

· данные вообще не были собраны (например, при анкетировании скрыт возраст);
· некоторые атрибуты могут быть неприменимы для некоторых объектов (например, атрибут «годовой доход» неприменим к ребенку).

Как мы можем поступить с пропущенными данными?

· Исключить объекты с пропущенными значениями из обработки.
· Рассчитать новые значения для пропущенных данных.
· Игнорировать пропущенные значения в процессе анализа.
· Заменить пропущенные значения на возможные значения.

Дублирование данных.

Набор данных может включать продублированные данные, т.е. дубликаты.

Дубликатами называются записи с одинаковыми значениями всех атрибутов. Наличие дубликатов в наборе данных может являться способом повышения значимости некоторых записей. Такая необходимость иногда возникает для особого выделения определенных записей из набора данных. Однако в большинстве случаев, продублированные данные являются результатом ошибок при подготовке данных.

Как мы можем поступить с продублированными данными?

Существует два варианта обработки дубликатов. При первом варианте удаляется вся группа записей, содержащая дубликаты. Этот вариант используется в том случае, если наличие дубликатов вызывает недоверие к информации, полностью ее обесценивает.

Второй вариант состоит в замене группы дубликатов на одну уникальную запись.

Шумы и выбросы.

Выбросы – резко отличающиеся объекты или наблюдения в наборе данных.

Шумы и выбросы являются достаточно общей проблемой в анализе данных. Выбросы могут как представлять собой отдельные наблюдения, так и быть объединенными в некие группы. Задача аналитика – не только их обнаружить, но и оценить степень их влияния на результаты дальнейшего анализа. Если выбросы являются информативной частью анализируемого набора данных, используют робастные методы и процедуры.

Достаточно распространена практика проведения двухэтапного анализа – с выбросами и с их отсутствием – и сравнение полученных результатов.

Визуализация данных позволяет представить данные, в том числе и выбросы, в графическом виде. Мы видим несколько наблюдений, резко отличающихся от других (находящихся на большом расстоянии от большинства наблюдений).

Очевидно, что результаты Data Mining на основе грязных данных не могут считаться надежными и полезными. Однако наличие таких данных не обязательно означает необходимость их очистки или же предотвращения появления. Всегда должен быть разумный выбор между наличием грязных данных и стоимостью и/или временем, необходимым для их очистки.

Очистка данных.

Очистка данных (data cleaning, data cleansing или scrubbing) занимается выявлением и удалением ошибок и несоответствий в данных с целью улучшения качества данных.

Метод очистки данных должен удовлетворять ряду критериев Он должен выявлять и удалять все основные ошибки и несоответствия, как в отдельных источниках данных, так и при интеграции нескольких

источников.

1. Метод должен поддерживаться определенными инструментами, чтобы сократить объемы ручной проверки и программирования, и быть гибким в плане работы с дополнительными источниками.

2. Очистка данных не должна производиться в отрыве от связанных со схемой преобразования данных, выполняемых на основе сложных метаданных.

3. Функции маппирования для очистки и других преобразований данных должны быть определены декларативным образом и подходить для использования в других источниках данных и в обработке запросов.

4. Инфраструктура технологического процесса должна особенно интенсивно поддерживаться для Хранилищ данных, обеспечивая эффективное и надежное выполнение всех этапов преобразования для множества источников и больших наборов данных.

На сегодняшний день интерес к очистке данных возрастает. Целый ряд исследовательских групп занимается общими проблемами, связанными с очисткой данных, в том числе, со специфическими подходами к Data Mining и преобразованию данных на основании сопоставления схемы. В последнее время некоторые исследования коснулись единого, более сложного подхода к очистке данных, включающего ряд аспектов преобразования данных, специфических операторов и их реализации.

Этапы очистки данных:

В целом, очистка данных включает следующие этапы

1. Анализ данных.
2. Определение порядка и правил преобразования данных.
3. Подтверждение.
4. Преобразования.
5. Противоток очищенных данных.

Анализ данных.

Подробный анализ данных необходим для выявления подлежащих удалению видов ошибок и несоответствий. Здесь можно использовать как ручную проверку данных или их шаблонов, так и специальные программы для получения метаданных о свойствах данных и определения проблем качества.

Определение порядка и правил преобразования данных.

В зависимости от числа источников данных, степени их неоднородности и загрязненности, данные могут требовать достаточно обширного преобразования и очистки. Иногда для отображения источников общей модели данных используется трансляция схемы; для Хранилищ данных обычно используется реляционное представление. Первые шаги по очистке могут уточнить или изменить описание проблем отдельных источников данных, а также подготовить данные для интеграции. Дальнейшие шаги должны быть направлены на интеграцию схемы/данных и устранение проблем множественных элементов, например, дубликатов. В некоторых случаях для очистки данных также применяются нейронные сети, обученные на обработку текстов, написанных в свободной форме. Для Хранилищ в процессе работы по определению ETL должны быть определены методы контроля и поток данных, подлежащий преобразованию и очистке.

Подтверждение.

На этом этапе определяется правильность и эффективность процесса и определений преобразования. Это осуществляется путем тестирования и оценивания, например, на примере или на копии данных источника, - чтобы выяснить, необходимо ли как-то улучшить эти определения. При анализе, проектировании и подтверждении может потребоваться множество итераций, например, в связи с тем, что некоторые ошибки становятся заметны только после проведения определенных преобразований.

Преобразования.

На этом этапе осуществляется выполнение преобразований либо в процессе ETL для загрузки и обновления Хранилища данных, либо при ответе на запросы по множеству источников.

Противоток очищенных данных.

После того как ошибки отдельного источника удалены, загрязненные данные в исходных источниках должны замениться на очищенные, для того чтобы улучшенные данные попали также в унаследованные приложения и в дальнейшем при извлечении не требовали дополнительной очистки. Для Хранилищ очищенные данные находятся в области хранения данных.

Вопрос 2. Оптимизация данных.

Оптимизация данных, как элемент предобработки, включает снижение размерности входных данных, выявление и исключение незначащих признаков. Основное отличие оптимизации от очистки в том, что факторы, устраняемые в процессе очистки, существенно снижают точность решения задачи или делают работу аналитических алгоритмов невозможной. Проблемы, решаемые при оптимизации, адаптируют данные к конкретной задаче и повышают эффективность их анализа.

Одним из широко применяемых методов оптимизации является факторный анализ.

Факторный анализ представляет собой группу методов, используемых для сокращения числа переменных и их обобщения.

Фактор – латентная переменная, конструируемая таким образом, чтобы можно было объяснить корреляцию между набором имеющихся переменных.

На (рис. 15) показана схема факторного анализа.

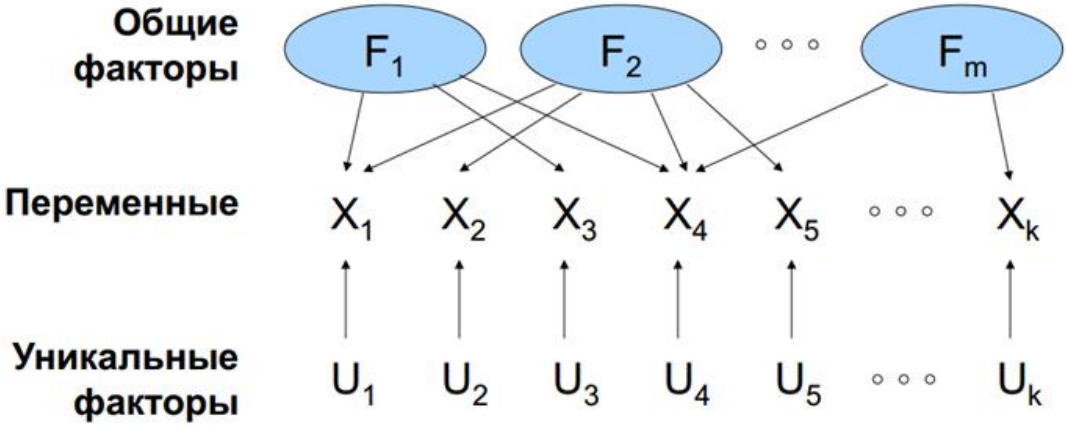

Рис. 15. Схема факторного анализа

В современной статистике под факторным анализом понимают совокупность методов, которые на основе реально существующих связей признаков, объектов или явлений позволяют выявлять *латентные* (скрытые и не доступные для непосредственного измерения) обобщающие характеристики организованной структуры и механизма развития изучаемых явлений или процессов.

