ПРИМЕНЕНИЕ ТЕХНОЛОГИЙ DATA MINING В БИЗНЕС-АНАЛИТИКЕ

А. Р. Филич, Д. А. Мисуно

Белорусский государственный университет, г. Минск;
alex.filich2001@gmail.com;midariam@gmail.com;
науч. рук. – Н. Н. Васюкевич

Как показывает международный опыт, отсутствие достаточно эффективных методов обработки данных зачастую вызывает критические проблемы бизнеса, которые ставят под угрозу само существование и операционную деятельность предприятия. Технологии анализа данных в бизнес-аналитике предлагают решение данных проблем, облегчая непосредственное исследование и интерпретацию бизнес-информации. В статье рассматриваются основные методы инструментов Data Mining и их применение на примере сервисов Microsoft для бизнес-аналитики.

Ключевые слова: Data Mining; интеллектуальный анализ данных; Knowledge Discovery Process; ETL-инструменты; Business Intelligence; бизнес-аналитика.

Под понятием Data Mining (от англ. «добыча/раскопка данных») как правило понимают интеллектуальный анализ данных, или совокупность методов обнаружения в данных ранее неизвестных, нетривиальных, практически полезных и доступных интерпретации знаний, необходимых для принятия решений в различных сферах деятельности. В Data Mining выделяют большое множество методов исследования данных (табл.) [1].

Основные типы методов исследования данных Data Mining

Метод	Характеристика
Статистический анализ	Включает различные виды статистических анализов. Такие методы исследования данных реализуются в виде статистических пакетов и программных продуктов (SPSS Statistics, Statistica, EViews и другое)
Алгоритмы k-ближайшего соседа	Совокупность конкретно заданных решений, которые заключаются в выборе близкого аналога исходных данных из уже имеющихся исторических данных
Нейронные сети	Такие сети рассматривают input как сигналы, преобразующиеся в соответствии с имеющимися связями между «нейронами», а в качестве output – отклик всей сети на исходные данные. Такие алгоритмы не программируются, а обучаются.
Деревья решений	Иерархическая структура принятия решений, которая основывается на структуре типа «листья» - «ветви», где «ветвями» выступают признаки, от которых зависит целевая функция, а в «листьях» указаны значения целевой функции.
Кластерные модели	Модели, в основе которых лежат принципы кластеризации, когда схожие данные объединяются в кластеры на основании схожих характеристик

Зачастую синонимично понятию Data Mining используют термин Knowledge Discovery Process (KDP), или Knowledge Discovery in Databases (KDD), который буквально подразумевает «обнаружение знаний (в базах данных)» (рис. 1) [2].

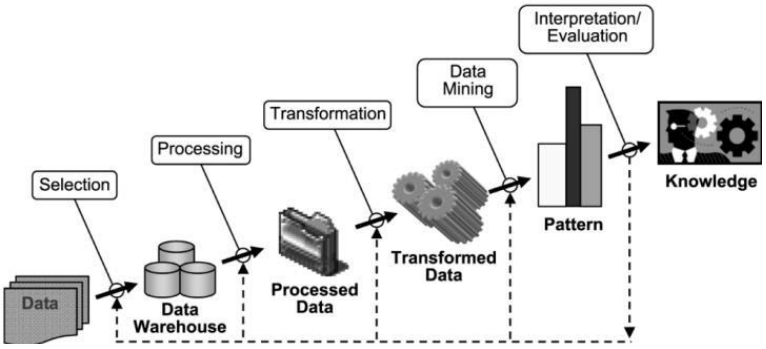

Рис. 1. Этапы Knowledge Discovery Process
Примечание – Источник: [2]

В первую очередь исходные данные (Data), поступая в обработку, проходят этап отбора (Selection): из источников извлекаются только необходимые и релевантные для задачи данные. Сформированные базы и хранилища данных (Data Warehouse) подвергаются очистке от шумов, выбросов и прочих неточностей (Processing). Полученные массивы данных (Processed Data) преобразуются в форму, пригодную для дальнейшего анализа инструментами Data Mining. После этого преобразованные данные (Transformed Data) проходят через самый важный этап – этап интеллектуального анализа данных (Data Mining), в результате алгоритмов которого происходит процесс извлечения потенциально полезных шаблонов (Patterns). По итогам интерпретации (Interpretation/Evaluation) происходит оценка этих паттернов и представление их в удобной для пользователя форме. На завершающем этапе процесса KDP (Knowledge) происходит внедрение извлеченных знаний в какое-либо прикладное решение, приложение, сервис и т.д., например, в виде дашбордов, отчетов и прочих визуализаций

Ни один инструмент Data Mining невозможно представить без технологии ETL. ETL (от англ. Extract, Transform, Load – «извлечение, преобразование, загрузка») – это совокупность инструментов по извлечению, преобразованию и загрузке данных из одного источника в другой. Современные инструменты ETL выполняют автоматическую и быструю пакетную обработку данных. Если останавливаться на стеке от Microsoft, то необходимые технологии представлены сервисами SSIS, SSAS и SSRS [3].

