Желание извлечь полезную информацию и знания из обширного объема данных стали главным мотивом, который привлек внимание исследователей в последнее время к добыче данных. Извлеченная информация и знания могут быть важными для приложений в диапазоне от управления малого бизнеса до сложного проектирования инженерных систем и научных исследований. Добыча данных (так же используются термины «Data mining» и «Интеллектуальный анализ данных») - это анализ и исследование огромных наборов данных с целью обнаружения значимых закономерностей и правил, которые ранее были неизвестны. Основная цель - использование обработки данных на компьютере в сочетании с человеческой способностью воспринимать закономерности (Хан и Кэмбер, 2001). Эпоха приложений добычи данных была заложена в 1980 году, главным образом инструментами, разработанными исследователями для выполнения отдельных задач (Пятецкий-Шапиро, 2000). В последнее время добыча данных стала доминирующей среди статистиков, специалистов по управлению информацией и аналитиков данных. На первом семинаре по KDD (**Knowledge** **Discovery** in **Databases**) в 1989 году Пятецкий-Шапиро придумал фразу «открытие знаний в базах данных». Популярность добычи данных и KDD не должна удивлять, учитывая масштаб данных, собираемых из различных источников. Собранные данные настолько огромны, что их нельзя рассматривать вручную, а многие автоматические анализы данных, поддерживаемые классической статистикой и машинным обучением, могут столкнуться с проблемами при работе с объемными и сложными наборами данных. В связи с этим сборища данных становятся "хранилищами данных", которые не часто посещаются.

В настоящее время методы и инструменты data mining активно используются во многих областях, включая бизнес-аналитику, финансы, маркетинг, медицину, геологию и другие. Data mining позволяет извлекать ценную информацию и знания из больших объемов данных, которые могут быть использованы для принятия эффективных решений и улучшения качества жизни людей. Однако, следует учитывать возможные проблемы в сфере конфиденциальности и этики, связанные с использованием личных данных людей в data mining процессе. В целом, data mining продолжает эволюционировать и представляет огромный потенциал для будущих исследований и разработок.

