ЛЕКЦИЯ 4. ЗАДАЧИ DATA MINING. ИНФОРМАЦИЯ И ЗНАНИЯ

Напомним, что в основу технологии *Data* Mining положена концепция шаблонов, представляющих собой закономерности. В результате обнаружения этих, скрытых от невооруженного глаза закономерностей решаются *задачи Data Mining*. Различным типам закономерностей, которые могут быть выражены в форме, понятной человеку, соответствуют определенные *задачи Data Mining*.

Задачи (tasks) Data Mining иногда называют закономерностями (regularity) или техниками (techniques).

Единого мнения относительно того, какие задачи следует относить к *Data* Mining, нет. Большинство авторитетных источников перечисляют следующие: *классификация*, *кластеризация*, *прогнозирование*, *ассоциация*, *визу ализация*, *анализ* и обнаружение отклонений, *оценивание*, *анализ связей*, *подведение итогов*.

Цель описания, которое следует ниже, - дать общее *представление* о *задачах Data Mining*, сравнить некоторые из них, а также представить некоторые методы, с помощью которых эти задачи решаются. Наиболее распространенные *задачи Data Mining* - *классификация*, *кластеризация*, *ассоциация*, *прогнозирование* и *визуализация* - будут подробно рассмотрены в последующих лекциях. Таким образом, задачи подразделяются *по* типам производимой *информации* , это наиболее общая *классификация задач Data Mining*. Дальнейшее детальное знакомство с методами решения *задач Data Mining* будет представлено в следующем разделе курса.

Задачи Data Mining

Классификация (*Classification*)

Краткое описание. Наиболее простая и распространенная задача *Data* Mining. В результате решения задачи классификации обнаруживаются признаки, которые характеризуют группы объектов исследуемого набора данных - классы; *по* этим признакам новый *объект* можно отнести к тому или иному классу.

Методы решения. Для решения задачи классификации могут использоваться методы: ближайшего соседа (Nearest Neighbor); k-ближайшего соседа (k-Nearest Neighbor); байесовские сети (*Bayesian Networks*); *индукция* деревьев решений; нейронные сети (*neural networks*).

Кластеризация (Clustering)

Краткое описание. *Кластеризация* является логическим продолжением идеи классификации. Это задача более сложная, особенность кластеризации заключается в том, что классы объектов изначально не предопределены. Результатом кластеризации является *разбиение* объектов на группы.

Пример метода решения задачи кластеризации: обучение "без учителя" особого вида нейронных сетей - *самоорганизующихся карт* Кохонена.

Ассоциация (Associations)

Краткое описание. В ходе решения задачи поиска ассоциативных правил отыскиваются закономерности между связанными событиями в наборе данных.

Отличие *ассоциации* от двух предыдущих *задач Data Mining*: *поиск* закономерностей осуществляется не на основе свойств анализируемого объекта, а между несколькими событиями, которые происходят одновременно.

Наиболее известный *алгоритм* решения задачи поиска ассоциативных правил - *алгоритм* Apriori.

Последовательность (*Sequence*), или последовательная *ассоциация* (*sequential association*)

Краткое описание. *Последовательность* позволяет найти временные закономерности между транзакциями. Задача *последовательности* подобна *ассоциации*, но ее целью является установление закономерностей не между одновременно наступающими событиями, а между событиями, связанными во времени (т.е. происходящими с некоторым определенным интервалом во времени). Другими словами, *последовательность* определяется высокой вероятностью цепочки связанных во времени событий. Фактически, *ассоциация* является частным случаем *последовательности* с временным шагом, равным нулю. Эту *задачу Data Mining* также называют задачей нахождения последовательных шаблонов (*sequential pattern*).

Правило *последовательности*: после события X через определенное время произойдет событие Y.

Пример. После покупки квартиры жильцы в 60% случаев в течение двух недель приобретают холодильник, а в течение двух месяцев в 50% случаев приобретается *телевизор*. Решение данной задачи широко применяется в маркетинге и менеджменте, например, при управлении циклом работы с клиентом (*Customer* Lifecycle *Management*).

Прогнозирование (*Forecasting*)

Краткое описание. В результате решения задачи прогнозирования на основе особенностей исторических данных оцениваются пропущенные или же будущие значения целевых численных показателей.

Для решения таких задач широко применяются методы математической статистики, нейронные сети и др.

