Лекция 1 - Технологии анализа данных. Методы и средства извлечения знаний

1.1. ВВЕДЕНИЕ В АНАЛИЗ ДАННЫХ

С 1960-х гг. информационно-коммуникационные технологии (ИКТ) последовательно эволюционировали от простых систем обработки файлов до сложных, мощных систем управления базами данных (БД). Исследования в области БД с 1970-х гг. смещались от ранних иерархических и сетевых баз данных к реляционным системам управления базами данных (СУБД), инструментам моделирования данных, а также к вопросам индексирования и организации данных. Пользователи получили гибкий и удобный интерфейс доступа к данным с помощью языков запросов (типа SQL), пользовательские интерфейсы, управление транзакциями и т.п. При этом создаваемые и поддерживаемые БД имели преимущественно ограниченный регистрирующий характер, поддерживая рутинные операции линейного персонала. Основными требованиями к таким системам были обеспечение транзакционности и оперативность выполнения всех изменений.

Технология баз данных, начиная с середины 1980-х гг., характеризовалась популяризацией, широким внедрением и концентрацией исследовательских усилий на новые, все более мощные СУБД. Появились новые модели данных, такие как объектно-ориентированные, объектно-реляционные, дедуктивные модели. Возникали различные предметно-ориентированные базы данных и СУБД (пространственные, временные, мультимедийные, научные и пр.). Эффективные методы онлайновой обработки транзакций *(on-line transaction processing - OLTP'")* внесли большой вклад в эволюцию и широкое внедрение реляционной технологии в качестве одного из главных универсальных инструментов эффективного хранения, извлечения и управления большими объемами структурированных данных реляционных СУБД.

С развитием сети Интернет получили развитие и вопросы построения распределенных баз данных, создания распределенных глобальных информационных систем. Многократно возросла интенсивность формирования и архивирования различных данных, за которыми следовало развитие масштабируемых программно-аппаратных комплексов, дорогостоящих мощных и недорогих пользовательских компьютеров и накопителей данных.

Все это способствовало всплеску развития индустрии ИКТ и сделало огромное количество баз данных доступными для хранения разнородной информации в значительных объемах и управления транзакциями в них. При этом все больше возникала потребность анализа имеющихся данных в разновременном аспекте, с возможностью построения произвольных запросов, при условии обработки сверхбольших объемов данных, полученных, в том числе, из различных регистрирующих БД. Использование для этих задач традиционных регистрирующих систем и БД крайне затруднительно. Например, в

регистрирующей системе информация актуальна исключительно на момент обращения к БД, а в следующий момент времени по тому же запросу можно ожидать другой результат. Интерфейс таких систем рассчитан на проведение определенных стандартизованных операций и возможности получения результатов на нерегламентированный произвольный запрос ограничены. Возможности обработки больших массивов данных также могут быть ограничены вследствие ориентации СУБД на нормализованные данные, характерные для стандартных реляционных регистрирующих БД.

Ответом на возникшую потребность стало появление новой технологии организации баз данных - технологии *хранилищ Данных* (англ. *Data Warehouse*[1]), предполагающей некоторую предварительную обработку данных и их интеграцию, а также онлайновую аналитическую обработку (англ. *On-Line Analytical Processing, OLAP*[2]).

Несмотря на очевидную пользу такого инструмента анализа данных, он ориентирован на хорошо нормализованные табличные данные и не предполагает использование целого ряда дополнительного аналитического инструментария типа классификации, кластеризации, регрессионного анализа, моделирования, прогнозирования и интерпретации многомерных данных и т.п.

Таким образом, сегодня наблюдается высокий уровень развития масштабируемой аппаратно-программной ИКТ инфраструктуры, позволяющей увеличивать и без того значительные архивы данных. Имеется достаточно существенный задел в области компьютерных наук и информационных технологий, разработаны теория и прикладные аспекты теории вероятности и математической статистики. Однако при этом следует признать, что присутствует заметный *избыток данных*[3] при *дефиците информации*[4] *и знаний*[5]. Быстро растущие объемы накопленных и пополняемых (автоматически, а не людьми - как это было когда-то) архивов данных пока существенно превышают способности человека в их практически полезной обработке. Для обострения этого тезиса иногда говорят, что «*большие базы данных стали могилами, которые редко посещаются*». Как следствие, важные решения порой принимаются не на основе аналитических выводов из информативных БД, а на основе интуиции человека, не имеющего подходящих инструментов для извлечения полезных знаний из имеющихся огромных объемов данных.

Поэтому в последние годы стремительное развитие получила область *Data Mining* (в отечественной литературе наиболее используемая аналогия -

---

[1] Предметно-ориентированная информационная база данных, главным образом предназначенная для поддержки принятия решений с помощью отчетов.