**ОБЗОР ЛИТЕРАТУРЫ**

В статье «От интеллектуального анализа данных к обнаружению знаний в базах данных» KDD было описано как «нетривиальный процесс распознавания действительных, новых, потенциально полезных и, наконец, понятных закономерностей в данных». Данные для уточнения определения представляли собой любой набор достоверных фактов, доступных в электронной форме. Шаблоны — это модели, выраженные на каком-либо языке в виде подмножества данных. Шаблоны должны быть действительными, чтобы они могли быть смоделированы для любых новых данных. Процесс включает в себя несколько шагов от подготовки данных до расширения знаний, которые периодически используются до тех пор, пока не будут достигнуты искомые результаты. Нетривиальность процесса указывает на то, что должно быть своего рода вычисление логического вывода, чтобы отличить его от традиционного вычисления значений. В статье KDD описана как «обобщенная процедура извлечения ценных знаний из данных, при этом майнинг является одним из других шагов в этом процессе, в котором используются некоторые алгоритмы для процесса извлечения знаний». В 1998 был предложен интеллектуальный анализ данных в качестве эффективного инструмента прямого маркетинга, чтобы улучшить маркетинг продукта в этот технологический век, когда традиционные средства маркетинга, такие как массовый маркетинг, демонстрируют тенденцию к упадку. Используя интеллектуальный анализ данных, мы можем определить модели покупателей, чтобы выделить потенциальных покупателей из списка клиентов. Было продемонстрировано, что интеллектуальный анализ данных как инструмент прямого маркетинга может принести больше прибыли, чем традиционные средства массового маркетинга, поскольку он ориентирован только на потенциальных покупателей. Майкл Гобель в своей статье «Обзор инструментов интеллектуального анализа данных и поиска знаний» представили обобщенный взгляд на общие задачи поиска знаний и различные методологии для их решения. Была предложена схема классификации признаков, которая использовалась для изучения программного обеспечения для интеллектуального анализа данных и знаний. Гобель указал некоторые из важных функций, которые следует считать важными для программного обеспечения для обнаружения знаний, чтобы его можно было эффективно использовать и решать больше вопросов, которые недостаточно изучены. Многие организации в мире сегодня имеют очень огромные базы данных, которые не имеют ограничений на рост. Новые данные добавляются в эти базы данных со скоростью миллионов записей в день. Эти типы баз данных предоставляют новую задачу и уникальные возможности для анализа этих потоков данных. Дэвид Хэнд (2001) определил интеллектуальный анализ данных как «анализ огромных наборов данных для обнаружения неожиданных взаимосвязей и более логического анализа данных, чтобы они служили желаемым результатам». Согласно Ригельски (2002), технология интеллектуального анализа данных добавила CRM новое измерение. Способность интеллектуального анализа данных извлекать прогностическую неизвестную информацию из обширных наборов данных нашла свое применение в CRM для выявления и оценки ценных клиентов, прогнозирования покупательского поведения клиентов, что помогает поставщикам принимать упреждающие и основанные на знаниях решения. Имонн Кеох (2004) рассматривал «интеллектуальный анализ данных без параметров», поскольку алгоритмы с нагруженными параметрами могут переоценивать или недооценивать определенные параметры, что приводит к шаблонам, которые могут быть не совсем точными. Майнинг без параметров может помешать нам применять наши собственные предположения или предубеждения. Они предложили парадигму сбора данных, основанную на сжатии. Потоковый интеллектуальный анализ данных считается сложной задачей по обнаружению знаний в базах данных, традиционные подходы к интеллектуальному анализу неприменимы для его решения, поскольку данные поступают в виде нескольких непрерывных и колеблющихся во времени потоков данных. Алхаммади и др. (2007) представили необычную методологию анализа эволюционирующих закономерностей в потоковых данных, которая, как доказано экспериментами, имеет лучшую сложность анализа и точность классификации. Поскольку большинство современных методологий интеллектуального анализа данных исследуют знания в одной таблице данных. Но в последнее время большинство этих методологий растянуты на реляционные случаи. Интеллектуальный анализ реляционных данных включает в себя применение подхода интеллектуального анализа данных к данным из нескольких таблиц для абстрагирования содержащихся в них знаний. Венкадатри M. и др. (2011), обсуждают соответствующие методы и методологии, которые потребуются в будущем для удовлетворения потребностей области интеллектуального анализа данных, поскольку она исследует все более и более сложные области, чтобы мы могли исследовать такие сложные ситуации, когда данные огромен, но полон скрытой информации. Дивья Томар (2013) представил интеллектуальный анализ данных как наиболее яркую и привлекательную область исследований, которая набирает популярность в медицинской сфере. Интеллектуальный анализ данных дает несколько преимуществ в области здравоохранения, в том числе делает медицинское обслуживание в экономичнее. Ананд В., Сауркар и др. (2014) определили интеллектуальный анализ данных как «междисциплинарную область, которая состоит из интегрированных баз данных, искусственного интеллекта, машинного обучения, статистики и т. д.». Они определили интеллектуальный анализ данных как многоэтапный процесс, который включает подготовку данных для майнинга, алгоритмы майнинга, анализ результатов и интерпретацию результатов. Способность интеллектуального анализа данных глубоко проникать в данные и извлекать из них скрытую информацию и знания привлекла огромное внимание со стороны бизнес-профессионалов для создания моделей, связанных с поведением клиентов, и прогнозирования будущих продаж и тенденций, а также для помощи лицам, определяющим политику, в принятии решений с помощью целью увеличения прибыли (Шрадха Сони 2015).

**АРХИТЕКТУРА DATA MINING**

Архитектура интеллектуального анализа состоит из следующих частей:

**База знаний:** служит инициатором всего процесса интеллектуального анализа данных. Она действует как руководство для поиска или оценки интереса результирующих шаблонов. Такой тип знаний может включать в себя иерархии понятий, которые организуют атрибуты или их значения в отдельные этапы абстракции.