Определение отклонений или выбросов (*Deviation Detection*), *анализ* отклонений или выбросов

Краткое описание. Цель решения данной задачи - обнаружение и *анализ данных*, наиболее отличающихся от общего *множества* данных, выявление так называемых нехарактерных шаблонов.

Оценивание (*Estimation*)

Задача *оценивания* сводится к предсказанию непрерывных значений признака.

Анализ связей (Link *Analysis*) - задача нахождения зависимостей в наборе данных.

Визуализация (*Visualization, Graph* Mining)

В результате *визуализации* создается графический образ анализируемых данных. Для решения задачи *визуализации* используются графические методы, показывающие наличие закономерностей в данных.

Пример методов *визуализации* - *представление* данных в 2D и *3D* измерениях.

Подведение итогов (*Summarization*) - задача, цель которой - описание конкретных групп объектов из анализируемого набора данных.

Классификация задач Data Mining

Согласно классификации по стратегиям, *задачи Data Mining* подразделяются на следующие группы:

- *обучение с учителем*;
- *обучение без учителя*;
- другие.

Категория *обучение с учителем* представлена следующими *задачами Data Mining*: *классификация*, оценка, *прогнозирование*.

Категория *обучение без учителя* представлена задачей кластеризации.

В категорию другие входят задачи, не включенные в предыдущие две стратегии.

Задачи Data Mining, в зависимости от используемых моделей, могут быть дескриптивными и прогнозирующими. Эти типы моделей будут подробно описаны в лекции, посвященной процессу Data Mining.

В соответствии с этой классификацией, *задачи Data Mining* представлены группами *описательных* и *прогнозирующих задач*.

В результате решения *описательных (descriptive) задач* аналитик получает шаблоны, описывающие данные, которые поддаются интерпретации.

Эти задачи описывают общую концепцию анализируемых данных, определяют информативные, итоговые, отличительные особенности данных. Концепция *описательных задач* подразумевает характеристику и сравнение наборов данных.

Характеристика набора данных обеспечивает краткое и сжатое описание некоторого набора данных.

Сравнение обеспечивает сравнительное описание двух или более наборов данных.

Прогнозирующие (predictive) основываются на анализе данных, создании модели, предсказании тенденций или свойств новых или неизвестных данных.

Достаточно близким к вышеупомянутой классификации является подразделение *задач Data Mining* на следующие: исследования и открытия, прогнозирования и классификации, объяснения и описания.

Автоматическое исследование и открытие (*свободный поиск*)

Пример задачи: обнаружение новых сегментов рынка.

Для решения данного класса задач используются методы кластерного анализа.

прогнозирование и *классификация*

Пример задачи: предсказание роста объемов продаж на основе текущих значений.

Методы: регрессия, нейронные сети, *генетические алгоритмы*, деревья решений.

Задачи классификации и прогнозирования составляют группу так называемого индуктивного моделирования, в результате которого обеспечивается изучение анализируемого объекта или системы. В процессе решения этих задач на основе набора данных разрабатывается общая модель или гипотеза.

Объяснение и описание

Пример задачи: характеристика клиентов по демографическим данным и историям покупок.

Методы: деревья решения, системы правил, правила *ассоциации*, *анализ связей*.

Если доход клиента больше, чем 50 условных единиц, и его возраст - более 30 лет, тогда класс клиента - первый.

В интерпретации обобщенной модели аналитик получает новое знание. Группировка объектов происходит на основе их сходства.

Связь понятий

Итак, в предыдущей лекции нами были рассмотрены методы Data Mining и действия, выполняемые в рамках стадий Data Mining. Только что мы рассмотрели основные *задачи Data Mining*.

Напомним, что главная ценность Data Mining - это практическая направленность данной технологии, путь от сырых данных к конкретному знанию, от постановки задачи к готовому приложению, при поддержке которого можно принимать решения.

Многочисленность понятий, которые объединились в Data Mining, а также разнообразие методов, поддерживающих данную технологию, начинающему аналитику могут напомнить мозаику, части которой мало связаны между собой.

Как же мы можем связать в одно целое задачи, методы, действия, закономерности, приложения, данные, информацию, решения?