[2] Технология анализа данных, предполагающая подготовку агрегированной структурированной многомерной информации на основе больших массивов данных *(OLAP-куба),* используемой в реляционной БД при построении сложных многотабличных запросов.

[3] Под *Данными* будем понимать представление некоторых фактов в формализованном виде, пригодном для хранения, обработки и передачи.

[4] Под *информацией* будем понимать сведения в любой форме; в отличие от данных, информация имеет некоторый *контекст.*

[5] Под *знаниями* будем понимать совокупность информации о мире, свойствах объектов, закономерностях процессов и явлений, а также правилах их использования *для принятия решений.*

интеллектуальный анализ Данных, ИАД), направленная на поиск и разработку методов извлечения из имеющихся данных *знаний*, позволяющих принимать на их основе конкретные, в высокой степени обоснованные, практически полезные управленческие решения.

На рисунке приведен пример обобщенного иерархического представления методологий обработки данных, начиная от интеграции разнородных источников данных и завершая использованием методов *Data Mining* для принятия управленческих решений.

1.2. МОДЕЛИ АНАЛИЗА ДАННЫХ

Предсказательные модели

Предсказательные (predictive) модели строятся на основании набора данных с известными результатами. Они используются для предсказания результатов на основании других наборов данных. При этом, естественно, требуется, чтобы модель работала максимально точно, была статистически значима и оправданна и т. д.

К таким моделям относятся следующие:

- модели классификации— описывают правила или набор правил, в соответствии с которыми можно отнести описание любого нового объекта к одному из классов. Такие правила строятся на основании информации о существующих объектах путем разбиения их на классы;
- модели последовательностей— описывают функции, позволяющие прогнозировать изменение непрерывных числовых параметров. Они строятся на основании данных об изменении некоторого параметра за прошедший период времени.

Описательные модели

Описательные (descriptive) модели уделяют внимание сути зависимостей в

наборе данных, взаимному влиянию различных факторов, т. е. построению эмпирических моделей различных систем. Ключевой момент в таких моделях— легкость и прозрачность для восприятия человеком. Возможно, обнаруженные закономерности будут специфической чертой именно конкретных исследуемых данных и больше нигде не встретятся, но это все равно может быть полезно, и потому должно быть известно.

К таким моделям относятся следующие виды:

- регрессионные модели— описывают функциональные зависимости между зависимыми и независимыми показателями и переменными в понятной человеку форме. Необходимо заметить, что такие модели описывают функциональную зависимость не только между непрерывными числовыми параметрами, но и между категориальными параметрами;

- модели кластеров— описывают группы (кластеры), на которые можно разделить объекты, данные о которых подвергаются анализу. Группируются объекты (наблюдения, события) на основе данных (свойств), описывающих сущность объектов. Объекты внутри кластера должны быть "похожими" друг на друга и отличаться от объектов, вошедших в другие кластеры. Чем сильнее "похожи" объекты внутри кластера и чем больше отличий между кластерами, тем точнее кластеризация;

- модели исключений — описывают исключительные ситуации в записях (например, отдельных пациентов), которые резко отличаются чем-либо от основного множества записей (группы больных). Знание исключений может быть использовано двояким образом. Возможно, эти записи представляют собой случайный сбой, например ошибки операторов, вводивших данные в компьютер. Характерный случай: если оператор, ошибаясь, ставит десятичную точку не в том месте, то такая ошибка сразу дает резкий "всплеск" на порядок. Подобную "шумовую" случайную составляющую имеет смысл отбросить, исключить из дальнейших исследований, поскольку большинство методов, которые будут рассмотрены, очень чувствительно к наличию "выбросов"— резко отличающихся точек, редких, нетипичных случаев. С другой стороны, отдельные, исключительные записи могут представлять самостоятельный интерес для исследования, т. к. они могут указывать на некоторые редкие, но важные аномальные заболевания. Даже сама идентификация этих записей, не говоря об их последующем анализе и детальном рассмотрении, может оказаться очень полезной для понимания сущности изучаемых объектов или явлений;

- итоговые модели— выявление ограничений на данные анализируемого массива. Например, при изучении выборки данных по пациентам не старше 30 лет, перенесшим инфаркт миокарда, обнаруживается, что все пациенты, описанные в этой выборке, либо курят более 5 пачек сигарет в день, либо имеют вес не ниже 95 кг. Подобные ограничения важны для понимания данных массива, по сути

дела это новое знание, извлеченное в результате анализа. Таким образом, построение итоговых моделей заключается в нахождении каких-либо фактов, которые верны для всех или почти всех записей в изучаемой выборке данных, но которые достаточно редко встречались бы во всем мыслимом многообразии записей такого же формата и, например, характеризовались бы теми же распределениями значений полей. Если взять для сравнения информацию по всем пациентам, то процент либо сильно курящих, либо чрезмерно тучных людей будет весьма невелик. Можно сказать, что решается как бы неявная задача классификации, хотя фактически задан только один класс, представленный имеющимися данными;

- ассоциативные модели— выявление закономерностей между связанными событиями. Примером такой закономерности служит правило, указывающее, что из события *X* следует событие У. Такие правила называются ассоциативными.