Понятие латентности является ключевым и означает не явность характеристик, раскрываемых при помощи методов факторного анализа.

Идея, заложенная в основе факторного анализа, достаточно проста. В результате измерения мы имеем дело с набором элементарных признаков X_i, измеренных по нескольким шкалам. Это – *явные переменные*. Если признаки изменяются согласованно, то можно предположить существование определенных общих причин этой изменчивости, т.е. существование некоторых скрытых (латентных) факторов. Задача анализа – найти эти факторы.

Так как факторы представляют собой объединение определенных переменных, то из этого следует, что эти переменные связаны друг с другом, т.е. обладают корреляцией (ковариацией), причем большей между собой, чем с другими переменными, входящими в другой фактор. Методы отыскания факторов и основываются на использовании коэффициентов корреляции (ковариации) между переменными. Факторный анализ дает нетривиальное решение, т.е. решение нельзя предвидеть, не применяя специальную технику извлечения факторов. Это решение имеет большое значение для характеристики явления, так как вначале оно описывалось достаточно большим числом переменных, а в результате применения анализа оказалось, что его можно охарактеризовать меньшим числом других переменных – факторов.

На рисунке 16 показаны стадии факторного анализа.

Рис. 16. Стадии факторного анализа

Наличие высокого коэффициента корреляции межу факторами указывает на то, что либо один зависит от другого, либо они оба зависят от некого третьего **латентного** фактора. Найдя латентные факторы, можно снизить размерность и упростить процесс анализа.

Рис. 17. Пример

На рисунке 17 описаны данные, полученные из анкет покупателей. В результате проведения факторного анализа были выявлены 3 латентные переменные: отношение в семье, отношение к покупкам и отношение к рекламе. Таким образом, число факторов было снижено более чем в 2 раза.

Вопросы для самопроверки:

1. Затратная ли процедура очистки данных?
2. Каковы причины появления грязных данных?
3. Как классифицируют разные виды грязных данных?
4. Каковы этапы очистки данных?
5. Каковы алгоритмы очистки у
6. Чем предобработка данных отличается от очистки?
7. Для чего используют факторный анализ?
8. Чем опасны грязные данные?

Литература по теме:

Основная литература:
1. Чубукова И.А. Data mining. – БИНОМ. Лаборатория знаний, 2008 г. Гл. 2 §§18, 19.

Дополнительная литература:
1. Иванов О. Факторный анализ http://msu-students.ru/Stat_lectures/stat34.pdf

Тема 5. Анализ данных

Цели и задачи изучения данной темы – получение общетеоретических знаний о методах и алгоритмах анализа данных. При успешном усвоении материала данной темы студенты познакомятся с основными методами бизнес-анализа, их сильными и слабыми сторонами.

В результате успешного изучения темы Вы:
Узнаете:
· методы и алгоритмы поиска ассоциаций;
· о применении найденных ассоциативных правил в широком круге задач;
· о подходах к классификации событий и объектов;
· о методах поиска общности и различий в неструктурированных данных;
· о прогнозировании с помощью статистических методов.

Приобретете следующие профессиональные компетенции:
· применение алгоритма Apriori при поиске ассоциативных правил;
· применение дерева решений к задачи классификации;
· поиск кластеров расчет их характеристик;
· оценка связи факторов по корреляции Пирсона;
· регрессионный анализ и прогнозирование с помощью пакета анализа MS Excel.

В процессе освоения темы акцентируйте внимание на следующих ключевых понятиях:
Транзакция – это множество событий, которые произошли одновременно.
Рыночная корзина – это набор товаров, приобретенных покупателем в течение одной транзакции.
Поддержка – количество или процент транзакций, содержащих определенный набор данных.
Достоверность правила показывает, какова вероятность того, что из события A следует событие B. Правило «Из A следует B» справедливо с достоверностью c, если c% транзакций из всего множества, содержащих набор элементов A, также содержат набор элементов B.
Классификация – упорядоченное по некоторому принципу множество объектов, которые имеют сходные классификационные признаки (одно или несколько свойств), выбранных для определения сходства или различия между этими объектами.
Различают:
· **вспомогательную (искусственную) классификацию**, которая производится по внешнему признаку и служит для придания множеству предметов (процессов, явлений) нужного порядка;
· **естественную классификацию**, которая производится по существенным признакам, характеризующим внутреннюю общность предметов и явлений. Она является результатом и важным средством научного исследования, т.к. предполагает и закрепляет результаты изучения закономерностей классифицируемых объектов.

Скорость характеризует время, которое требуется на создание модели и ее использование.
Робастность, т.е. устойчивость к каким-либо нарушениям исходных предпосылок, означает возможность работы с зашумленными данными и пропущенными значениями в данных.
Интерпретируемость обеспечивает возможность понимания модели аналитиком.
Свойства классификационных правил:
· размер дерева решений;
· компактность классификационных правил.

Кросс-проверка – это процедура оценки точности классификации на данных из тестового множества, которое также называют кросс-проверочным множеством.

Вопросы темы:

1. Поиск ассоциаций.
2. Кластеризация и классификация.
3. Обзор статистических методов анализа и их применение.

Теоретический материал

Вопрос 1. Поиск ассоциаций.

Большинство организаций накапливают за время своей деятельности огромные объемы данных, но единственное что они хотят от них получить – это информация. Как можно узнать из данных о том, что нужно наиболее предпочтительным для организации клиентам, как разместить ресурсы наиболее эффективным образом или как минимизировать потери? Новейшая технология, адресованная к решению этих проблем – это технология data mining. Она использует сложный статистический анализ и моделирование для нахождения моделей и отношений, скрытых в базе данных – таких моделей, которые не могут быть найдены обычными методами.

Модель – это абстрактное описание реальности. Существуют два вида моделей: **предсказательные** и **описательные**. Первые используют один набор данных с известными результатами для построения моделей, которые явно предсказывают результаты для других наборов, а вторые описывают зависимости в существующих данных, которые в свою очередь используются для принятия управленческих решений или действий.

Рассмотрим построении модели на основе исторических сведений о покупательских трансакциях.

Транзакция – это множество событий, которые произошли одновременно.

Рыночная корзина – это набор товаров, приобретенных покупателем в рамках одной отдельно взятой транзакции.

Очень часто покупатели приобретают не один товар, а несколько. В большинстве случаев между этими товарами существует взаимосвязь. Так, например, покупатель, приобретающий макаронные изделия, скорее всего, захочет приобрести также кетчуп. Эта информация может быть использована для размещения товара на прилавках.

Часто встречающиеся приложения с применением ассоциативных правил:

· розничная торговля: определение товаров, которые стоит продвигать совместно; выбор местоположения товара в магазине; анализ потребительской корзины; прогнозирование спроса;

· перекрестные продажи: если есть информация о том, что клиенты приобрели продукты А, Б и В, то какие из них вероятнее всего купят продукт Г?

· маркетинг: поиск рыночных сегментов, тенденций покупательского поведения;

· сегментация клиентов: выявление общих характеристик клиентов компании, выявление групп покупателей;

· оформление каталогов, анализ сбытовых кампаний фирмы, определение последовательностей покупок клиентов (какая покупка последует за покупкой товара А);

· анализ Web-логов.

Приведем простой пример ассоциативного правила: покупатель, приобретающий банку краски, приобретет кисточку для краски с вероятностью 50%.

Рассмотрим набор транзакций из нижеприведенной таблицы.

№ транз.	Рыночная корзина
1.	Шампунь, Кондиционер, Бритва
2.	Кондиционер, Зубная Паста
3.	Кондиционер, Шампунь, Зубная Паста, Бритва
4.	Мыло, Зубная Паста
5.	Кондиционер, Шампунь, Зубная Паста, Бритва
6.	Мочалка, Зубная Паста

Поддержка – количество или процент транзакций, содержащих определенный набор данных.

В нашем случае Поддержка Зуб. пасты = 5 = 83%.