Рассмотрим два потока:

1. **ДАННЫЕ - ИНФОРМАЦИЯ - ЗНАНИЯ И РЕШЕНИЯ**
2. **ЗАДАЧИ - ДЕЙСТВИЯ И МЕТОДЫ РЕШЕНИЯ - ПРИЛОЖЕНИЯ**

Эти потоки являются "двумя сторонами одной медали", отображением одного процесса, результатом которого должно быть знание и принятие решения.

От данных к решениям

Для начала рассмотрим первый *поток*. На рис. 4.1.показана *связь* понятий "данные", "*информация*" и "решения", которая возникает в процессе *принятия решений*.

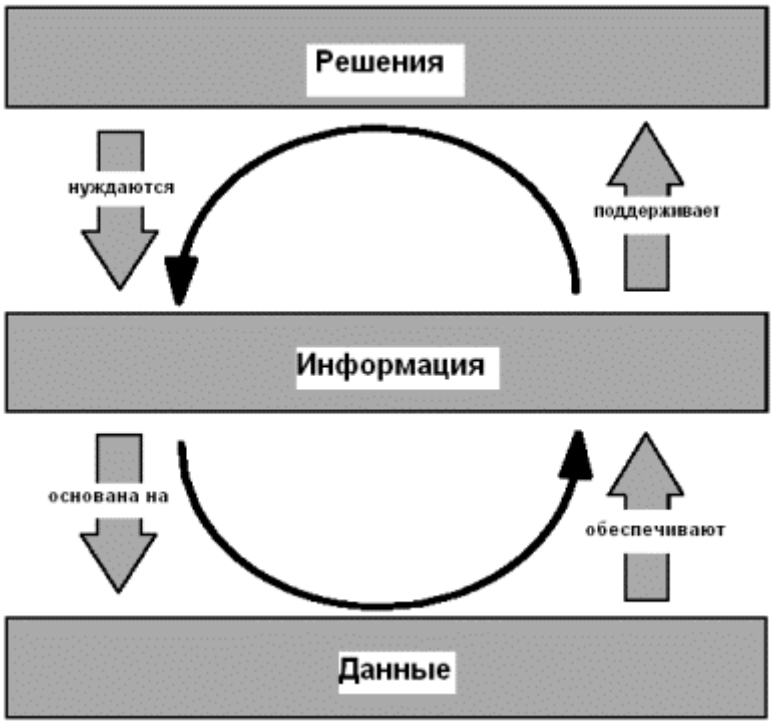

Рис. 4.1. Решения, информация и данные

Как видно из рисунка, данный процесс является циклическим. Принятие решений требует *информации*, которая основана на данных. Данные обеспечивают информацию, которая поддерживает решения, и т.д.

Рассмотренные понятия являются составной частью так называемой ***информационной пирамиды***, в основании которой находятся данные, следующий уровень - это *информация*, затем идет решение, завершает пирамиду уровень знания. *По* мере продвижения вверх *по информационной пирамиде* объемы данных переходят в ценность решений, т.е. ценность для бизнеса. А, как известно, целью *Business Intelligence* является преобразование объемов данных в ценность бизнеса.

От задачи к приложению

Теперь подойдем к этому же процессу с другой стороны. Рассмотрим рис. 4.2. *По* словам авторов, он не претендует на полноту, зато отображает все уровни, которые затрагивает *Data* Mining.