Для построения рассмотренных моделей используются различные методы и алгоритмы Data Mining. Ввиду того, что технология Data Mining развивалась и развивается на стыке таких дисциплин, как статистика, теория информации, машинное обучение и теория баз данных, вполне закономерно, что большинство алгоритмов и методов Data Mining были разработаны на основе различных технологий и концепций. Далее рассмотрим технологии, наиболее часто реализуемые методами Data Mining.

1.3. ЭТАПЫ ИНТЕЛЛЕКТУАЛЬНОГО АНАЛИЗА ДАННЫХ

Выделяют следующие типовые этапы, сопровождающие решение задач интеллектуального анализа данных:

1. Анализ предметной области, формулировка целей и задач исследования.
2. Извлечение и сохранение данных.
3. Предварительная обработка данных:

- очистка (англ. *cleaning;* исключение противоречий, случайных выбросов и помех[6], пропусков);
- интеграция (англ. *integration*; объединение данных из нескольких возможных источников в одном хранилище);
- преобразование (англ. *transformation*; может включать агрегирование и сжатие данных, дискретизацию атрибутов и сокращение размерности и т. п.).

4. Содержательный анализ данных методами *Data Mining* (установление общих закономерностей или решение более конкретных, частных задач).
5. Интерпретация полученных результатов с помощью их представления в удобном формате (визуализация и отбор полезных паттернов, формирование информативных графиков и / или таблиц).
6. Использование новых знаний для принятия решений.

[6] Если они сами не являются предметом анализа в данном случае.

1.4. ОБЩИЕ ТИПЫ ЗАКОНОМЕРНОСТЕЙ ПРИ АНАЛИЗЕ ДАННЫХ

Как правило, выделяют пять стандартных типов закономерностей, которые позволяют относить используемые методы к методам *Data Mining*:

1. Ассоциация.
2. Последовательность.
3. Классы.
4. Кластеры.
5. Временные ряды.

Ассоциация (англ. *Association*) имеет место в случае, если несколько событий связаны друг с другом. Например, исследование показывает, что 75% покупателей, приобретавших кукурузные чипсы, приобретают и «колу». Это ассоциация позволяет предложить скидку за такой тип продуктового «комплекта» и, возможно, увеличить тем самым объемы продаж.

В случае если несколько событий связаны друг с другом во времени, то имеет место тип зависимости, именуемый *последовательностью* (англ. *Sequential Patterns*). Например, после покупки дома в 45% случаев в течение месяца приобретается и новая кухонная плита, а в пределах двух недель 60% новоселов обзаводятся холодильником.

Закономерность *классы* (англ. *Classes*) появляется в случае, если имеется несколько заранее сформированных классов (групп, типов) объектов. Отнесение нового объекта к какому-либо из существующих классов выполняется путем *классификации*. Закономерность *кластеры* (англ. *Clusters*) отличается тем, что классы (группы, типы) заранее не заданы, а их количество и состав определяются автоматически в результате процедуры *кластеризации*.

Хранимая ретроспективная информация позволяет определить еще одну закономерность, заключающуюся в поиске существующих *временных рядов* (англ. *Time Series*) и прогнозировании динамики значений в них на будущие периоды времени.

1.5. ГРУППЫ ЗАДАЧ АНАЛИЗА ДАННЫХ

Наряду с поиском самых общих типов закономерностей, которые могут присутствовать в данных, также выделяют группы более конкретных, частных задач анализа данных. Несмотря на обширную сферу применения *Data Mining* в бизнесе, медицине или государственном управлении, подавляющее большинство этих задач может быть объединено в сравнительно небольшое число групп (табл. 1).

