**Обзор литературы по data mining**

**Аннотация**

С развитием технологий, особенно за последние три десятилетия, огромное количество информации было переведено в цифровую форму, что привело к формированию огромных хранилищ данных. По мере накопления информации в этих хранилищах сохранялась проблема извлечения из нее значимых знаний. Для решения этой ситуации был использован интеллектуальный анализ данных в качестве инструмента. Интеллектуальный анализ данных, рассматриваемый как ступень к процедуре обнаружения знаний в базах данных, представляет собой процедуру извлечения скрытой информации из огромных наборов баз данных для выявления шаблонов и правил. Интеллектуальный анализ данных (data mining) в настоящее время стал незаменимым компонентом практически во всех сферах человеческой жизни. В данном параграфе представлен анализ доступной литературы по интеллектуальному анализу данных. Кратко излагается концепция интеллектуального анализа данных, а также его различные методологии. Также были проиллюстрированы некоторые приложения, задачи и проблемы, связанные с ИТ.

**Введение**

Наличие большого объема данных практически в каждой области и желание извлекать из них полезную информацию и знания были главной мотивацией, которая в недавнем прошлом привлекала внимание исследователей к интеллектуальному анализу данных. Извлеченные информация и знания могут оказаться чрезвычайно полезными для различных приложений, начиная от управления малым бизнесом и заканчивая сложным инженерным проектированием и научными исследованиями. Интеллектуальный анализ данных - это анализ и тщательное изучение огромных массивов данных с целью выявления важных закономерностей и правил, которые ранее были неизвестны. Основная цель - использовать возможности компьютера по обработке данных в сочетании со способностью человека воспринимать шаблоны (Хан и Кэмбер 2001). Эпоха приложений интеллектуального анализа данных началась в 1980 году преимущественно исследовательскими инструментами, ориентированными на выполнение рутинной работы в одиночку (Пятецкий-Шапиро 2000). В последнее время интеллектуальный анализ данных становится доминирующим среди статистиков и аналитиков данных. В 1989 году Пятецкий-Шапиро ввел в обиход KDD (Knowledge Discovery in Database / Поиск Знаний в Базах Данных). Признание интеллектуального анализа данных и KDD не должно вызывать удивления, учитывая масштаб данных, собранных из различных доступных источников, собранные данные проверяются вручную, и часто автоматический анализ данных, поддерживаемый классической статистикой и машинным обучением, может вызвать проблемы, если процедура сложная, а собранные знания включают в себя проблемные сущности. Огромный объем данных, собранных из многочисленных источников и хранящихся в обширных и разнообразных хранилищах. Таким образом, сбор данных превышает возможности человека для анализа без мощного инструмента анализа, в результате чего эти хранилища становятся "хранилищами данных", которые посещаются нечасто. Поскольку лицам, принимающим решения, не хватает инструментов для извлечения ценных знаний, содержащихся в огромном объеме данных, следовательно, при принятии жизненно важных решений не используются богатые информацией данные. Инструменты интеллектуального анализа данных выполняют анализ данных и определяют важные закономерности, которые ранее оставались не замечены. Поскольку каждая сфера человеческой жизни стала насыщенной данными, это привело к тому, что интеллектуальный анализ данных стал незаменимым компонентом. KDD можно рассматривать как всеобъемлющую процедуру извлечения полезных знаний из данных, в то время как интеллектуальный анализ данных можно рассматривать как ядро KDD, которое включает алгоритмы, которые исследуют данные, строят модели и обнаруживают неизвестные закономерности.