**Механизм интеллектуального анализа данных:** является основным компонентом системы интеллектуального анализа данных и состоит из всех необходимых модулей, таких как определение характеристик, прогнозирование, кластерный анализ, анализ выбросов и анализ эволюции для выполнения задач интеллектуального анализа данных.

**Модуль оценки шаблона:** этот модуль обычно связан с мерами интереса. Он постоянно взаимодействует с механизмом интеллектуального анализа данных, чтобы оставаться сосредоточенным на поиске интересных закономерностей. Во многих случаях он использует пороговые значения для отсеивания обнаруженных шаблонов или может использовать модуль оценки шаблонов, интегрированный с модулем интеллектуального анализа данных, в зависимости от используемого метода интеллектуального анализа данных.

**Пользовательский интерфейс:** модуль действует как соединение между пользователями и системой интеллектуального анализа данных. Это облегчает взаимодействие пользователей с системой простым и эффективным способом, не беспокоя пользователя о свертках, стоящих за процессом.

**Источники данных** (хранилище данных, база данных, другие репозитории): это фактические источники данных, и для успешного анализа данных требуется огромный объем исторических данных. Организации обычно хранят данные в базах данных или хранилищах данных. Иногда в хранилище данных содержится более одной базы данных, текстовых файлов или электронных таблиц.

**Сервер базы данных или хранилища данных:** он содержит материальные данные, которые должны быть извлечены. Получение данных по запросу пользователей является его ключевой обязанностью.

**Другие процессы:** прежде чем данные будут переданы на сервер хранилища данных, данные необходимо очистить и интегрировать, поскольку данные собираются из разных источников и имеют разные форматы, поэтому их нельзя использовать непосредственно для процесса майнинга. Данные должны быть очищены, интегрированы, и только надежные данные должны быть выбраны и переданы на сервер хранилища данных. Процесс может потребовать ряда методов очистки, интеграции и отбора.

**ПРОЦЕСС ДОБЫЧИ ДАННЫХ.**

Файяд и др. (1996) определили «интеллектуальный анализ данных как один из нескольких шагов в процессе обнаружения знаний, он включает в себя применение анализа данных и поиск алгоритмов, которые дают точное перечисление закономерностей в данных при любой приемлемой эффективности вычислений». Эта процедура является совместной и повторяющейся и включает в себя множество шагов с решениями, принимаемыми пользователем, с попытками, предпринимаемыми на каждом этапе для выполнения конкретной задачи обнаружения, каждая из которых выполняется путем применения метода обнаружения. Интеллектуальный анализ данных, который некоторые синонимично используют для процесса обнаружения фразовых знаний в базах данных (KDD), наоборот, многие считают его важным шагом KDD, который приводит к полезным шаблонам или моделям для данных. Рассмотрим процессы интеллектуального анализа данных.

**Отбор:** выбор подходящих данных из различных источников для процесса майнинга.

**Предварительная обработка:** поскольку данные собираются из разных источников, они содержат несоответствия, для устранения которых на этом этапе выполняются различные действия, исправляются или удаляются ошибочные данные, удаляются шумы и несоответствия, а данные из разных источников объединяются.

**Преобразование:** здесь данные преобразуются в подходящую форму для майнинга. Могут использоваться выбор признаков, выборка, агрегирование.

**Интеллектуальный анализ данных:** это важный шаг, когда выбирается алгоритм интеллектуального анализа данных, который соответствует образцу данных. Также осуществляется извлечение шаблонов данных.

**Интерпретация и оценка:** Распознавание и преобразование результатов или закономерностей добычи в знания путем устранения избыточности и нерелевантных закономерностей. Здесь используется набор визуализаций и приемов графического интерфейса для преобразования выгодных шаблонов в понятные человеку термины.

**ЗАДАЧА ДОБЫЧИ ДАННЫХ**

Задачи интеллектуального анализа данных сгруппированы в две основные категории: 1: Предиктивные(предсказательные) 2: Описательные. Эти две задачи считаются основными задачами интеллектуального анализа данных. Файяд et.al. 1996 определяют шесть основных функций интеллектуального анализа данных: 1. Классификация 2. Регрессия 3. Кластеризация 4. Моделирование зависимостей 5. Обнаружение отклонений. 6.Подведение итогов.