Т а б л и ц а 1 Основные группы задач анализа данных

Группа задач (англ.)	**Аналог в отечественной литературе**	**Пояснение**	**Пример задачи**
Classification and Prediction	Классификация и прогнозирование	Индуктивно разрабатывается обобщенная модель или формулируется некоторая гипотеза, описывающая принадлежность объектов к соответствующим классам	Предсказание роста объемов продаж на основе текущих значений, отнесения претендента на кредит к известным классам кредитоспособности, выявление лояльных или нелояльных держателей кредитных карт, классификация стран по климатическим зонам и т.п.
Clustering	Кластеризация	Выделение некоторого количества групп, имеющих сходные в некотором смысле признаки. Основной принцип - максимизация межклассового и минимизация внутриклассового расстояния	Обнаружение новых сегментов рынка, совершенствование рекламных стратегий для различных групп потребителей
Associations, Link Analysis	Ассоциации, анализ взаимозависимостей	Поиск интересных ассоциаций и / или корреляционных связей	95% покупателей автомобильных шин и автоаксессуаров также приобретали пакет сервисного обслуживания автомобиля, 80% покупателей газировки приобретают и «воздушную» кукурузу
Visualization	Визуализация	С использованием графических методов визуализации информации создается графический образ анализируемых	Визуализация некоторых зависимостей с использованием 2D- и 3D- измерений
		данных, отражающий имеющиеся в данных интересные закономерности	
Summarization	Подведение итогов	Интегральное (генерализованное) описание конкретных групп объектов из анализируемого набора данных	Суммирование данных сетевого трафика при оценке эффективности каналов связи, подготовка краткого реферата по тексту значительного объема, визуализация многомерных данных большого объема
Deviation (Anomaly) Detection, Outlier Analysis	Определение и анализ отклонений и / или выбросов в данных	Обнаружение фрагментов данных, существенно отличающихся от общего множества данных,	Применимо при анализе наличия шума / ошибок, а также при выявлении мошеннических действий

		выявление нехарактерных паттернов (шаблонов)	
Estimation	Оценивание	Предсказание непрерывных значений признака	Оценка производительности процессора на определенных задачах по ряду параметров процессора, оценка числа детей в семье по уровню образования матери, оценка дохода семьи по количеству в ней автомобилей, оценка стоимости недвижимости в зависимости от ее удаленности от бизнес центра
Feature Selection, Feature Engineering	Отбор значимых признаков	Применяется при анализе признаковых пространств большой размерности путем сокращения размерности и / или выбора значимых признаков с трансформацией признакового пространства или без трансформации	Как правило, применяется как вспомогательный метод на этапе предварительной обработки данных, а также для повышения эффективности методов визуализации в многомерных признаковых пространствах

1.6. КЛАССИФИКАЦИЯ МЕТОДОВ

Существует большое количество различных оснований для стратификации, категоризации, классификации значительного количества существующих и вновь разрабатываемых методов *Data Mining*. Например, можно встретить классификации по принципу работы с исходными обучающими данными (подвергаются они или нет в результате обработки изменениям), по типу получаемого результата (предсказательные и описательные, рисунок), по видам применяемого математического аппарата (статистические и кибернетические) и др.

Например, по типу используемого математического аппарата, как правило, выделяют следующие основные группы методов *Data Mining*:

1. Дескриптивный анализ и описание исходных данных, предварительный анализ природы статистических данных (проверка гипотез стационарности, нормальности, независимости, однородности, оценка вида функции распределения, ее параметров и т.п.).
2. Многомерный статистический анализ (линейный и нелинейный дискриминантный анализ, кластерный анализ, компонентный анализ, факторный анализ и т. п.).
3. Поиск связей и закономерностей (линейный и нелинейный регрессионный анализ, корреляционный анализ и т.п.).
4. Анализ временных рядов (динамические модели и прогнозирование).

Детализируя используемый математический аппарат, являющийся важнейшим компонентом практически любых современных методов *Data Mining*, можно получить существенно более глубокую классификацию существующих методов (табл. 2).

Т а б л и ц а 2 Пример классификации методов *Data Mining* по математическому аппарату

Раздел	**Методы, способы**
Метрические методы классификации	Метод ближайших соседей и его обобщения, отбор эталонов и оптимизация метрики
Логические методы классификации	Понятия закономерности и информативности, решающие списки и деревья
Линейные методы классификации	Градиентные методы, метод опорных векторов
Байесовские методы классификации	Оптимальный байесовский классификатор, параметрическое и непараметрическое оценивание плотности, разделение смеси распределений, логистическая регрессия
Методы регрессионного анализа	Многомерная линейная регрессия, нелинейная параметрическая регрессия, непараметрическая регрессия, неквадратичные функции потерь, прогнозирование временных рядов
Нейросетевые методы классификации и регрессии	Многослойные нейронные сети
Композиционные методы классификации и регрессии	Линейные композиции, бустинг, эвристические и стохастические методы, нелинейные алгоритмические композиции
Критерии выбора моделей и методы отбора признаков	Задачи оценивания и выбора моделей, теория обобщающей способности, методы отбора признаков
Ранжирование	
Обучение без учителя	Кластеризация, сети Кохонена, таксономия; поиск ассоциативных правил, задачи с частичным обучением, коллаборативная фильтрация, тематическое моделирование, обучение с подкреплением

1.7. СРАВНИТЕЛЬНЫЕ ХАРАКТЕРИСТИКИ ОСНОВНЫХ МЕТОДОВ

В завершение различных подходов к классификации методов *Data Mining* приведем пример сравнительного анализа наиболее широко используемых методов между собой, применяя в качестве характеристики каждого из атрибутов следующую шкалу оценок: «чрезвычайно низкая, очень низкая, низкая / нейтральная, нейтральная / низкая, нейтральная, нейтральная / высокая, высокая, очень высокая» (табл. 3). Видно, что ни один из методов нельзя признать единственно эффективным, имеющим очевидное превосходство над другими методами.