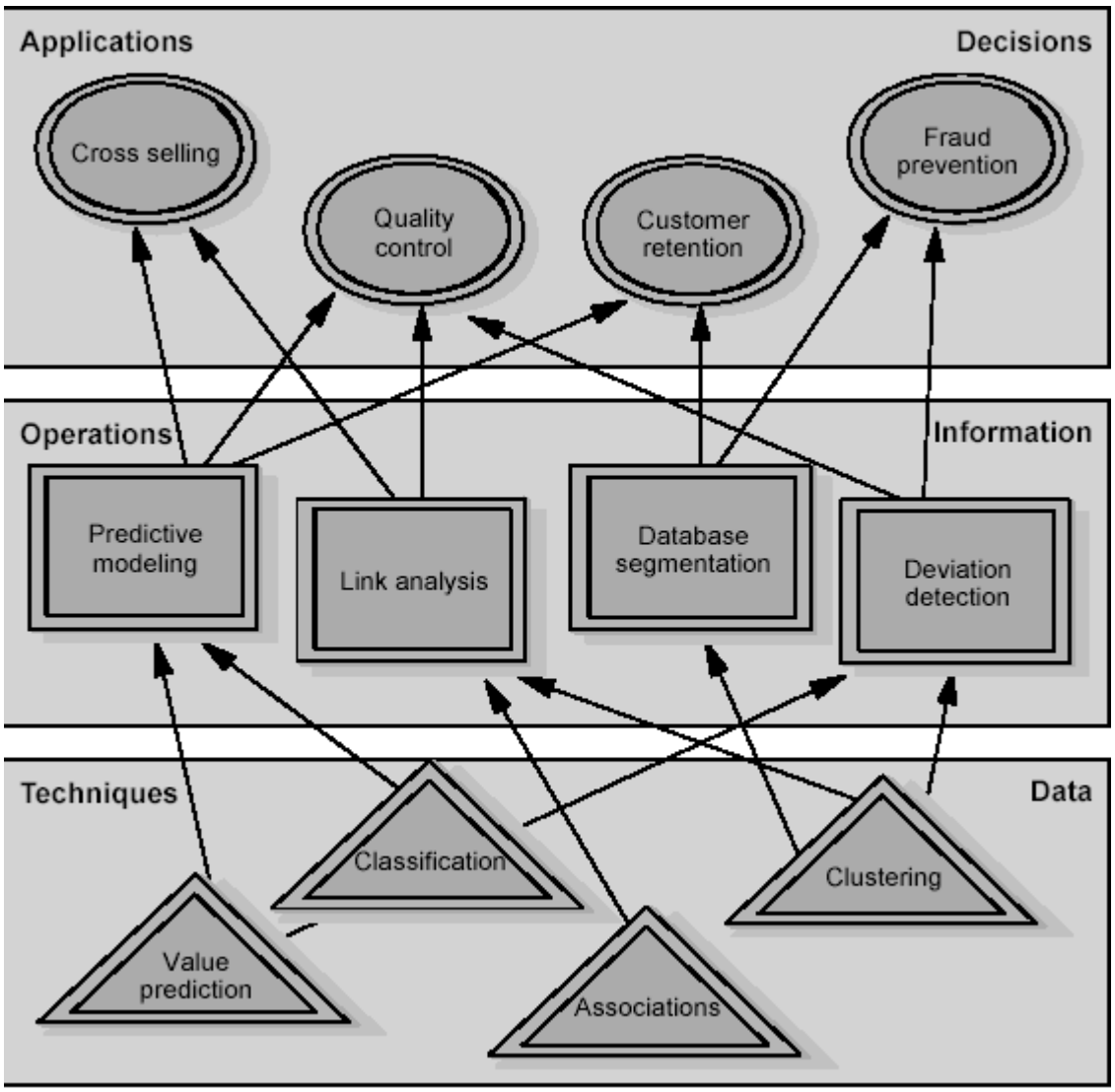

Рис. 4.2. Задачи, действия, приложения

Следует отметить, что уровни анализа (данные, *информация*, *знания*) практически соответствуют этапам эволюции анализа данных, которая происходила на протяжении последних лет.

Верхний - уровень приложений - является уровнем бизнеса (если мы имеем дело с задачей бизнеса), на нем менеджеры принимают решения. Приведенные примеры приложений: перекрестные продажи, *контроль* качества, удерживание клиентов.

Средний - уровень действий - *по* своей сути является уровнем *информации*, именно на нем выполняются действия *Data* Mining; на рисунке приведены такие действия: *прогностическое моделирование* (было рассмотрено в предыдущей лекции), *анализ связей*, *сегментация* данных и другие.

Нижний - уровень определения *задачи Data Mining*, которую необходимо решить применительно к данным, имеющимся в наличии; на рисунке приведены задачи предсказания числовых значений, *классификация*, *кластеризация*, *ассоциация*.

Рассмотрим таблицу, демонстрирующую *связь* этих понятий.

Таблица 4.1. Уровни Data Mining				
уровень 3	приложения	удержание клиентов	знания	Data Mining результат
уровень 2	действия	*прогностическое моделирование*	информация	метод анализа
уровень 1	задачи	классификация	данные	запросы

Напомним, что для решения задачи классификации результаты работы первой стадии (индукции правил) используются для отнесения нового объекта, с определенной уверенностью, к одному из известных, предопределенных классов на основании известных значений.

Рассмотрим задачу удержания клиентов (определения надежности клиентов фирмы).