**Обзор литературы**

Файяд и др. в своей статье “От интеллектуального анализа данных к обнаружению знаний в базах данных” описали KDD как “нетривиальный\* процесс распознавания действительных, новых, потенциально полезных и, наконец, понятных шаблонов в данных”. При разработке определения данными были любые достоверные факты, доступные в электронной форме. Шаблоны - это модели, выражаемые некоторым языком в виде подмножества данных. Шаблоны должны быть допустимыми, чтобы они были истинными и могли быть смоделированы для любых новых данных. Процесс включает в себя множество этапов от подготовки данных до расширения знаний, которые используются повторно до тех пор, пока не будут достигнуты желаемые результаты. Нетривиальный\* указывает на то, что должно существовать своего рода вычисление вывода, чтобы отличать его от традиционного вычисления значений. Файяд и Столорц (1997) в своей статье описали KDD как “обобщенную процедуру извлечения ценных знаний из данных, при этом интеллектуальный анализ является одним из других этапов этого процесса, который использует некоторые алгоритмы для процесса извлечения знаний”. Чарльз и др. (1998) предложили интеллектуальный анализ данных в качестве эффективного инструмента прямого маркетинга, чтобы улучшить маркетинг продукции в век, когда традиционные средства маркетинга, такие как массовый маркетинг, демонстрируют тенденцию к упадку. Используя интеллектуальный анализ данных, можно определить модели поведения покупателей, чтобы выделить потенциальных покупателей из списка клиентов. Было продемонстрировано, что интеллектуальный анализ данных как инструмент прямого маркетинга может приносить больше прибыли, чем традиционные средства массового маркетинга, поскольку он ориентирован только на потенциальных покупателей. Майкл Гебель и др. (1999) в своей статье “Обзор инструментов интеллектуального анализа данных и обнаружения знаний” представили обобщенное представление об общих задачах обнаружения знаний и различных методологиях их решения. Была предложена схема классификации признаков, которая использовалась для изучения знаний и программного обеспечения для интеллектуального анализа данных. Они указали некоторые важные функции, которые следует считать важными для программного обеспечения для обнаружения знаний, чтобы оно могло эффективно использоваться и решало больше проблем, которые были недостаточно изучены. Многие организации в современном мире располагают огромными базами данных, которые не имеют никаких ограничений на рост. Новые данные добавляются в эти БД со скоростью миллионов записей в день. Эти типы баз данных создают новую проблему и уникальные возможности для обработки этих потоков данных. Отличая data mining от информации Хенд и др. (2001) определили интеллектуальный анализ данных как “анализ огромных наборов данных с целью обнаружения неожиданных взаимосвязей и более логичного анализа данных, чтобы это привело к желаемым результатам”. По словам Ригельского и др. (2002), технология data mining добавила новое измерение в CRM. Возможности интеллектуального анализа данных по извлечению прогнозируемой неизвестной информации из обширных наборов данных нашли свое применение в CRM для выявления и оценки ценных клиентов, прогнозирования покупательского поведения клиентов, что помогает поставщикам принимать упреждающие решения, основанные на знаниях. Имон Киф и др. (2004), обсуждался “интеллектуальный анализ данных без параметров”, поскольку алгоритмы, нагруженные параметрами, могут завышать или занижать оценку определенных параметров, что приведет к шаблонам, которые могут быть не совсем точными. Майнинг без параметров может помешать нам применять наши собственные предположения или предубеждения. Они предложили парадигму обработки данных, основанную на сжатии. Потоковый анализ данных считается сложной задачей при обнаружении знаний в базах данных, традиционные подходы интеллектуального анализа данных неосуществимы для решения этой задачи, поскольку данные поступают в виде множества, непрерывных и изменяющихся во времени потоков данных. Альхаммади и др. (2007) представили необычную методологию интеллектуального анализа эволюционирующих шаблонов потоковых данных, которая обладает большей сложностью интеллектуального анализа и точностью классификации, что было доказано экспериментально, поскольку большинство современных методологий интеллектуального анализа данных исследуют знания в единой таблице данных. Но в последнее время большинство из этих методологий применяются к реляционным случаям. Реляционный data mining предполагает применение подхода интеллектуального анализа данных к данным из нескольких таблиц для абстрагирования содержащихся в них знаний (Сасо Дзероски 2010). Венкатадри и др. (2011), обсуждают, какие соответствующие методы и методологии необходимы в будущем для удовлетворения потребностей области интеллектуального анализа данных, поскольку она исследует все более сложные области, чтобы мы могли исследовать такие сложные ситуации, когда данные огромны и полны скрытой информации. Типаван Сильваттананусарн и др. (2012) в своей работе исследовали применение методов интеллектуального анализа данных для поддержки процесса управления знаниями. Ими было показано, что интеллектуальный анализ данных может быть интегрирован и внедрен в систему управления знаниями. Дивья Томар и др. (2013) представили интеллектуальный анализ данных как наиболее динамичную и привлекательную область исследований, которая набирает популярность в медицинской сфере. Интеллектуальный анализ данных дает ряд преимуществ в области здравоохранения. Это повышает эффективность медицинских услуг с точки зрения затрат. Ананд 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 многоступенчатый процесс, который включает подготовку данных для интеллектуального анализа, алгоритмы интеллектуального анализа данных, анализ и интерпретацию результатов. Способность интеллектуального анализа данных глубоко копаться в данных и извлекать из них скрытую информацию и знания привлекла огромное внимание бизнес-профессионалов, чтобы генерировать модели, связанные с поведением клиентов, и прогнозировать будущие продажи и тенденции, а также помогать политикам в принятии решений с целью увеличения прибыли (Шраддха Сони, 2015).

