Результатом развития информационных технологий и сети Internet является количество информации, накопленная человечеством в электронном виде: тексты, изображения, аудио, видео, гипертекстовые документы, базы данных и т. д. Современные системы извлечения информации используют основанные на методах искусственного интеллекта средства представления и интерпретации для поиска в терабайтных хранилищах весьма ценную информацию.

Большинство современных программ контент-анализа ограничены обработкой текста, однако их возможности гораздо шире. Примером технологии этого поколения является технология "добычи" данных или Text Mining. Вообще результатом естественной эволюции информационных технологий стали облачные технологии и методы (классификация, кластеризация, прогнозирование) и технологии (Data Mining, Text Mining, Web Mining, OLAP) интеллектуального анализа данных Причинами их популярности стали следующие: стремительное накопление данных; всеобщая компьютеризация; проникновения Интернет во все сферы деятельности; прогресс в области информационных технологий (совершенствование СУБД и хранилищ данных); прогресс в области производственных технологий (рост производительности компьютеров, объемов накопителей, внедрение Grid-систем).

Несмотря на количество методов Data Mining, приоритет все больше смещается в сторону логических алгоритмов поиска данных if-then алгоритмов,

с помощью которых решаются задачи прогнозирования, классификации, распознавания образов, сегментации БД, извлечения из данных скрытых знаний, интерпретации данных, установления ассоциаций в БД и прочее. Результаты таких алгоритмов эффективны и легко интерпретируются. Но главной проблемой логических методов выявления закономерностей проблема перебора вариантов за ограниченное время. Эти методы искусственно ограничивают такой перебор и строят дерева решений с принципиальными ограничениями эффективности поиска if-then правил.

Основная особенность Data Mining - это сочетание широкого математического инструментария (от классического статистического анализа к новым кибернетических методов) и последних достижений в сфере информационных технологий. В технологии Data Mining гармонично объединились строго формализованные методы и методы неформального анализа, т.е. количественный и качественный анализы данных. Большинство аналитических методов, используемых в технологии Data Mining - это известные математические алгоритмы и методы. Новым является то, что их можно применять при решении тех или иных конкретных проблем. Это обусловлено новыми свойствами технических и программных средств.

Knowledge Discovery in Databases (дословно: «выявление знаний в базах данных» - KDD) - аналитический процесс исследования больших объемов информации с привлечением средств автоматизации, имеет целью выявить скрытые в множестве данных структуры, зависимости и взаимосвязи. При этом предполагается полная или частичная отсутствие априорных представлений о

характере скрытых структур и зависимостей. KDD предполагает, что человек предварительно осмысливает задачу и подает неполное (в терминах целевых переменных) ее формулировки, преобразует данные в формат пригодного для их автоматизированного анализа и предварительной обработки, проявляет средствами автоматического исследования данных скрытые структуры и зависимости, апробирует обнаружены модели на новых данных, неиспользуемых для построения моделей, и интерпретирует обнаружены модели и результаты.

Итак, KDD – это синтетическая технология, сочетающая в себе последние достижения искусственного интеллекта, многочисленных математических методов, статистики и эвристических подходов. Методы KDD особенно стремительно развиваются в течение последних 20 лет, а ранее задачи компьютерного анализа баз данных выполнялись преимущественно с помощью разного рода стандартных статистических методов.

Технология KDD позволяет не только подтверждать (отбрасывать) эмпирические выводы, но и строить новые, неизвестные ранее модели. Найдена модель не сможет основном претендовать на абсолютное знание, но она предоставляет аналитику некоторые преимущества уже благодаря самому факту обнаружения альтернативной статистически значимой модели, а также, возможно, становится поводом для поиска ответа на вопрос: действительно ли существует выявлена взаимосвязь и является ли он причинным? А это, в свою очередь, стимулирует углубленные исследования, способствуя более глубокому пониманию изучаемого явления.

Итак, важнейшая цель применения технологии KDD к исследованию реальных систем – это улучшение понимания сути их функционирования. Отметим, что процесс выявления знаний не вполне автоматизированным - он требует участия пользователя (эксперта, принимающего решение). Пользователь должен четко осознавать, что он ищет, основываясь на собственных догадках. В конце концов вместо того, чтобы подтверждать имеющуюся гипотезу, процесс поиска часто способствует появлению ряда новых гипотез. Все это обозначается термином «discovery-driven data mining» (DDDM), и сроки Data Mining, Knowledge Discovery в общем случае относятся к технологии DDDM.