Первый уровень. Данные - *база данных по* клиентам. Есть данные о клиенте (возраст, пол, *профессия*, доход). Определенная часть клиентов, воспользовавшись продуктом фирмы, осталась ей верна; другие клиенты больше не приобретали продукты фирмы. На этом уровне мы определяем тип задачи - это задача классификации.

На **втором уровне** определяем действие - *прогностическое моделирование*. С помощью *прогностического моделирования* мы с определенной долей уверенности можем отнести новый *объект*, в данном случае, нового клиента, к одному из известных классов - постоянный клиент, или это, скорее всего, его разовая покупка.

На **третьем уровне** мы можем воспользоваться приложением для принятия решения. В результате приобретения знаний, *фирма* может существенно снизить *расходы*, например, на рекламу, зная заранее, каким из клиентов следует активно рассылать рекламные материалы.

Таким образом, на протяжении нескольких лекций мы определились с понятиями "данные", "задачи", "методы", "действия".

Информация

Сейчас остановимся на еще не рассмотренном понятии *информации*. Несмотря на распространенность данного понятия, мы не всегда можем точно его определить и отличить от понятия данных. *Информация, по* своей сути, имеет многогранную природу. С развитием человечества, в том числе, с развитием компьютерных технологий, *информация* обретает все новые и новые свойства.

Обратимся к словарю. ***Информация*** (лат. informatio) -

1. любые сообщения о чем-либо;
2. сведения, являющиеся объектом хранения, переработки и передачи (например генетическая информация);
3. в математике (кибернетике) - количественная мера устранения неопределенности (энтропия), мера организации системы; в теории *информации* - раздел кибернетики, изучающий количественные закономерности, которые связаны со сбором, передачей, преобразованием и вычислением *информации*.

Информация - любые, неизвестные ранее сведения о каком-либо событии, сущности, процессе и т.п., являющиеся объектом некоторых операций, для которых существует содержательная *интерпретация*.

Под операциями здесь подразумевается восприятие, передача, преобразование, хранение и использование. Для восприятия *информации* необходима некоторая воспринимающая система, которая может интерпретировать ее, преобразовывать, определять соответствие определенным правилам и т.п. Таким образом, понятие *информации* следует рассматривать только при наличии источника и получателя *информации*, а также канала связи между ними.

Свойства информации

• Полнота *информации*.

Это свойство характеризует качество *информации* и определяет достаточность данных для принятия решений, т.е. *информация* должна содержать весь необходимый набор данных.

Пример. "Продажи товара А начнут сокращаться" Эта *информация* неполная, поскольку неизвестно, когда именно они начнут сокращаться.

Пример полной *информации*. "Начиная с первого квартала, продажи товара А начнут сокращаться." Этой *информации* достаточно для принятия решений.

• Достоверность *информации*.

Информация может быть достоверной и недостоверной. В недостоверной *информации* присутствует *информационный шум*, и чем он выше, тем ниже достоверность *информации*.

• Ценность *информации*.

Ценность *информации* не может быть абстрактной. Информация должна быть полезной и ценной для определенной категории пользователей.

• Адекватность *информации*.

Это свойство характеризует степень соответствия *информации* реальному объективному состоянию. Адекватная *информация* - это полная и достоверная *информация*.

• Актуальность *информации*.

Информация должна быть актуальной, т.е. не устаревшей. Это свойство *информации* характеризует степень соответствия *информации* настоящему моменту времени.

• Ясность *информации*.

Информация должна быть понятна тому кругу лиц, для которого она предназначена.

• Доступность *информации*.

Доступность характеризует меру возможности получить определенную информацию. На это свойство *информации* влияют одновременно доступность данных и доступность адекватных методов.

• Субъективность *информации*.

Информация носит субъективный характер, она определяется степенью восприятия субъекта (получателя *информации*).

Требования, предъявляемые к информации

• Динамический характер *информации*.

Информация существует только в момент взаимодействия данных и методов, т.е. в момент информационного процесса. Остальное время она пребывает в состоянии данных.

• Адекватность используемых методов.

Информация извлекается из данных. Однако в результате использования одних и тех же данных может появляться разная *информация*. Это зависит от адекватности выбранных методов обработки исходных данных.

Данные, по своей сути, являются объективными. Методы являются субъективными, в основе методов лежат алгоритмы, субъективно составленные и подготовленные. Таким образом, *информация* возникает и существует в момент диалектического взаимодействия объективных данных и субъективных методов.