**Архитектура data mining**

Архитектура data mining представлена на схеме ниже.

База знаний

Пользовательский интерфейс

Модуль оценки шаблонов

Сервер БД / Хранилища данных

Прочие хранилища

Всемирная паутина

Хранилище данных

База данных

Механизм data mining’a

Архитектура data mining по Хану и Кэмберу

**База знаний:** Она служит отправной точкой для всего процесса интеллектуального анализа данных. Он действует как руководство для поиска или оценки интересности полученных паттернов. Такой тип знаний может включать в себя иерархии понятий, которые организуют атрибуты или их значения в различные стадии абстракции.

**Механизм data mining’a:** Он является основным компонентом системы интеллектуального анализа данных, он состоит из всех необходимых модулей, таких как характеристика, прогнозирование, кластерный анализ, анализ выбросов и эволюционный анализ для выполнения задач интеллектуального анализа данных.

**Модуль оценки шаблонов:** Этот модуль обычно связан с показателями интересности. Он постоянно взаимодействует с механизмом интеллектуального анализа данных для поиска закономерностей. Часто он использует пороговые значения для отсеивания обнаруженного шаблона или может использовать модуль оценки шаблона, интегрированный с модулем интеллектуального анализа данных, в зависимости от используемого метода интеллектуального анализа данных.

**Пользовательский интерфейс:** Модуль служит связующим звеном между пользователями и системой интеллектуального анализа данных. Это облегчает взаимодействие пользователей с системой простым и эффективным способом, не беспокоя пользователя о сложностях, стоящих за процессом.

**Источники данных** (www, хранилище данных, база данных, другие хранилища). Организации обычно хранят данные в базах данных или хранилищах данных. Иногда в хранилище данных содержится более одной базы данных, текстовых файлов или электронных таблиц.

**База данных или сервер хранилища данных:** Они содержит данные, которые настроены для извлечения. Получение данных по запросу пользователей является их ключевым назначением.

**Другие обработки:** перед передачей данных на сервер хранилища данных данные необходимо очистить и интегрировать, поскольку данные собираются из разных источников и в разных форматах, поэтому их нельзя использовать непосредственно для процесса интеллектуального анализа данных. Данные должны быть очищены, интегрированы, и только надежные данные должны быть отобраны и переданы на сервер хранилища данных. Для этого процесса может потребоваться ряд методов очистки, интеграции и выбора.

**Процесс data mining**

Файяд и др. (1996) определили “интеллектуальный анализ данных как один из нескольких этапов в процессе обнаружения знаний, он включает в себя применение анализа данных и обнаружение алгоритмов, которые дают точное перечисление шаблонов по данным 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 шаге для выполнения конкретной задачи обнаружения, каждая из которых выполняется с помощью применения метода обнаружения. Интеллектуальный анализ данных, синонимично используемый некоторыми для процесса обнаружения знаний в базах данных (KDD), и наоборот, многие рассматривают его как важный этап KDD, который приводит к созданию полезных шаблонов или моделей для данных. Различные этапы data mining показаны на схеме ниже.

Отбор

Данные

Целевая выборка

Препроцессинг

Предварительно обработанные данные

Преобразование

Преобразованные данные

Шаблоны

Data mining

Знания

Интерпретация и оценка

**Отбор**: Отбор соответствующих данных из различных источников для процесса интеллектуального анализа данных.

**Предварительная обработка (препроцессинг):** Поскольку данные собираются из разных источников, они содержат несоответствия, для устранения которых на этом этапе выполняются различные действия, исправляются или удаляются дефектные данные, устраняются шумы и расхождения и данные из разных источников объединяются.

**Преобразование:** Здесь данные передаются в соответствующую форму для интеллектуального анализа данных. Могут использоваться отбор признаков, выборка, агрегирование.