Достоверность правила показывает, какова вероятность того, что из события A следует событие B. Правило «Из A следует B» справедливо с достоверностью c, если c% транзакций из всего множества, содержащих набор элементов A, также содержат набор элементов B.

В нашем случае Достоверность «из Зуб. паста à Кондиционер» = 3/5 = 60%.

Достоверность «из Кондиционер àЗуб. паста» = 3/4=75%.

Существует множество алгоритмов поиска ассоциативных правил, например, AIS, STM, Apriori.

Рассмотрим работу алгоритма Apriori.

На первом этапе происходит формирование одноэлементных кандидатов. Далее алгоритм подсчитывает поддержку одноэлементных наборов. Наборы с уровнем поддержки меньше установленного,

Далее происходит формирование двухэлементных кандидатов, подсчет их поддержки и отсечение наборов с уровнем поддержки, меньшим заданного. Оставшиеся двухэлементные наборы товаров, считающиеся часто встречающимися двухэлементными наборами, принимают участие в дальнейшей работе алгоритма.

Отсечение кандидатов происходит на основе предположения о том, что у часто встречающегося набора товаров все подмножества должны быть часто встречающимися. Если в наборе находится подмножество, которое на предыдущем этапе было определено как нечасто встречающееся, этот кандидат уже не включается в формирование и подсчет кандидатов.

Вопрос 2. Классификация и кластеризация.

Задача классификации.

Классификация является наиболее простой и одновременно наиболее часто решаемой задачей Data Mining. Ввиду распространенности задач классификации необходимо четкое понимания сути этого понятия.

Классификация – упорядоченное по некоторому принципу множество объектов, которые имеют сходные классификационные признаки (одно или несколько свойств), выбранных для определения сходства или различия между этими объектами.

Классификация требует соблюдения следующих правил:

· в каждом акте деления необходимо применять только одно основание;

· деление должно быть соразмерным, т.е. общий объем видовых понятий должен равняться объему делимого родового понятия;

· члены деления должны взаимно исключать друг друга, их объемы не должны перекрещиваться;

· деление должно быть последовательным.

Различают:

· **вспомогательную (искусственную) классификацию**, которая производится по внешнему признаку и служит для придания множеству предметов (процессов, явлений) нужного порядка;

· **естественную классификацию**, которая производится по существенным признакам, характеризующим внутреннюю общность предметов и явлений. Она является результатом и важным средством научного исследования, т.к. предполагает и закрепляет результаты изучения закономерностей классифицируемых объектов.

В зависимости от выбранных признаков, их сочетания и процедуры деления понятий классификация может быть:

· **простой** – деление родового понятия только по признаку и только один раз до раскрытия всех видов. Примером такой классификации является дихотомия, при которой членами деления бывают только два понятия, каждое из которых является противоречащим другому (т.е. соблюдается принцип: «A и не A»);

· **сложной** – применяется для деления одного понятия по разным основаниям и синтеза таких простых делений в единое целое. Примером такой классификации является периодическая система химических элементов.

Под классификацией будем понимать отнесение объектов (наблюдений, событий) к одному из заранее известных классов.

Классификация – это закономерность, позволяющая делать вывод относительно определения характеристик конкретной группы. Таким образом, для проведения классификации должны присутствовать

признаки, характеризующие группу, к которой принадлежит то или иное событие или объект (обычно при этом на основании анализа уже классифицированных событий формулируются некие правила).

Классификация относится к стратегии обучения с учителем (supervised learning), которое также именуют контролируемым или управляемым обучением.

Задачей классификации часто называют предсказание категориальной зависимой переменной (т.е. зависимой переменной, являющейся категорией) на основе выборки непрерывных и/или категориальных переменных.

Например, можно предсказать, кто из клиентов фирмы является потенциальным покупателем определенного товара, а кто – нет, кто воспользуется услугой фирмы, а кто – нет, и т.д. Этот тип задач относится к задачам бинарной классификации, в них зависимая переменная может принимать только два значения (например, да или нет, 0 или 1).

Другой вариант классификации возникает, если зависимая переменная может принимать значения из некоторого множества предопределенных классов. Например, когда необходимо предсказать, какую марку автомобиля захочет купить клиент. В этих случаях рассматривается множество классов для зависимой переменной.

Классификация может быть **одномерной** (по одному признаку) и **многомерной** (по двум и более признакам).

Рассмотрим задачу классификации на простом примере. Допустим, имеется база данных о клиентах туристического агентства с информацией о возрасте и доходе за месяц. Есть рекламный материал двух видов: более дорогой и комфортный отдых и более дешевый, молодежный отдых. Соответственно, определены два класса клиентов: класс 1 и класс 2.

Определить, к какому классу принадлежит новый клиент и какой из двух видов рекламных материалов ему стоит отсылать.

Для наглядности представим нашу базу данных в двухмерном измерении (возраст и доход), в виде множества объектов, принадлежащих классам 1 (оранжевая метка) и 2 (серая метка). На рис. 18 приведены объекты из двух классов.

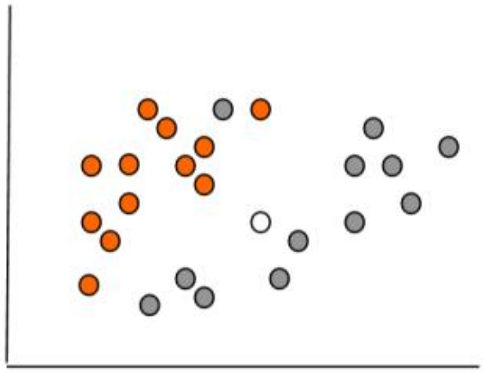

Рис. 18. Множество объектов базы данных в двухмерном измерении

Решение задачи будет состоять в том, чтобы определить, к какому классу относится новый клиент, на рисунке обозначенный белой меткой.

Процесс классификации.

Цель процесса классификации состоит в том, чтобы построить модель, которая использует прогнозирующие атрибуты в качестве входных параметров и получает значение зависимого атрибута.

Для проведения классификации с помощью математических методов необходимо иметь формальное описание объекта, которым можно оперировать, используя математический аппарат классификации. Таким описанием в нашем случае выступает база данных. Каждый объект (запись базы данных) несет информацию о некотором свойстве объекта.

Набор исходных данных (или выборку данных) разбивают на два множества: обучающее и тестовое.

Обучающее множество – множество, которое включает данные, использующиеся для обучения (конструирования) модели. Такое множество содержит входные и выходные (целевые) значения примеров. Выходные значения предназначены для обучения модели.

Тестовое множество также содержит входные и выходные значения примеров. Здесь выходные значения используются для проверки работоспособности модели.

Процесс классификации состоит из двух этапов: конструирования модели и ее использования.

1. Конструирование модели: описание множества предопределенных классов.

· Каждый пример набора данных относится к одному предопределенному классу.
· На этом этапе используется обучающее множество, на нем происходит конструирование модели.
· Полученная модель представлена классификационными правилами, деревом решений или математической формулой.

2. Использование модели: классификация новых или неизвестных значений.
· Оценка правильности (точности) модели: известные значения из тестового примера сравниваются с результатами использования полученной модели.
· Уровень точности – процент правильно классифицированных примеров в тестовом множестве.
· Тестовое множество, т.е. множество, на котором тестируется построенная модель, не должно зависеть от обучающего множества.

3. Если точность модели допустима, возможно использование модели для классификации новых примеров, класс которых неизвестен.

Методы, применяемые для решения задач классификации.

Для классификации используются различные методы. Выбор метода определяется особенностями структуры классов и получаемой модели. Основные из них:
· классификация с помощью деревьев решений;
· байесовская (наивная) классификация;
· классификация при помощи искусственных нейронных сетей;
· классификация методом опорных векторов;
· статистические методы, в частности, линейная регрессия;
· классификация при помощи метода ближайшего соседа;
· классификация CBR-методом;
· классификация при помощи генетических алгоритмов.

Классификация с помощью деревьев решений – наиболее простой и наглядный метод, но он не всегда работает. Подробнее об этом методе [2]. Модель строится в виде графа. Вершина графа – *корень*, *ветвление* происходит, когда при выбранных значениях параметра есть представители разных классов. Полная сортировка оканчивается *листом*.

Точность классификации: оценка уровня ошибок.