Т а б л и ц а 3 Пример сравнительного анализа методов *Data Mining*

Метод Характеристика	**Линейная регрессия**	**Нейронные сети**	**Методы визуализации**	**Деревья решений**	**К-ближай-шего соседа**
Точность	Нейтральная	Высокая	Низкая	Низкая	Низкая
Масштабируе-мость	Высокая	Низкая	Очень низ кая	Высокая	Очень низкая
Интерпретиру-емость	Высокая / нейтральная	Низкая	Высокая	Высокая	Высокая / нейтральна я
Пригодность к использованию	Высокая	Низкая	Высокая	Высокая / нейтральная	Нейтральна я
Трудоемкость	Нейтральная	Нейтральная	Очень высокая	Высокая	Низкая / нейтральна я
Разносторонность	Нейтральная	Низкая	Низкая	Высокая	Низкая
Быстрота	Высокая	Очень низкая	Чрезвычайно низкая	Высокая/ней тральная	Высокая
Популярность	Низкая	Низкая	Высокая / нейтральная	Высокая / нейтральная	Низкая

Это подтверждает тезис о том, что залогом успешного решения задач *Data Mining* является необходимость погружения не только в особенности предметной области, но и в математические основы различных методов обработки и анализа данных.

1.8. ПРЕДВАРИТЕЛЬНАЯ ОБРАБОТКА ДАННЫХ

Практическое применение методов *Data Mining* предполагает многоэтапную процедуру. Одним из ключевых этапов этой процедуры, предваряющей, собственно, применение методов *Data Mining*, является этап предварительной обработки данных, включающий различные типы преобразований. Рассмотрим их более подробно.

Одним из ключевых преобразований этапа предварительной обработки данных является *«очистка» Данных* (англ. *Data Cleaning, Data Cleansing, Data Scrubbing),* предполагающая обнаружение и корректировку / удаление поврежденных элементов данных. Данные, имеющие такие повреждения (неточные, неполные, дублированные, противоречивые, зашумленные), называют «*грязными*». Источниками «грязных» данных могут быть поврежденные инструменты сбора

данных, проблемы во введении исходных данных, «человеческий фактор» в случае неавтоматического варианта формирования данных, проблемы в каналах передачи данных, ограничения технологий передачи данных, использование разных наименований в пределах одной номенклатуры и т. п.

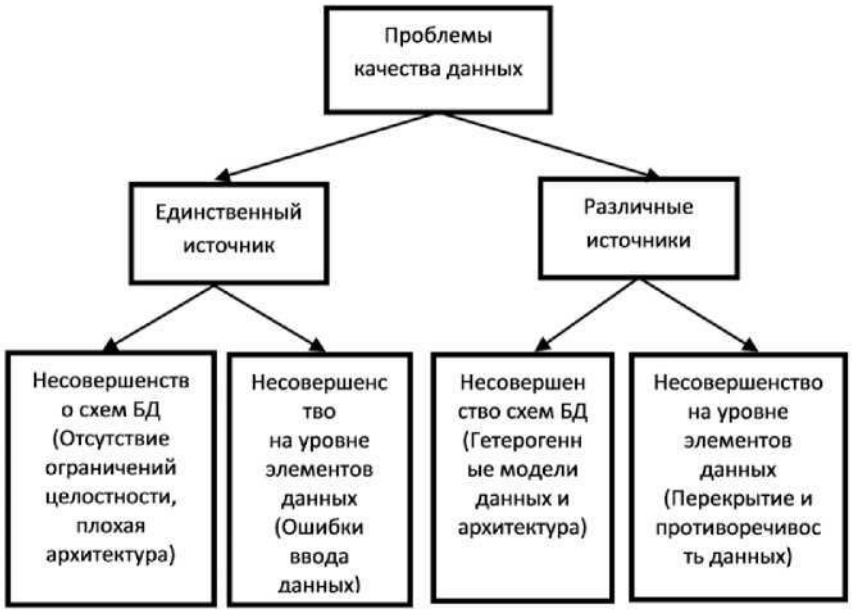
Особую актуальность наличие очистки грязных данных подтверждает известное в информатике выражение - *«Мусор на входе - мусор на выходе»* (англ. *Garbage In - Garbage Out, GIG).* Оно означает, что при неверных входных данных будут получены неверные результаты работы, в принципе, верного алгоритма. Действительно, практически полезными результаты применения каких бы то ни было методов *Data Mining* будут только в случае использования ими корректных достоверных данных. Учитывая то, что такие данные могут быть доставлены из разных источников и быть достаточно существенными в объеме, задача получения и обработки «чистых» данных может быть крайне непростой.