**Интеллектуальный анализ данных:** это важный шаг, при котором выбирается алгоритм интеллектуального анализа данных, соответствующий шаблону в данных. Также осуществляется извлечение шаблонов данных.

**Интерпретация и оценка:** распознавать и преобразовывать результаты интеллектуального анализа данных или закономерности в знания путем устранения избыточности и нерелевантных закономерностей. Здесь используется множество стратегий визуализации и графического интерфейса для преобразования выгодных шаблонов в понятные человеку термины.

**Задачи data mining**

Задачи data mining сгруппированы в две основные категории:

1: Прогнозирующие;

2: Описательные.

Файяд и др. (1996) определяют шесть основных функций майнинга данных:

1. Классификация

2. Регрессия

3. Кластеризация

4. Моделирование зависимостей

5. Обнаружение отклонений.

6. Обобщение.

Классификация, регрессия и обнаружение отклонений отнесены к прогностической категории, в то время как кластеризация, моделирование зависимостей отнесены к описательной категории. Прогностическая модель составляет прогнозы, используя некоторую переменную в наборе данных, чтобы предсказать неизвестные значения другой релевантной переменной, в то время как описательная модель классифицирует закономерности или взаимосвязи и охватывает понятные человеку закономерности и тенденции в данных (Флорин Горунеску 2011).

**Классификация:** классификация относится к классическим методам интеллектуального анализа данных, основанным на машинном обучении. Он находит общие свойства среди набора объектов в базе данных и классифицирует их по различным классам в соответствии с моделью классификации. Его основная цель - тщательно проанализировать данные обучения и разработать точное описание или модель для каждого класса, используя характерные черты присутствующие в данных. Этот метод использует математические методы, такие как деревья решений, нейронные сети и статистика (Минг-Сьян и др. 1996).

**Регрессия:** Это один из методов интеллектуального анализа данных, который определяет связь между зависимыми и независимыми переменными. Прогнозирование осуществляется с помощью поддержки регрессий. Статистически регрессия - это математическая модель, которая устанавливает связь между значениями зависимой переменной и значениями другого предиктора или независимой переменной. При регрессии прогнозируемая переменная может быть непрерывной переменной. При регрессии вещественнозначные прогнозирующие переменные сопоставляются с элементами обучающей функции. Статистическая регрессия, нейронная сеть, машинная регрессия опорных векторов - вот некоторые из часто используемых стратегий регрессии. Более сложные методы, такие как логистическая регрессия, деревья решений или нейронные сети, также могут быть использованы для прогнозирования будущих значений, эти методы также можно комбинировать для достижения лучшего результата.

**Кластеризация:** это метод интеллектуального анализа данных, который группирует физические или абстрактные объекты в классы похожих объектов . Кластеризация - это метод разделения набора данных (записей / кортежей / объектов/выборок) на несколько групп (кластеров) на основе предопределенного сходства. Основной целью кластеризации является нахождение групп (кластеров) объектов на основе сходства таким образом, чтобы внутри отдельного кластера было большое сходство друг с другом, в то время как кластеры достаточно отличаются друг от друга. В терминологии машинного обучения кластеризация - это форма неконтролируемого обучения.

**Моделирование зависимостей** (анализ ассоциативных правил): это один из лучших признанных методов интеллектуального анализа данных, который относится к неконтролируемому методу интеллектуального анализа данных, который направлен на поиск связей между элементами или записями, принадлежащими к большому набору данных, и выявляет существенные зависимости между переменными. Интеллектуальный анализ правил ассоциации является следствием формы X → Y, где x и y являются различными элементами или наборами элементов, содержащими инструкции if-then относительно значений атрибутов. При анализе рыночной корзины это правило широко используется, оно пытается проанализировать покупателей, покупающих определенные товары, и дает представление о сочетаниях, которые клиенты часто покупают вместе.

**Обнаружение аномалий:** синонимом его названия является обнаружение наиболее существенных изменений или отклонений от стандартного поведения.

**Обобщение:** Хотя это и не относится к методам интеллектуального анализа данных, но является результатом этих методов и имеет дело с определением компактного описания для подмножества данных, синонимично называемых обобщением или описанием.

**Последовательные шаблоны**: Обнаружение последовательности - это метод интеллектуального анализа данных, который используется для определения последовательных шаблонов, ассоциаций или регулярных событий / тенденций между переменными полями данных в течение бизнес-периода.