Оценка точности классификации может проводиться при помощи кросс-проверки. **Кросс-проверка** – это процедура оценки точности классификации на данных из тестового множества, которое также называют кросс-проверочным множеством. Точность классификации тестового множества сравнивается с точностью классификации обучающего множества. Если классификация тестового множества дает приблизительно такие же результаты по точности, как и классификация обучающего множества, считается, что данная модель прошла кросс-проверку.

Разделение на обучающее и тестовое множества осуществляется путем деления выборки в определенной пропорции, например обучающее множество – две трети данных и тестовое – одна треть данных. Этот способ следует использовать для выборок с большим количеством примеров. Если же выборка имеет малые объемы, рекомендуется применять специальные методы, при использовании которых обучающая и тестовая выборки могут частично пересекаться.

Оценивание классификационных методов.

Оценивание методов следует проводить, исходя из следующих характеристик: скорость, робастность, интерпретируемость, надежность.

Скорость характеризует время, которое требуется на создание модели и ее использование.

Робастность, т.е. устойчивость к каким-либо нарушениям исходных предпосылок, означает возможность работы с зашумленными данными и пропущенными значениями в данных.

Интерпретируемость обеспечивает возможность понимания модели аналитиком.

Свойства классификационных правил:
· размер дерева решений;
· компактность классификационных правил.

В [2] описан алгоритм дихотомической классификации на основе дерева решений.

Задача кластеризации.

Только что мы изучили задачу классификации, относящуюся к стратегии «обучение с учителем».

В этой части лекции мы введем понятия кластеризации, кластера, кратко рассмотрим классы методов, с помощью которых решается задача кластеризации, некоторые моменты процесса кластеризации, а также разберем примеры применения кластерного анализа.

Задача кластеризации сходна с задачей классификации, является ее логическим продолжением, но ее отличие в том, что классы изучаемого набора данных заранее не предопределены.

Синонимами термина « кластеризация « являются «автоматическая классификация «, «обучение без учителя» и «таксономия».

Кластеризация предназначена для разбиения совокупности объектов на однородные группы (кластеры или классы). Если данные выборки представить как точки в признаковом пространстве, то задача кластеризации сводится к определению «сгущений точек».

Цель кластеризации – поиск существующих структур.

Кластеризация является описательной процедурой, она не делает никаких статистических выводов, но дает возможность провести разведочный анализ и изучить «структуру данных».

Само понятие «кластер» определено неоднозначно: в каждом исследовании свои « кластеры «. Переводится понятие кластер (cluster) как «скопление», «гроздь».

Кластер можно охарактеризовать как группу объектов, имеющих общие свойства.

Характеристиками кластера можно назвать два признака:

· внутренняя однородность;

· внешняя изолированность.

Наибольшее применение кластеризация первоначально получила в таких науках как биология, антропология, психология. Для решения экономических задач кластеризация длительное время мало использовалась из-за специфики экономических данных и явлений.

В таблице приведено сравнение некоторых параметров задач классификации и кластеризации.

Сравнение классификации и кластеризации		
Характеристика	**Классификация**	**Кластеризация**
Контролируемость обучения	Контролируемое обучение	Неконтролируемое обучение
Стратегия	Обучение с учителем	Обучение без учителя
Наличие метки класса	Обучающее множество сопровождается меткой, указывающей класс, к которому относится наблюдение	Метки класса обучающего множества неизвестны
Основание для классификации	Новые данные классифицируются на основании обучающего множества	Дано множество данных с целью установления существования классов или кластеров данных

Кластеры могут быть непересекающимися, или эксклюзивными, и пересекающимися.

Следует отметить, что в результате применения различных методов кластерного анализа могут быть получены кластеры различной формы. Например, возможны кластеры»цепочного» типа, когда кластеры представлены длинными «цепочками», кластеры удлиненной формы и т.д., а некоторые методы могут создавать кластеры произвольной формы.

Различные методы могут стремиться создавать кластеры определенных размеров (например, малых или крупных) либо предполагать в наборе данных наличие кластеров различного размера.

Некоторые методы кластерного анализа особенно чувствительны к шумам или выбросам, другие – менее.

В результате применения различных методов кластеризации могут быть получены неодинаковые результаты, это нормально и является особенностью работы того или иного алгоритма.

Данные особенности следует учитывать при выборе метода кластеризации.

На сегодняшний день разработано более сотни различных алгоритмов кластеризации.

Приведем краткую характеристику подходов к кластеризации.

1. Алгоритмы, основанные на разделении данных, в том числе итеративные:

а) разделение объектов на k кластеров;

б) итеративное перераспределение объектов для улучшения кластеризации.

2. Иерархические алгоритмы:

а) агломерация: каждый объект первоначально является кластером, кластеры, соединяясь друг с другом, формируют больший кластер и т.д.;

б) методы, основанные на концентрации объектов;

в) основаны на возможности соединения объектов;

г) игнорируют шумы, нахождение кластеров произвольной формы.

3. Грид-методы (Grid-based methods):

а) квантование объектов в грид-структуры.

4. Модельные методы (Model-based):

а) использование модели для нахождения кластеров, наиболее соответствующих данным.

Оценка качества кластеризации.

Оценка качества кластеризации может быть проведена на основе следующих процедур:

· ручная проверка;

· установление контрольных точек и проверка на полученных кластерах;

· определение стабильности кластеризации путем добавления в модель новых переменных;

· создание и сравнение кластеров с использованием различных методов. Разные методы кластеризации могут создавать разные кластеры, и это является нормальным явлением. Однако создание схожих кластеров различными методами указывает на правильность кластеризации.

Процесс кластеризации.

Процесс кластеризации зависит от выбранного метода и почти всегда является итеративным. Полученные результаты требуют дальнейшей интерпретации, исследования и изучения свойств и характеристик объектов для возможности точного описания сформированных кластеров.

Применение кластерного анализа.

Кластерный анализ применяется в различных областях. Так, в медицине используется кластеризация заболеваний, лечения заболеваний или их симптомов, а также таксономия пациентов, препаратов и т.д. В археологии устанавливаются таксономии каменных сооружений и древних объектов и т.д. В маркетинге это может быть задача сегментации конкурентов и потребителей. В менеджменте примером задачи кластеризации будет разбиение персонала на различные группы, классификация потребителей и поставщиков, выявление схожих производственных ситуаций, при которых возникает брак. В медицине - классификация симптомов. В социологии задача кластеризации - разбиение респондентов на однородные группы.

Кластерный анализ в маркетинговых исследованиях.

В маркетинговых исследованиях кластерный анализ применяется достаточно широко - как в теоретических исследованиях, так и практикующими маркетологами, решающими проблемы группировки различных объектов. При этом решаются вопросы о группах клиентов, продуктов и т.д.

Так, одной из наиболее важных задач при применении кластерного анализа в маркетинговых исследованиях является анализ поведения потребителя, а именно: группировка потребителей в однородные классы для получения максимально полного представления о поведении клиента из каждой группы и о факторах, влияющих на его поведение.

Важной задачей, которую может решить кластерный анализ, является позиционирование, т.е. определение ниши, в которой следует позиционировать новый продукт, предлагаемый на рынке. В результате применения кластерного анализа строится карта, по которой можно определить уровень конкуренции в различных сегментах рынка и соответствующие характеристики товара для возможности попадания в этот сегмент.

Кластерный анализ также может быть удобен, например, для анализа клиентов компании. Для этого все клиенты группируются в кластеры, и для каждого кластера вырабатывается индивидуальная политика. Такой подход позволяет существенно сократить объекты анализа, и, в то же время, индивидуально подойти к каждой группе клиентов.

Вопрос 3. Обзор статистических методов анализа и их применение.

Статистический анализ включает большое разнообразие методов, даже для поверхностного знакомства с которыми объема одной лекции слишком мало. Цель данной темы – дать самое общее

представление о понятиях корреляции, регрессии, а также познакомиться с описательной статистикой. Примеры, рассмотренные в лекции, намеренно упрощены.

Корреляционный анализ применяется для количественной оценки взаимосвязи двух наборов данных, представленных в безразмерном виде. Корреляционный анализ дает возможность установить, ассоциированы ли наборы данных по величине. Коэффициент корреляции, всегда обозначаемый латинской буквой r, используется для определения наличия взаимосвязи между двумя свойствами.