Классификация, регрессия и обнаружение аномалий относятся к категории прогнозирования, в то время как кластеризация и моделирование зависимостей относятся к описательной категории. Прогностическая модель прогнозирует с использованием некоторой переменной в наборе данных, чтобы предсказать неизвестные значения другой соответствующей переменной, в то время как описательная модель классифицирует закономерности или отношения и охватывает понятные человеку закономерности и тенденции в данных (Горунеску Флорин, 2011 г.)

**Классификация:** классификация относится к классическим методам интеллектуального анализа данных, основанным на машинном обучении. Он находит взаимные свойства среди набора объектов в базе данных и распределяет их по разным классам в соответствии с моделью классификации. Его основная цель — тщательно изучить обучающие данные и разработать точное описание или модель для каждого класса, используя функции, доступные в данных. В этом методе используются математические методы, такие как деревья решений, нейронные сети и статистика (Минг-Сян и др., 1996).

**Регрессия:** это один из методов интеллектуального анализа данных, который определяет связь между зависимыми и независимыми переменными. Прогнозирование выполняется с поддержкой регрессии. Статистическая регрессия — это математическая модель, которая устанавливает связь между значениями зависимой переменной и значениями другого предиктора или независимой переменной. В регрессии прогнозируемая переменная может быть непрерывной переменной. В регрессии действительные переменные предсказания сопоставляются с элементами функции обучения. Статистическая регрессия, нейронная сеть, регрессия машины опорных векторов являются одними из наиболее часто используемых стратегий регрессии. Более сложные методы, такие как логистическая регрессия, деревья решений или нейронные сети, также могут использоваться для прогнозирования будущих значений, эти методы также могут быть объединены для достижения лучшего результата.

**Кластеризация:** это метод интеллектуального анализа данных, который группирует физические или абстрактные объекты в классы подобных объектов. кластеризация — это метод разделения набора данных (записей/кортежей/объектов/выборок) на несколько групп (кластеров) на основе предопределенных сходств. Основная цель кластеризации - найти группы (кластеры) объектов на основе сходства, чтобы внутри отдельного кластера было большое сходство друг с другом, в то время как кластеры достаточно отличались друг от друга. В терминологии машинного обучения кластеризация — это форма обучения без учителя.

**Моделирование зависимостей** (интеллектуальный анализ правил ассоциации): это один из лучших признанных методов интеллектуального анализа данных, который относится к категории неконтролируемого метода интеллектуального анализа данных, который направлен на поиск связей или отношений между элементами или записями, принадлежащими большому набору данных, и помечает существенные зависимости между переменными. Интеллектуальный анализ ассоциативных правил представляет собой импликацию формы X → Y, где x и y — различные элементы или наборы элементов, производящие операторы «если-то» относительно значений атрибутов. В анализе потребительской корзины это правило обычно используется, оно пытается проанализировать клиентов, покупающих определенные товары, и дает представление о комбинациях, которые клиенты часто покупают вместе.

**Обнаружение аномалий:** синоним его названия, он занимается обнаружением наиболее существенных изменений или отклонений от стандартного поведения.

**Подведение итогов**: хотя и не входит в число методов интеллектуального анализа данных, но является результатом этих методов и имеет дело с определением компактного описания для подмножества данных, синонимично именуемого обобщением или описанием.

**ПРОБЛЕМЫ В DATA MINING**

Интеллектуальный анализ данных хорошо развит, но при практической реализации он по-прежнему сталкивается с множеством проблем, некоторые из которых обсуждаются ниже:

**Вопросы безопасности:** безопасность является наиболее важным и жизненно важным вопросом, касающимся любого процесса передачи данных, учитывая чрезвычайно конфиденциальный характер данных, следует предотвращать потенциальный незаконный доступ к знаниям и обеспечивать конфиденциальность.