Более того, следует отметить, что наличие «грязных» данных может быть порой более проблематичным, чем их отсутствие вовсе - извлечение полезных знаний из таких данных может потребовать значительного времени, причем безрезультатно. При этом еще более проблематичным будет успешное извлечение из таких данных недостоверных знаний и дальнейшее их практическое использование с трудно предсказуемыми последствиями. Именно поэтому этапу получения «чистых», готовых к анализу данных придают большое значение, а по затратам времени этот этап может быть одним из самых длительных.

Сегодня проблемам получения «чистых» данных посвящены отдельные достаточно емкие исследования. В них обсуждается целый спектр различных особенностей этой проблематики, начиная от концептуальных вопросов и завершая деталями современных технологических решений в базах данных и хранилищах данных. Отметим здесь некоторые наиболее принципиальные моменты.

Все проблемы очистки данных разделяют на две группы, вызванные *интеграцией различных источников данных* (англ. *MultiSource Problems*) или обусловленные проблемами *единственного источника данных* (англ. *Single-Source Problems*). В свою очередь каждая из групп может быть разделена на две другие группы, определяемые либо *несовершенством схем интегрируемых баз данных* (англ. *Schema Level*), либо *несовершенством на уровне собственно элементов данных* (англ. *Instance Level*, записей, объектов и т.п.). Далее каждая из ветвей полученного дерева классификации детализируется конкретным перечнем возможных проблем очистки данных (рисунок).

В табл. 4 и 5 приведены некоторые примеры «грязных» данных, порожденные на разных уровнях - *Schema Level* и *Instant Level.*



Т а б л и ц а 4 Примеры «грязных» данных единственного источника на уровне схемы данных

Проблема		**«Грязные» данные**	**Причины**
Атрибут	Недопустимые значения	дата рождения= 30.13.70	Значение за пределами диапазона
Запись	Нарушение зависимости атрибутов	возраст=22, дата рождения=12.02.70	Возраст = (текущая дата - дата рождения)
Тип записи	Нарушение уникальности	сотр.1=(имя=Иван, SSN=123) сотр.2=(имя=Петр, SSN=123)	SNN должен быть уникальным
Источник	Нарушение ссылочной целостности	сотр.1=(имя=Иван, отд.=789)	Отдела с номером 789 не существует

Т а б л и ц а 5 Примеры «грязных» данных единственного источника на уровне записей

Причина		**«Грязные» данные**	**Причина**
Атрибут	Пропущенное значение	тел.=9999-999999	Недопустимые (некорректные, null и т. п.) значения при вводе
	Орфографические ошибки	город=Тамск город=Москваа	Орфографическая ошибка
	Сокращения и аббревиатуры	должность=А, отдел=ЛТО	
	Объединенные значения	имя=Иван 12.07.70 Томск	Несколько значений в атрибуте
Запись	Нарушение зависимости атрибутов	город=Томск, инд.=666777	Город и индекс не соответствуют друг другу

Тип записи	Дубликаты записей	сотр.1=(имя=Иван, SSN=123) сотр.2=(имя=Иван, SSN=123)	
	Противоречащие записи	сотр.1=(имя=Иван, SSN=123) сотр.1=(имя=Иван, SSN=321)	Записи одного и того же сотрудника с разным SSN
Источник	Неверные ссылки	сотр.=(имя=Иван, отд.=789)	Отдела с номером 789 существует, но указан неверно

Выделяют следующее этапы очистки данных:

1. **Анализ данных** (англ. *Data analysis*). Для того чтобы определить, какие виды ошибок и несоответствий должны быть удалены, требуется детальный анализ данных. В дополнение к инспекции данных или отдельных выборок данных «вручную», следует использовать и метаданные.

2. **Определение способов трансформации потоков данных и правил отображения** (англ. *Definition of transformation workflow and mapping rules*). На данном этапе выполняется оценка количества источников данных, степени их неоднородности и «загрязненности». На основе этой информации создаются схемы потоков данных, позволяющих преобразовать множество источников данных в один, избегая создания ошибок *Multi-Source* слияния (например, появление дублирующих записей).

3. **Верификация** (англ. *Verification*). Оценка корректности и результативности выполнения предыдущего этапа (например, на небольшой выборке данных). При необходимости производиться возврат к этапу 2 для его повторного выполнения.

4. **Трансформация** (англ. *Transformation*). Загрузка данных в единое хранилище с использованием правил трансформации, определенных и отлаженных на этапах 2-3. Очистка данных уровня *Single-Source.*

5. **Обратная загрузка очищенных данных** (англ. *Backflow of cleaned data*). Имея на этапе 4 очищенный набор данных в едином хранилище, целесообразно этими «чистыми» данными заменить аналогичные «грязные» данные в исходных источниках. Это позволит в будущем во многом не выполнять повторно все этапы преобразований по очистке данных.

Реализовать эти этапы можно самыми различными путями с использованием существующих и созданных специально способов и технологий. Рассмотрим наиболее интересные из них.