Коэффициент корреляции Пирсона.

Коэффициент корреляции Пирсона r, который является безразмерным индексом в интервале от -1,0 до 1,0 включительно, отражает степень линейной зависимости между двумя множествами данных.

Связь между признаками (по шкале Чеддока) может быть сильной, средней и слабой. Тесноту связи определяют по величине коэффициента корреляции, который может принимать значения от -1 до +1 включительно. Критерии оценки тесноты связи показаны на рис. 19.

Величина коэффициента корреляции	0.1 - 0.3	0.3 - 0.5	0.5 – 0.7	0.7 – 0.9	0.9 – 1.0
Характеристика силы связи	слабая	умеренная	заметная	высокая	весьма высокая
		средняя		сильная	

Рис. 19. Количественные критерии оценки тесноты связи

Регрессионный анализ.

Основная особенность регрессионного анализа: при его помощи можно получить конкретные сведения о том, какую форму и характер имеет зависимость между исследуемыми переменными.

Последовательность этапов регрессионного анализа.

· Формулировка задачи. На этом этапе формируются предварительные гипотезы о зависимости исследуемых явлений.

· Определение зависимых и независимых (объясняющих) переменных.

· Сбор статистических данных. Данные должны быть собраны для каждой из переменных, включенных в регрессионную модель.

· Формулировка гипотезы о форме связи (простая или множественная, линейная или нелинейная).

· Определение функции регрессии (заключается в расчете численных значений параметров уравнения регрессии).

· Оценка точности регрессионного анализа.

· Интерпретация полученных результатов. Полученные результаты регрессионного анализа сравниваются с предварительными гипотезами. Оценивается корректность и правдоподобие полученных результатов.

· Предсказание неизвестных значений зависимой переменной.

При помощи регрессионного анализа возможно решение задачи прогнозирования и классификации. Прогнозные значения вычисляются путем подстановки в уравнение регрессии параметров значений объясняющих переменных. Решение задачи классификации осуществляется таким образом: линия регрессии делит все множество объектов на два класса, и та часть множества, где значение функции больше нуля, принадлежит к одному классу, а та, где оно меньше нуля, – к другому классу.

Определение функции регрессии.

Вторая задача сводится к выяснению действия на зависимую переменную главных факторов или причин, при неизменных прочих равных условиях, и при условии исключения воздействия на зависимую переменную случайных элементов. Функция регрессии определяется в виде математического уравнения того или иного типа.

Оценка качества модели.

Величина R-квадрат, называемая также мерой определенности, характеризует качество полученной регрессионной прямой. Это качество выражается степенью соответствия между исходными данными и регрессионной моделью (расчетными данными). Мера определенности всегда находится в пределах интервала [0;1].

В большинстве случаев значение R-квадрат находится между этими значениями, называемыми экстремальными, т.е. между нулем и единицей.

Если значение R-квадрата близко к единице, это означает, что построенная модель объясняет почти всю изменчивость соответствующих переменных. И наоборот, значение R-квадрата, близкое к нулю, означает плохое качество построенной модели.

Оценка неизвестных значений зависимой переменной.

Решение этой задачи сводится к решению задачи одного из типов:

· Оценка значений зависимой переменной внутри рассматриваемого интервала исходных данных, т.е. пропущенных значений; при этом решается задача интерполяции.

· Оценка будущих значений зависимой переменной, т.е. нахождение значений вне заданного интервала исходных данных; при этом решается задача экстраполяции.

Обе задачи решаются путем подстановки в уравнение регрессии найденных оценок параметров значений независимых переменных. Результат решения уравнения представляет собой оценку значения целевой (зависимой) переменной.

Имея уравнение регрессии, задача прогнозирования сводится к решению уравнения Y= F(x) с известными значениями x.

Вопросы для самопроверки:

1. Откуда берет знания информационная система?
2. В чем смысл поиска ассоциаций?
3. Где применяются найденные ассоциации?
4. Какова последовательность действий в процессе классификаций?
5. В чем отличие кластеризации от классификации?
6. Где используются методы классификации и кластеризации?
7. Что дают статистические методы?
8. Какую информацию несет коэффициент корреляции?

Литература по теме:

Основная литература:

1. Чубукова И.А. Data mining. – БИНОМ. Лаборатория знаний, 2008 г. Гл. 2 §§9, 10, 15.

Дополнительная литература:

1. Алексеева Т. В., Амириди Ю. В., Дик В. В. Информационные аналитические системы. – Синергия, 2013.

2. Решение специальных задач в приложениях «1С: Предприятие». Гл. 2. http://langslab.com/ebooks/spectask/solutions-ch2.

3. Паклин Н., Орешков В. Бизнес-аналитика. От данных к знаниям (+ CD-ROM).– Питер, 2013.

Тема 6. Data mining и Text mining

Цели и задачи изучения данной темы – получение общетеоретических знаний о технологиях обработки информации с целью получения знаний. Изучение последней темы познакомит студентов, в частности, с новыми возможностями использования информации, накопленной в Интернете, для оценки и прогнозирования экономического. политических и социального состояния общества.

В результате успешного изучения темы Вы:

Узнаете:

· компоненты и алгоритмы Data Mining;
· понятие искусственной нейронной сети;
· методы обучения нейронной сети;
· методы анализа неструктурированного текста;
· о методах повышения эффективности поисковой машины;
· какие знания спрятаны в больших объемах текстовой информации.

Приобретете следующие профессиональные компетенции:

· методы создания поисковых запросов;

· способы создания обучающих выборок;
· оценка качества обучения ИНС;
· оценка качества обработки запросов.

В процессе освоения темы акцентируйте внимание на следующих ключевых понятиях:

Итерация – это циклическая управляющая структура, она содержит выбор между альтернативами и следование избранной.

Адекватность модели – соответствие *модели* моделируемому объекту или процессу.

Эпоха – одна итерация в процессе обучения, включающая предъявление всех примеров из обучающего множества и, возможно, проверку качества обучения на контрольном множестве.

Обучающая выборка включает входные значения и соответствующие им выходные значения набора данных. В ходе обучения *нейронная сеть* находит некие зависимости выходных полей от входных.

Функция ошибок – это целевая функция, требующая минимизации в процессе управляемого обучения нейронной сети.

Переобучение, или чрезмерно близкая подгонка – излишне точное соответствие нейронной сети конкретному набору обучающих примеров, при котором сеть теряет способность к обобщению.

Text mining – интеллектуальный анализ неструктурированного текста, при котором происходит обнаружение принципиально новых, потенциально полезных паттернов, тенденций и взаимосвязей, способных помочь в принятии стратегических решений.

Контент-анализ – это содержательный анализ информационных потоков с целью получения необходимых качественных и количественных срезов, который производится постоянно, т.е. на протяжении не определяемого заранее промежутка времени.

Релевантность – соответствие документа запросу. Различают **техническую релевантность,** формальное соответствие запросу, и **семантическую релевантность** (соответствие смыслу, а не форме запроса).

Пертинентные документы – документы, действительно соответствующие потребности пользователя.

Полнота поиска $=p/P$, где p – найденные релевантные документы, P – все имеющиеся в системе релевантные документы.

Шум поиска $= н/(р+н)$, где $н$ – найденные нерелевантные документы, $р+н$ – все найденные документы.

Орант – местоимение, заменяющее субъект.

Вопросы темы:

1. Data mining.
2. Обучение нейронной сети.
3. Text mining.
4. Повышение эффективности информационно-поисковой машины.

Теоретический материал по теме

Вопрос 1. Data mining.

Data Mining является современной технологией анализа информации с целью нахождения в накопленных данных ранее неизвестных, нетривиальных и практически полезных знаний, необходимых для принятия оптимальных решений в различных областях человеческой деятельности.

В процессе «интеллектуальной раскопки данных» анализируется информация с целью автоматического поиска шаблонов (паттернов), характерных для каких-либо фрагментов неоднородных многомерных данных. В отличие от оперативной аналитической обработки данных (OLAP) в Data Mining задача по формулировке гипотез и выявлению необычных шаблонов переложена на информационную систему.