**Вопросы методологии интеллектуального анализа данных:** поскольку разные пользователи интересуются различными видами знаний, интеллектуальный анализ данных должен охватывать широкий спектр задач анализа данных и обнаружения знаний, которые могут использовать одну и ту же базу данных по-разному и требуют разработки многочисленных методов интеллектуального анализа данных.

**Проблемы с пользовательским интерфейсом:** обнаружение знаний с помощью инструментов интеллектуального анализа данных является выгодным и выразительным только в том случае, если оно представлено пользователю явно и увлекательно. Поскольку трудно понять, что можно обнаружить в базе данных, процесс майнинга должен быть интерактивным, данные должны быть представлены на языке высокого уровня, визуальными представлениями или другими формами графического выражения, чтобы пользователь мог понять и интерпретировать их и использовать по мере необходимости.

**Обработка зашумленных и неполных данных:** данные, хранящиеся в базах данных, могут различаться, поскольку с источниками данных связаны различные проблемы, данные могут быть неполными, данные могут содержать случаи, которые могут вызывать исключения. Интеллектуальный анализ данных с такими несоответствиями вызывает неоднозначность в процессе, в результате чего построенная модель знаний переопределяет данные и снижает точность полученных знаний, поэтому требуются методы интеллектуального анализа, которые могут справиться с этими несоответствиями.

**Проблемы с производительностью:** Эффективность и масштабируемость интеллектуального анализа данных являются стратегическими факторами для интеллектуального анализа данных, реализующего систему баз данных. Информация должна быть эффективно и умело извлечена из баз данных, поскольку они огромны по объему. Используемые алгоритмы должны быть эффективными и масштабируемыми, время их работы должно быть предсказуемым и приемлемым для больших баз данных.

**ПРИМЕНЕНИЕ DATA MINING**

**Применение в здравоохранении:** интеллектуальный анализ данных может иметь значительные преимущества в системе здравоохранения, но его успех зависит от наличия чистых данных. В здравоохранении используется для диагностики и прогнозирования заболеваний, а также может быть установлена принадлежность заболеваний друг к другу. Врачи могут определить эффективные и передовые методы, чтобы пациент получал более качественные и разумные услуги. Поскольку огромные объемы медицинских данных сложны и обширны для обработки и анализа, интеллектуальный анализ данных предоставляет методологию и инструменты для преобразования данных в информацию для эффективного принятия решений.

**Применение в образовательных системах:** интеллектуальный анализ данных в образовательной системе является развивающейся областью, и исследователи проявляют к ней большой интерес. Поскольку миллионы студентов ежегодно зачисляются в различные учебные заведения, это добавляет огромный объем данных. Методы интеллектуального анализа данных могут помочь заполнить пробелы в знаниях в образовательной системе, выявив завуалированные закономерности, коннотации и различия. Это помогает заинтересованным сторонам повысить эффективность принятия решений, что приводит к совершенствованию системы образования.

**Применение в CRM:** интеллектуальный анализ данных в CRM (Customer Relationship Management) в настоящее время является наиболее обсуждаемой темой исследований в промышленности и академических кругах с целью предоставления сводки исследований по использованию методов интеллектуального анализа данных в области CRM.

**Применение в анализе рыночной корзины (MBA):** для анализа рыночной корзины используются различные методы интеллектуального анализа данных. наблюдает за покупательскими привычками клиентов. Коммерческие дома могут использовать методы интеллектуального анализа данных для выявления моделей покупок и поведения клиентов, на основе которых покупателю может быть представлен ряд вариантов в соответствии с его привычкой покупать.

**Применение в спортивных данных:** Методы интеллектуального анализа данных также проникли в область спорта. В каждом виде спорта проводится огромное количество игр, генерирующих огромное количество статистических данных. Эти огромные данные необходимо поддерживать в отношении планирования событий и статистики отдельных игроков в этих событиях. Интеллектуальный анализ данных можно использовать для прогнозирования и анализа производительности, а также для планирования стратегии.