**Проблемы data mining**

Интеллектуальный анализ данных хорошо развит, но он все еще сталкивается с целым рядом проблем при его практической реализации. Некоторые из них обсуждаются ниже: **Проблемы безопасности:** Безопасность является наиболее важной проблемой, касающейся любого процесса обработки данных, учитывая чрезвычайно конфиденциальный характер данных, потенциальный незаконный доступ к знаниям должен быть предотвращен, и тайна должна соблюдаться.

**Проблемы методологии интеллектуального анализа данных:** поскольку разные пользователи проявляют интерес к различным видам знаний, интеллектуальный анализ данных должен охватывать широкий спектр задач анализа данных и обнаружения знаний, которые могут по-разному использовать одну и ту же базу данных и требуют разработки многочисленных методов интеллектуального анализа данных.

**Проблемы с пользовательским интерфейсом**: Обнаружение знаний с помощью инструментов майнинга данных выгодно и выразительно только до тех пор, пока они представлены явно и привлекательно для пользователя. Поскольку трудно понять, что может быть обнаружено в базе данных, процесс интеллектуального анализа данных должен быть интерактивным, данные должны быть представлены на языке высокого уровня, в визуальных представлениях или других графических формах выражения, чтобы пользователь мог понять и интерпретировать их и использовать по мере необходимости.

**Обработка зашумленных и неполных данных:** Данные, хранящиеся в базах данных, могут отличаться, поскольку различные проблемы связаны с источниками данных, данные могут быть неполными, данные могут содержать случаи, которые могут вызывать исключения. Интеллектуальный анализ данных с такими нарушениями приводит к возникновению неоднозначностей в процессе, в результате чего построенная модель знаний перегружает данные и снижает точность результирующих знаний, поэтому требуются методы интеллектуального анализа данных, которые могут справиться с этими несоответствиями.

**Проблемы с производительностью:** Для обработки данных эффективность и масштабируемость являются стратегическими факторами при внедрении data mining’a баз данных. Информация должна эффективно и квалифицированно извлекаться из баз данных, поскольку их объем достаточно велик. Используемые алгоритмы должны быть масштабируемыми, время их выполнения должно быть предсказуемым и приемлемым для больших баз данных.

**Применение data mining**

**Применение в здравоохранении:** data mining / интеллектуальный анализ данных может быть существенно выгоден в системе здравоохранения, но его успех зависит от наличия чистых данных. В здравоохранении он используется для диагностики и прогноза заболевания, также может быть установлена взаимосвязь между заболеваниями. Врачи могут определить эффективные методы лечения для каждого пациента. Поскольку огромное количество медицинских данных является сложным и обширным для обработки и анализа, data mining предоставляет методологию и инструменты для преобразования данных в информацию для принятия эффективных решений.

**Применение в образовательных системах:** Интеллектуальный анализ данных в образовательных системах - развивающаяся область, к которой исследователи проявляют большой интерес. Методы интеллектуального анализа данных могут помочь восполнить пробел в знаниях в системе образования путем выявления скрытых закономерностей, коннотаций и различий. Это помогает заинтересованным сторонам повысить эффективность принятия решений, что приводит к совершенствованию образовательной системы.

**Применение в CRM**: Интеллектуальный анализ данных в CRM в настоящее время является наиболее обсуждаемой темой исследований в промышленности и академических кругах с целью предоставления резюме исследований по использованию методов интеллектуального анализа данных в области CRM.

**Применение при анализе рыночной корзины (MBA**): Для анализа рыночной корзины используются различные методы интеллектуального анализа данных -MBA, метод, который помогает выявить связь между различными товарами, которые покупатель кладет в свою корзину во время покупок, он отслеживает покупательские привычки покупателей. Торговые дома могут использовать методы сбора данных для выявления моделей покупок и поведения клиента, на основе которых клиенту может быть представлен широкий выбор в соответствии с его привычкой к покупкам.

**Применение в спортивных данных:** Методы интеллектуального анализа данных также проникли в сферу спорта. Проводится огромное количество игр, и каждый вид спорта генерирует огромное количество статистических данных. Эти массивные данные необходимо поддерживать в отношении планирования событий и статистики отдельного игрока в этих событиях. Интеллектуальный анализ данных может использоваться для прогнозирования и анализа производительности, а также для стратегического планирования.