Процесс Data Mining является своего рода исследованием. Как любое исследование, этот процесс состоит из определенных этапов, включающих элементы сравнения, типизации, классификации, обобщения, абстрагирования, повторения.

Процесс Data Mining неразрывно связан с процессом принятия решений. Процесс Data Mining строит модель, а в процессе принятия решений эта модель эксплуатируется.

Исследование – это процесс познания определенной *предметной области*, объекта или явления с определенной целью. Процесс исследования заключается в наблюдении свойств объектов с целью

выявления и оценки важных, с точки зрения субъекта-исследователя, закономерных отношений между показателями данных свойств.

Любое исследование проходит последовательно через ряд этапов.

Этап 1. Выявление предметной области.

Решение любой задачи в сфере разработки программного обеспечения должно начинаться с изучения *предметной области*.

Предметная область – это мысленно ограниченная область реальной действительности, подлежащая описанию или моделированию и исследованию.

Предметная область состоит из объектов, различаемых по свойствам и находящихся в определенных отношениях между собой или взаимодействующих каким-либо образом.

Этап 2. Постановка задачи.

Постановка задачи Data Mining включает следующие шаги:

· формулировка задачи;

· формализация задачи.

Постановка задачи включает также описание статического и динамического поведения исследуемых объектов. Пример задачи. При продвижении нового товара на рынок необходимо определить, какая группа клиентов фирмы будет наиболее заинтересована в данном товаре.

Описание статики подразумевает описание объектов и их свойств.

Пример. Клиент является объектом. Свойства объекта «клиент»: семейное положение, доход за предыдущий год, место проживания.

Динамика описывает изменение свойств или отношений между ними во времени.

Этап 3. Моделирование.

В широком смысле слова *моделирование* – это научная дисциплина, цель которой, изучение методов построения и использования *моделей* для познания реального мира.

Моделирование – единственный к настоящему времени систематизированный способ увидеть варианты будущего и определить потенциальные последствия альтернативных решений, что позволяет их объективно сравнивать. *Моделирование* – достаточно популярный и эффективный метод исследования данных, который является основой анализа данных.

Моделирование как процесс представляет собой построение *модели* и изучение ее свойств, которые подобны наиболее важным, с точки зрения поставленной задачи.

Создание и использование Data Mining *модели* является ключевым моментом для начала понимания, осмысления и прогнозирования тенденций анализируемого объекта.

Построение *моделей* Data Mining осуществляется с целью исследования или изучения моделируемого объекта, процесса, явления и получения новых знаний, необходимых для принятия решений. Использование *моделей* Data Mining позволяет определить наилучшее решение в конкретной ситуации.

Построенные *модели* могут иметь различную сложность. Сложность построенной *модели* зависит от используемых методов, а также от сложности объекта, который анализируется.

Под ***сложным объектом*** понимается объект сложной структуры, который характеризуется большим количеством входных переменных, изменчивостью внутренней структуры и внешних факторов, нелинейностью взаимосвязей и др.

Классификация типов *моделей* в зависимости от характерных свойств, присущих изучаемому объекту или системе, такова

1. *динамические (системы*, изменяющиеся во времени) и статические;
2. стохастические и детерминированные;
3. непрерывные и дискретные;
4. линейные и нелинейные;
5. статистические; экспертные; *модели*, основанные на методах Data Mining;
6. *прогнозирующие* (классификационные) и описательные.

Этап 4. *Data Mining* является итеративным процессом.

Итерация – это циклическая управляющая структура, она содержит выбор между альтернативами и следование избранной.

Выбор между альтернативами в нашем случае – это этап *оценки модели*. Если *модель* приемлема, возможно ее использование.

Этапы подготовки данных, построения *модели*, *оценки модели* и выбора лучшей из них представляют собой цикл. Если по каким-либо причинам построенная *модель* оказалось неприемлемой, цикл повторяется и следует один из следующих этапов:

- подготовка данных (если причина некорректности *модели* в данных);
- построение *модели* (если причина некорректности во *внутренних параметрах* самой *модели*).

Иногда имеет смысл использовать несколько методов параллельно для возможности сравнения и анализа данных с различных точек зрения.

Этап 5. *Проверка модели* подразумевает *проверку* ее достоверности или *адекватности*. Эта *проверка* заключается в определении степени соответствия *модели* реальности. *Адекватность модели* проверяется путем тестирования.

Понятия достоверности и *адекватности* являются условными, поскольку мы не можем рассчитывать на полное соответствие *модели* реальному объекту, иначе это был бы сам объект, а не *модель*. Поэтому в процессе *моделирования* следует учитывать *адекватность* не *модели* вообще, а именно тех ее свойств, которые являются существенными с точки зрения проводимого исследования. В процессе *проверки модели* необходимо установить включение в *модель* всех существенных факторов. Сложность решения этой проблемы зависит от сложности решаемой задачи.

Проверка модели также подразумевает определение той степени, в которой она действительно помогает менеджеру при принятии решений.

Оценка модели подразумевает *проверку* ее правильности. *Оценка* построенной *модели* осуществляется путем ее тестирования.

Тестирование *модели* заключается в «прогонке» построенной *модели*, заполненной данными, с целью определения ее характеристик, а также в *проверке* ее работоспособности. Тестирование *модели* включает в себя проведение множества экспериментов. На вход *модели* могут подаваться выборки различного объема. С точки зрения статистики, точность *модели* увеличивается с увеличением количества исследуемых данных. Алгоритмы, являющиеся основой для построения *моделей* на сверхбольших базах данных, должны обладать свойством масштабирования.

Если в результате *моделирования* нами было построено несколько различных *моделей*, то на основании их *оценки* мы можем осуществить выбор лучшей из них.

Этап 6. Применение модели.

После тестирования, *оценки* и выбора *модели* следует этап применения *модели*. На этом этапе выбранная *модель* используется применительно к новым данным с целью решения задач, поставленных в начале процесса Data Mining. Для классификационных и *прогнозирующих моделей* на этом этапе прогнозируется целевой (выходной) атрибут

Этап 7. Коррекция и обновление модели.

По прошествии определенного установленного промежутка времени с момента начала использования *модели* Data Mining следует проанализировать полученные результаты, определить, действительно ли она «успешна» или же возникли проблемы и сложности в ее использовании.

Существует много причин, требующих обучить *модель* заново, т.е. обновить ее, чтобы отразить определенные изменения.

Основными причинами являются следующие:

- изменились входящие данные или их поведение;
- появились дополнительные данные для обучения;
- изменились требования к форме и количеству выходных данных;
- изменились цели бизнеса, которые повлияли на критерии принятия решений;
- изменилось внешнее окружение или среда (макроэкономика, политическая ситуация, научно-технический прогресс, появление новых конкурентов и товаров и т.д.).

Причины, перечисленные выше, могут обесценить допущения и исходную информацию, на которых основывалась *модель* при построении.

Погрешности в процессе Data Mining.

Процесс Data Mining может быть успешным и неуспешным. Использование Data Mining не является гарантией получения исключительно достоверных знаний и принятия на основе этих знаний абсолютно верных решений.

Построенная *модель* может обладать рядом погрешностей. Вот некоторые из них: недостоверные исходные допущения при построении *модели*; ограниченные возможности при сборе необходимых данных; неуверенность и страхи пользователя системы, и, в силу этого, слабое их применение; неоправданно высокая стоимость.

Наиболее распространенной погрешностями *модели* являются:

· **Неверные** или **недостоверные исходные допущения**. Некоторые допущения поддаются объективной предварительной *проверке*, другие не могут быть заранее проверены. Если *модель* Data Mining основана на допущениях, естественно, ее точность зависит от точности допущений. Если допущения предыдущих периодов при использовании *модели* не оправдались, т.е. оказались неточны, то следует отказаться от «продления» этих допущений на будущие периоды.

· **Неоправданно высокая стоимость.** В результате процесса Data Mining должна быть получена выгода (конечно, если речь не идет о научных исследованиях). Полученная прибыль должна оправдать расходы на процесс Data Mining, а это не только стоимость программного обеспечения для Data Mining, но и затраты на подготовку данных, обучение, консультирование и т.д. Стоимость проекта зависит от его длительности, типа конечного приложения, уровня подготовки пользователей, варианта внедрения (готовый продукт, разработка «под ключ», адаптация под конкретную задачу).