Этап анализа данных предполагает анализ использования метаданных, которых, как правило, недостаточно для оценки качества данных из имеющихся источников. Поэтому важно анализировать реальные примеры данных, оценивая их характеристики и сигнатуры значений. Это позволяет находить взаимосвязи между атрибутами в схемах данных различных источников. Выделяют два подхода решения этой задачи - *профилирование данных* (англ. *data profiling*) и *извлечение данных* (англ. *data mining*).

Профилирование данных сориентировано на анализ индивидуальных атрибутов, характеризующихся их конкретными свойствами: тип данных, длина, диапазон

значений, частота встречаемости дискретных значений, дисперсия, уникальность, встречаемость «*null*» значений, типичная сигнатура записи (например, у телефонного номера). Именно набор подобных свойств (профиль) позволяет оценить различные аспекты качества данных.

Извлечение Данных предполагает поиск взаимосвязей между несколькими атрибутами достаточно большого набора данных. Учитывая то, что этот способ получил название *data mining*, здесь используют упоминавшиеся выше (см. табл. 1) методы *кластеризации*, *подведения итогов*, *поиска ассоциаций* и *последовательностей*. Кроме того, для дополнения пропущенных значений, корректировки недопустимых значений или идентификации дубликатов могут быть использованы существующие ограничения целостности (англ. *integrity constraints*), принятые в реляционных базах данных, наложенные дополнительно на бизнес-связи между атрибутами. Например, известно, что *«Total = Quantity*Unit_Price».* Все записи, не удовлетворяющие этому условию, должны быть изучены более внимательно, исправлены или исключены из рассмотрения.

Для разрешения проблем очистки данных в одном источнике (*single-source problems*), в том числе перед его интеграцией с другими источниками данных, реализуют следующие этапы:

- **Извлечение значений из атрибутов свободной формы** (разбиение атрибутов, англ. *Extracting values from free-form attributes (attribute split)).* В данном случае речь может идти о строковых значениях, сохраняющих несколько слов подряд (например, адрес или полное имя человека). В данном случае требуется четкое понимание того, на какой позиции этого значения находится интересующая нас часть атрибута. Возможно, потребуется даже сортировка составных частей такого атрибута.
- **Валидация и коррекция** (англ. *Validation and correction).* Данный этап предполагает поиск ошибок ввода данных и их исправление наиболее автоматическим способом. Например, используя автоматическую проверку правописания во избежание орфографических ошибок и опечаток. Словарь географических названий и почтовых кодов также следует использовать для корректировки значений вводимых адресов. Зависимость атрибутов (дата рождения - возраст, *Total = Quantity* Unit Price* и т.п.) также способствует избеганию множества ошибок в данных.
- **Стандартизация** (англ. *Standardization).* Этот этап предполагает приведение всех данных к единому универсальному формату. Примерами таких форматов являются формат написания даты и времени, размер регистра в написании строковых значений. Текстовые поля должны исключать префиксы и суффиксы, аббревиатуры в них должны быть унифицированы, исключены проблемы с различной кодировкой.

Одной из основных проблем, вызванных интеграцией различных источников (*multi-source problems*) данных, является устранение дублирования записей. Этот этап выполняется после подавляющего большинства преобразований и чисток. Он предполагает сначала идентификацию сходных в некотором смысле записей, а

затем их слияние с объединением атрибутов. Очевидно, решение этой задачи при наличии у дублирующих записей первичного ключа достаточно просто. Если такого однозначно идентифицирующего признака нет, то задача устранения дубликатов значительно осложняется, требуя применения нечетких (англ. *fuzzy*) подходов сравнения (близости в некотором смысле) записей между собой.

1.9. ПРОЦЕСС ОБНАРУЖЕНИЯ ЗНАНИЙ

Основные этапы анализа

Для обнаружения знаний в данных недостаточно просто применить методы Data Mining, хотя, безусловно, этот этап является основным в процессе интеллектуального анализа. Весь процесс состоит из нескольких этапов. Рассмотрим основные из них, чтобы продемонстрировать, что без специальной подготовки аналитика методы Data Mining сами по себе не решают существующих проблем. Итак, весь процесс можно разбить на следующие этапы:

- понимание и формулировка задачи анализа;
- подготовка данных для автоматизированного анализа (препроцессинг);
- применение методов Data Mining и построение моделей;
- проверка построенных моделей;
- интерпретация моделей человеком.

На первом этапе выполняется осмысление поставленной задачи и уточнение целей, которые должны быть достигнуты методами Data Mining, важно правильно сформулировать цели и выбрать необходимые для их достижения методы, г. к. от этого зависит дальнейшая эффективность всего процесса.

Второй этап состоит в приведении данных к форме, пригодной для применения конкретных методов Data Mining. Данный процесс далее будет описан более

подробно, здесь заметим только, что вид преобразований, совершаемых над данными, во многом зависит от используемых методов, выбранных на предыдущем этапе.