SQL Server Integration Services (SSIS) – это платформа для построения интеграции и преобразования данных на уровне предприятия. Она представлена пакетами, способными извлекать и видоизменять данные, до-

бытые из всевозможных источников. После этого происходит загрузка данных на одно или несколько хранилищ.

SQL Server Analysis Services (SSAS) – инструмент из стека Microsoft Business Intelligence для разработки аналитических решений на основе технологии Online Analytical Processing (OLAP), в результате чего создаются многомерные модели данных – OLAP-кубы, предназначенные для более глубокого и быстрого анализа данных. При этом, для извлечения данных из куба используется Multi-dimensional expression (MDX) – специальный SQL-подобный язык запросов, предназначенный для получения доступа к многомерным структурам данных.

SQL Server Reporting Services (SSRS) – программное обеспечение Microsoft, которое представляет собой инструмент по созданию отчетов на основании данных, проанализированных и подсчитанных на предыдущих этапах ETL-процесса.

Проведя характеристику каждой технологии, следует объединить все компоненты сервисов вместе с такой СУБД, как Microsoft SQL Server, в одну общую архитектуру (рис. 2).

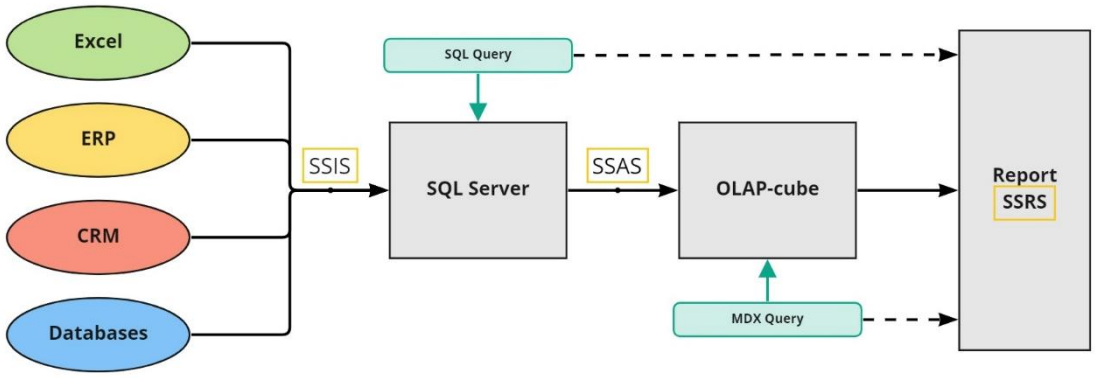

Рис. 2. Сервисы Microsoft SQL Server для бизнес-аналитики
Примечание – Источник: разработка автора на основе [3]

Данные, извлеченные из разнородных источников посредством инструментов SSIS, изменяются и загружаются на SQL Server. Далее сформированная SSAS-инструментами OLAP-структура предоставляет данные SSRS-платформе для создания конечных отчетов. На этапе SSRS доступ к данным может быть предоставлен как через SQL-запросы (к данным из SQL Server), так и через MDX-запросы (к данным OLAP-куба).

Наконец, стоит рассмотреть преимущества и недостатки стека технологий от корпорации Microsoft. Если рассматривать положительные стороны данных сервисов, то в первую очередь необходимо отметить, что данное ПО использует передовые технологии в DWH, при этом поддерживая широкий набор документации и используя новые подходы стан-

дартизированной интеграции данных и оперативного мониторинга их колебаний. Более того, отмечается высокая степень простоты имплементации платформы по сравнительно недорогой цене по рынку. Главный недостаток, как и ожидалось, заключается в возникновении проблем при работе на операционных системах, отличных от Windows [4].

Подводя итог о современных инструментах Data Mining в бизнес-аналитике, можно заключить следующее. Технологии интеллектуального анализа данных дают возможность открывать ранее неизвестные, нетривиальные, практически полезные и доступные интерпретации знания в данных, что значительным образом помогает оптимизировать операционную деятельность бизнеса. В то же время, такие инструменты входят в стек технологий Business Intelligence, что обеспечивает полноценность KDD-процессов. Яркими примерами инструментов ETL являются сервисы Microsoft SQL Server (SSIS, SSAS, SSRS). Исходя из этого, закономерным является увеличение объемов рынка инструментов Data Mining, подтверждаемое многочисленными исследованиями.

Библиографические ссылки

1. *Тучкова, А. С.* Термин "Data Mining". Задачи, решаемые методами Data Mining / А. С. Тучкова, П. П. Кондрашева // Тенденции развития науки и образования. – 2019. – № 55-2. – С. 27–30.
2. Data Mining: Practical Machine Learning Tools and Techniques / Ian H. Witten [и др.]. – Chennai: Morgan Kaufmann, 2017. – С. 4–9.
3. Difference Between Microsoft SSRS, SSAS and SSIS [Электронный ресурс]. – Режим доступа: https://codestoresolutions.com/difference-between-microsoft-ssrs-ssas-and-ssis/. – Дата доступа: 11.05.2022.
4. *Geetha, S.* Data Analysis and ETL Tools in Business Intelligence / S. Geetha, D. Keval Rajesh, D. Param Pankaj // International Research Journal of Computer Science (IRJCS). – 2020. – Т. 07, № 05. – С. 127–132.