Вопрос 2. Нейронные сети.

Вернемся к процессу создания модели. Одной из самых распространенных методов создания модели в Data Mining – обучение нейронных сетей.

Нейронная сеть – модель или компьютерная программа, имитирующая работу биологической нейронной сети (рис. 20).

Свойства нейронной сети:

1. В сети распознают входной (рецепторный) слой, воспринимающий сигналы внешнего возбуждения (например, экран, на который подается видеоизображение), и выходной слой, определяющий результат решения задачи распознавания или принятия решений. Работа сети тактируется для имитации прохождения по ней возбуждения и управления им.

2. Каждый нейрон обрабатывает сигнальную информацию (это важнейший принцип логической нейронной сети!) в диапазоне от нуля до условной единицы. Исходные данные в виде сигналов поступают (от пользователя, от блока обработки ситуации на входе, от другой нейронной сети и т.д.) или формируются (например, с помощью видео ввода) на рецепторном слое.

3. Функции активации бывают различны, но просты по объему вычислений. В простейшем случае такая функция совпадает с линейной формой, где аргументы, показанные на рис. 20, связаны операцией вычитания. Часто удобно не вычитать порог, а только лишь сравнивать с ним указанную сумму.

4. Найденная взвешенная сумма, превысившая порог, или величина превышения порога, является величиной возбуждения нейрона либо определяет значение величины возбуждения (например, в некоторых моделях величина возбуждения всегда равна единице, отсутствие возбуждения соответствует нулю). В некоторых моделях допускают и отрицательную величину возбуждения. Значение возбуждения передается через ветвящийся аксон в соответствии со связями данного нейрона с другими нейронами.

5. По дендритам может передаваться как возбуждающее, так и тормозящее воздействие. Первое может соответствовать положительному значению веса синапсической связей, второе — отрицательному. В нейронной сети возможны обратные связи.

6. Нейронная сеть работает в двух режимах: в режиме обучения и в режиме распознавания (рабочем режиме).

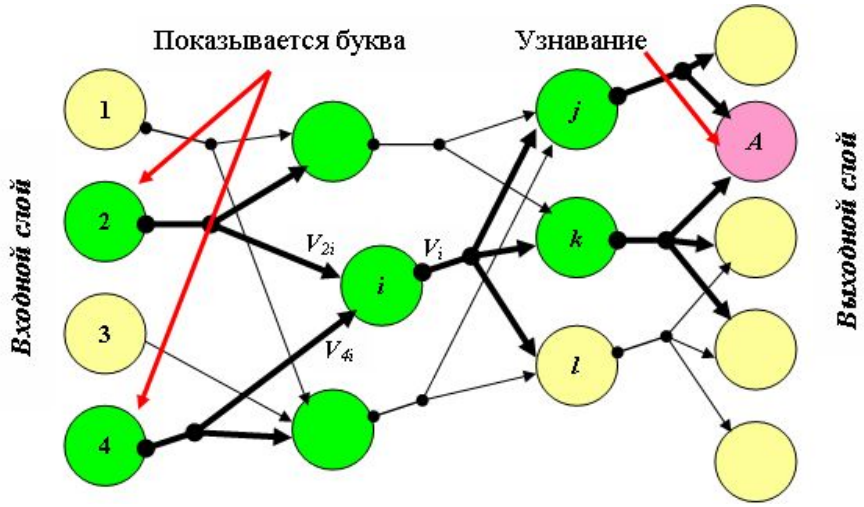

Рис. 20. Фрагмент нейронной сети

Алгоритм работы *нейронной сети* является итеративным, его шаги называют ***эпохами*** или циклами.

Эпоха – одна итерация в процессе обучения, включающая предъявление всех примеров из обучающего множества и, возможно, проверку качества обучения на контрольном множестве.

Процесс обучения осуществляется на обучающей выборке.

Обучающая выборка включает входные значения и соответствующие им выходные значения набора данных. В ходе обучения *нейронная сеть* находит некие зависимости выходных полей от входных.

Таким образом, перед нами ставится вопрос, какие входные поля (признаки) нам необходимо использовать. Первоначально выбор осуществляется эвристически, далее количество входов может быть изменено.

Сложность может вызвать вопрос о количестве наблюдений в наборе данных. И хотя существуют некие правила, описывающие связь между необходимым количеством наблюдений и размером сети, их верность не доказана.

Количество необходимых наблюдений зависит от сложности решаемой задачи. При увеличении количества признаков количество наблюдений возрастает нелинейно, эта проблема носит название «проклятие размерности». При недостаточном количестве данных рекомендуется использовать *линейную модель*.

Аналитик должен определить количество *слоев* в сети и количество нейронов в каждом *слое*.

Далее необходимо назначить такие значения весов и смещений, которые смогут минимизировать ошибку решения. Веса и смещения автоматически настраиваются таким образом, чтобы минимизировать разность между желаемым и полученным на выходе сигналами, которая называется ошибка обучения.

Ошибка обучения для построенной нейронной сети вычисляется путем сравнения выходных и целевых (желаемых) значений. Из полученных разностей формируется функция ошибок.

Функция ошибок – это целевая функция, требующая минимизации в процессе управляемого обучения нейронной сети.

С помощью функции ошибок можно оценить качество работы нейронной сети во время обучения. Например, часто используется сумма квадратов ошибок.

От качества обучения нейронной сети зависит ее способность решать поставленные перед ней задачи. Однако, этот процесс нельзя необоснованно увеличивать.

При обучении нейронных сетей часто возникает серьезная трудность, называемая проблемой переобучения.

Переобучение, или чрезмерно близкая подгонка – излишне точное соответствие нейронной сети конкретному набору обучающих примеров, при котором сеть теряет способность к обобщению.

Переобучение возникает в случае слишком долгого обучения, недостаточного числа обучающих примеров или переусложненной структуры нейронной сети.

Переобучение связано с тем, что выбор обучающего (тренировочного) множества является случайным. С первых шагов обучения происходит уменьшение ошибки. На последующих шагах с целью уменьшения ошибки (целевой функции) параметры подстраиваются под особенности обучающего множества. Однако при этом происходит «подстройка» не под общие закономерности ряда, а под особенности его части – обучающего подмножества. При этом точность прогноза уменьшается.

Один из вариантов борьбы с переобучением сети – деление обучающей выборки на два множества (обучающее и тестовое). На обучающем множестве происходит обучение нейронной сети. На тестовом множестве осуществляется проверка построенной модели. Эти множества не должны пересекаться.

С каждым шагом параметры модели изменяются, однако постоянное уменьшение значения целевой функции происходит именно на обучающем множестве. При разбиении множества на два мы можем наблюдать изменение ошибки прогноза на тестовом множестве параллельно с наблюдениями над обучающим множеством. Какое-то количество шагов ошибки прогноза уменьшается на обоих множествах. Однако на определенном шаге ошибка на тестовом множестве начинает возрастать, при этом ошибка на обучающем множестве продолжает уменьшаться (рис. 21). Этот момент считается концом реального или настоящего обучения, с него и начинается переобучение.

Описанный процесс проиллюстрирован на рис. 21.

Рис. 21. Определение момента прекращения обучения.

Вопрос 3. Text mining.

Text mining – нетривиальный процесс обнаружения действительно новых, потенциально полезных и понятных паттернов в неструктурированных текстовых данных: web-страницах, электронной почте, СМИ, нормативных документах, научных текстах.

Основную часть знаний аналитики получают в результате сравнения, анализа и синтеза информации из разрозненных фактов, размещенных в текстах. При работе с большими потоками документов процесс автоматического структурирования текстовой информации заменяет экспертный процесс выделения фактографической информации и объектов, выполняемый вручную. В статье рассматриваются примеры использования новых технологий извлечения знаний из текстов на русском языке, ориентированных на работу с большими хранилищами данных.

До 85% новых знаний аналитики получают, изучая тексты и в ближайшем будущем наиболее востребованными будут системы с максимально автоматизированными ETL-процессами (extract, transfer, load - сбор, выделение, преобразование, загрузка) структурирования контента. Другой важной чертой современных технологий является наличие функции оперативного анализа информации, полученной по запросу для выбора дальнейшего направления исследования документов (автопилотирование направления исследования), выполняемой с помощью методов интеллектуального анализа текста.