Третий этап— это собственно применение методов Data Mining. Сценарии этого применения могут быть самыми различными и могут включать сложную комбинацию разных методов, особенно если используемые методы позволяют проанализировать данные с разных точек зрения.

Следующий этап —- проверка построенных моделей. Очень простой и часто используемый способ заключается в том, что все имеющиеся данные, которые необходимо анализировать, разбиваются на две группы. Как правило, одна из них большего размера, другая меньшего. На большей группе, применяя тс или иные методы Data Mining, получают модели, а на меньшей — проверяют их. Но разнице в точности между тестовой и обучающей группами можно судить об адекватности построенной модели.

Последний этап — интерпретация полученных моделей человеком в целях их использования для принятия решений, добавление получившихся правил и зависимостей в базы знаний и т. д. Этот этап часто подразумевает использование методов, находящихся на стыке технологии Data Mining и технологии экспертных систем. Ог того, насколько эффективным он будет, в значительной степени зависит успех решения поставленной задачи.

Этим этапом завершается цикл Data Mining. Окончательная оценка ценности добытого нового знания выходит за рамки анализа, автоматизированного или традиционного, и может быть проведена только после претворения в жизнь решения, принятого на основе добытого знания, после проверки нового знания практикой. Исследование достигнутых практических результатов завершает оценку ценности добытого средствами Data Mining нового знания.

Подготовка исходных данных

Как уже отмечалось ранее, для применения того или иного метода Data Mining к данным их необходимо подготовить к этому. Например, поставлена задача построить фильтр электронной почты, не пропускающий спам. Письма представляют собой тексты в электронном виде. Практически ни один из существующих методов Data Mining не может работать непосредственно с текстами. Чтобы работать с ними, необходимо из исходной текстовой информации предварительно получить некие производные параметры, например: частоту встречаемости ключевых слов, среднюю длину предложений, параметры, характеризующие сочетаемость тех или иных слов в предложении, и т. д. Другими словами, необходимо выработать некий четкий набор числовых или нечисловых параметров, характеризующих письмо. Эта задача наименее автоматизирована в том смысле, что выбор системы данных параметров производится человеком, хотя, конечно, их значения могут вычисляться автоматически. После выбора

описывающих параметров изучаемые данные могут быть представлены в виде прямоугольной таблицы, где каждая строка представляет собой отдельный случай, объект или состояние изучаемого объекта, а каждая колонка — параметры, свойства или признаки всех исследуемых объектов. Большинство методов Data Mining работают только с подобными прямоугольными таблицами.

Полученная прямоугольная таблица пока еще является слишком сырым материалом для применения методов Data Mining, и входящие в нее данные необходимо предварительно обработать. Во-первых, таблица может содержать параметры, имеющие одинаковые значения для всей колонки. Если бы исследуемые объекты характеризовались только такими признаками, они были бы абсолютно идентичны, значит, эти признаки никак не индивидуализируют исследуемые объекты. Следовательно, их надо исключить из анализа. Во- вторых, таблица может содержать некоторый категориальный признак, значения которого во всех записях различны. Ясно, что мы никак не можем использовать это поле для анализа данных и его надо исключить. Наконец, просто этих полей может быть очень много, и если все их включить в исследование, то это существенно увеличит время вычислений, поскольку практически для всех методов Data Mining характерна сильная зависимость времени от количества параметров (квадратичная, а нередко и экспоненциальная). В то же время зависимость времени от количества исследуемых объектов линейна или близка к линейной. Поэтому в качестве предварительной обработки данных необходимо, во-первых, выделить то множество признаков, которые наиболее важны в контексте данного исследования, отбросить явно неприменимые из-за константности или чрезмерной вариабельности и выделить те, которые наиболее вероятно войдут в искомую зависимость. Для этого, как правило, используются статистические методы, основанные на применении корреляционного анализа, линейных регрессий ит. д. Такие методы позволяют быстро, хотя и приближенно оценить влияние одного параметр; н\ другой.

Мы обсудили очистку данных по столбцам таблицы (признакам). Точно так же бывает необходимо провести предварительную очистку данных по строкам таблицы (записям). Любая реальная база данных обычно содержит ошибки, очень приблизительно определенные значения, записи, соответствующие каким-то редким, исключительным ситуациям, и другие дефекты, которые могут резко понизить эффективность методов Data Mining, применяемых на следующих этапах анализа. Такие записи необходимо отбросить. Даже если подобные "выбросы[4] не являются ошибками, а представляют собой редкие исключительные ситуации, они все равно вряд ли могут быть использованы, поскольку по нескольким точкам статистически значимо судить об искомой зависимости невозможно. Эта предварительная обработка или препроцессинг данных и составляет второй этап процесса обнаружения знаний.