К наиболее актуальным средствам интеллектуального анализа текстов относятся технологии выделения фактографической информации об объектах с учетом анафорических ссылок (ссылочные местоимения на объект, поименованный в тексте ранее) на них [2]; нечеткий поиск; тематическое и тональное (точность и полнота) рубрицирование; кластерный анализ хранилищ и подборок документов; выделение ключевых тем; построение аннотаций; построение многомерных частотных распределений документов и их исследование с помощью OLAP технологии; использование методов интеллектуального анализа текста для определения направления исследования больших подборок документов и извлечения новых знаний.

В современных системах используется двухфазная технология аналитической обработки. В первой фазе (ETL) производится автоматизированный анализ отдельных документов, структуризация их контента и формирование хранилищ исходной и аналитической информации. Во второй фазе (OLAP, Text Mining и Data Mining) – извлечение в оперативном режиме знаний из хранилища или из полученной по запросу подборки документов.

В области обработки текстовой информации успешно работает ряд систем, отвечающих современным требованиям как по архитектуре, так и по функциональным возможностям. На наш взгляд, к интересным изделиям относятся: инструменты компании ClearForest, Convera RetrievalWare, Hummingbird KM, IBM Text Miner, инструменты компании IQMen, Inxight Smart Discovery Extraction Server, Ontos Miner, Oracle Text, ODB-Text, TextAnalyst, инструменты компании Smartware, XANALYS Link Explorer, «Аналитический курьер», инструменты компании Гарант-Парк-Интернет, «Медиалогия», «Система управления досье X-Files», и целый ряд других. Каждый из представленных изделий имеет свои преимущества в определенных технологиях. Авторы анализируют состояние технологий на примерах систем «Аналитический курьер» и «Система управления досье X-Files».

В ходе аналитической обработки происходит выделение текста фактографической информации об объекте, причем с учетом всех ссылок. Для этого сначала выделяются все предложения с упоминаниями об объекте (создается дайджест), в которых могут встречаться названия объекта («Иванов»), ссылки на него (**анафория**: «он», «который», …), а также обобщающие определения (**кореференты**: «воин», «семьянин», …). Нахождение и разрешение кореферентов и *анафор* дает увеличение объема дайджеста на 15-30%, а, значит, и объема фактографической информации. Для решения этих задач в программе «Аналитический курьер» за счет функции разрешения анафорических ссылок достигнуто приращение объема выделяемой фактографической информации (на 10-20%), повышение качества определения тональности публикаций -- на 30% (качество рубрицирования может быть оценено, например, как произведение точности и полноты рубрицирования), при увеличении времени обработки потока -- на 50%.

В начале исследования аналитики в первую очередь стремятся к полноте запроса, а не к его точности, поэтому объем релевантной подборки документов составляет сотни или тысячи единиц. Дальнейшее исследование проблемы производится уже после получения подборки документов с помощью кластерных, семантических карт или других методов. Такая технология работы аналитика сегодня типична как для работы в Сети, так и при работе со специализированными системами. Русский язык плохо поддаётся описанию формализмами различных уровней: морфологией, синтаксисом, семантикой. Например, для идентификации морфологических признаков лексемы на русском языке необходимо для снятия омонимии выполнить также предсинтаксический анализ предложения, и т.д. В любом случае реализации этих формализмов используют нечеткую модель анализа текста.

К наиболее актуальным направлениям извлечения знаний из текста на сегодня относятся:

1. *Аналитическая обработка фактов. Ведение досье.*
2. *Извлечение и структурирование фактографической информации.*
3. *Поиск информации по запросам на естественном языке с использованием тезаурусов.*
4. *Направления поиска информации, объектов в хранилище документов, в подборке документов.*
5. *Аннотирование документов, построение дайджестов по объектам.*
6. *Проведение тематического анализа документов (кластеризация и рубрицирование).*
7. *Построение и динамический анализ семантической структуры текстов.*
8. *Выделение ключевых тем и информационных объектов.*
9. *Определение общей и объектной тональности сообщений.*
10. *Исследование частотных характеристик текстов.*

Вопрос 4. Повышение эффективности информационно-поисковой машины.

Функция информационной поисковой системы состоит в выделении из поискового массива таких документов, которые содержат информацию, удовлетворяющую *информационную потребность* пользователя.

Документы, действительно соответствующие потребности пользователя, называются *пертинентными*. Сама информационная потребность представляет собой весьма сложное психическое явление, и проблема повышения степени пертинентности выдачи оказывается не только трудной для достижения, но её даже трудно чётко поставить как практическую задачу.

Документы, соответствующие запросу, называются **релевантными**. Однако суждение о релевантности будет зависеть от того, кто это суждение выносит. Таким образом, надо различать **техническую релевантность** и **семантическую релевантность** (соответствие смыслу, а не форме запроса).

Обработка запроса системой характеризуется:

Полнота поиска =р/Р, где *р* – найденные релевантные документы, Р – все имеющиеся в системе релевантные документы.

Шум = н/(р+н), где *н* – найденные нерелевантные документы.

Доказано, что увеличение полноты приводит к увеличению шума и наоборот.

Для увеличения *пертинентности* поисковые машины используют подсказки (*асессоры)*, соответствующие наиболее часто встречающимся запросам.

Одним из направлений повышения пертинентности используются плагины, отслеживающие движение пользователя по ссылкам. Потом вся статистика обрабатывается и при необходимости корректируется в зависимости от стиля формулировки запроса, новостных лент и пр. На основании этого происходит ранжирование списка ссылок. Ранжирование очень важно, так как 75% пользователей не доходит дальше первых трех документов из списка.

Вопросы для самопроверки:

1. Технологии получения знаний.
2. Что такое искусственная нейронная сеть?
3. Как происходит ее обучение?
4. Какие опасности таит неправильное обучение ИНС?
5. Как подготовить неструктурированный текст для анализа?
6. Какие знания можно получить из больших объемов текстов?
7. Как оценивается качество поисковой машины?

Литература по теме:

Основная литература:

1. Чубукова И.А. Data mining. – БИНОМ. Лаборатория знаний, 2008 г. Гл. 2 §22– 25.

Дополнительная литература:

1. Михаилян А. Некоторые методы автоматического анализа естественного языка, используемые в промышленных продуктах. http://citforum.ru/programming/digest/avtestlang.shtml

2. Чапайкина Н. Семантический анализ текстов. Основные положения [Текст] / Н. Е. Чапайкина // Молодой ученый. – 2012. – №5. – С. 112-115.

3. Большакова Е.И. и др. Автоматическая обработка текстов на естественном языке и компьютерная лингвистика: учеб. пособие. – М.: МИЭМ, 2011. – 272 с.

4. П. Браславский, И. Колычев. Автоматическое реферирование веб-документов с учётом запроса. Грант ООО «Яндекс» №102707.

5. Киселёв С.Л. Модель информационной системы бизнес-разведки. Открытые системы, #05-06/2005.

6. Илья Сегалович, Михаил Маслов, Денис Нагорнов. Как работают новые Яндекс. Новости.

7. Удо Хан, Индерджиет Мани. Системы автоматического реферирования. Открытые системы,#12/2000.

8. Шаров С.А. «Средства компьютерного представления лингвистической информации» // М: РНИИИИ, 1996.

9. Сэлтон Г. Автоматическая обработка, хранение и поиск информации. М. Сов. радио, 1973, 560 стр.

Напишите небольшое эссе (объемом в 1-2 страницы) по одному из перечисленных ниже вопросов:

1. Рынок систем поиска знаний.
2. Свободные ресурсы обработки знаний.
3. Интерфейс обучения и использования искусственной нейронной сети свободного доступа.
4. Text mining для … – любой вариант использования в одной из предметных областей.
5. Персональные данные и Text mining: нарушаются ли права личности?
6. Data mining и игровые технологии.
7. Нейронные сети для домашних роботов.
8. Text mining и социальные сети.
9. «Одноклассники» как источник знаний.
10. Мои претензии к поисковым системам.
11. Text mining и политология.
12. Text mining и политтехнологии.
13. Data mining в решении психологических задач.