Для бизнеса *информация* является исходной составляющей принятия решений.

Всю информацию, возникающую в процессе функционирования бизнеса и управления им, можно классифицировать определенным образом. В зависимости от источника получения, информацию разделяют на внутреннюю и внешнюю (например, *информация*, описывающая явления, происходящие за пределами фирмы, но имеющие к ней непосредственное отношение).

Также *информация* может быть классифицирована на фактическую и прогнозную. К фактической *информации* о бизнесе относится *информация*, характеризующая свершившиеся факты; она является точной. Прогнозная *информация* является рассчитываемой или предполагаемой, поэтому ее нельзя считать точной, она может иметь определенную погрешность.

Знания

Знания - совокупность фактов, закономерностей и эвристических правил, с помощью которых решается поставленная задача.

Итак, формирование *информации* происходит в процессе сбора и передачи, т.е. обработки данных. Каким же образом из *информации* получают *знания*?

Все чаще истинные *знания* образуются на основе распределенных взаимосвязей разнородной *информации* . Когда *информация* собрана и передана для получения явно не определенного заранее результата, то вы получаете *знания*. Сама *по* себе *информация* в чистом виде бессмысленна. Отсюда следует *вывод*, что *информация* - это чье-то тактическое *знание*, передаваемое в виде символов и при помощи каких-либо прикладных средств.

По определению Денхема Грэя, " *знания* - это абсолютное использование *информации* и данных, совместно с потенциалом практического опыта людей, способностями, идеями, интуицией, убежденностью и мотивациями".

Знания имеют определенные свойства, которые отличают их от *информации* .

1. **Структурированность.** *Знания* должны быть "разложены по полочкам".

2. **Удобство доступа и усвоения**. Для человека - это способность быстро понять и запомнить или, наоборот, вспомнить; для компьютерных знаний - средства доступа к *знаниям*.

3. **Лаконичность**. Лаконичность позволяет быстро осваивать и перерабатывать *знания* и повышает "коэффициент полезного содержания". В данный список лаконичность была добавлена из-за всем известной проблемы шума и мусорных документов, характерной именно для компьютерной *информации* - Internet и *электронного документооборота*.

4. **Непротиворечивость.** *Знания* не должны противоречить друг другу.

5. **Процедуры обработки**. *Знания* нужны для того, чтобы их использовать. Одно из главных свойств знаний - возможность их передачи другим и способность делать выводы на их основе. Для этого должны существовать процедуры обработки знаний. Способность делать выводы означает для машины наличие процедур обработки и вывода и подготовленность структур данных для такой обработки, т.е. наличие специальных форматов знаний.

Сопоставление и сравнение понятий "информация", "данные", "знание"

Для того чтобы уверенно оперировать понятиями "*информация*", "данные", "*знание*", необходимо не только понимать суть этих понятий, но и прочувствовать отличия между ними. Однако, одной интуитивной интерпретации этих понятий здесь недостаточно. Сложность понимания отличий вышеупомянутых понятий - в их кажущейся синонимичности. Вспомним, что понятие *Data* Mining переводится на русский язык при помощи этих же трех понятий: как добыча данных, извлечение *информации*, раскопка знаний.

Для начала сделаем попытку разобраться в этих терминах на простых примерах.

1. Студент, который сдает экзамен, нуждается в данных.
2. Студент, который сдает экзамен, нуждается в *информации*.
3. Студент, который сдает экзамен, нуждается в *знаниях*.

При рассмотрении первого варианта - студент нуждается в данных - возникает мысль, что студенту нужны данные, например, для вычислений. Информацией во втором варианте может выступать конспект или учебник. В результате их использования студент получает лишь информацию, которая в определенных случаях может перейти в *знания*. Третий вариант звучит наиболее логично.

Информация, в отличие от данных, имеет смысл.

Понятия "*информация*" и "*знания*", с философской точки зрения, являются понятиями более высокого уровня, чем "данные", которое возникло относительно недавно.

Понятие "*информации*" непосредственно связано с сущностью процессов внутри информационной системы, тогда так понятие "*знание*" скорее

ориентировано на качество процессов. Понятие "*знание*" тесно связано с процессом *принятия решений*.

Несмотря на различия, рассмотренные понятия, как уже отмечалось ранее, не являются разрозненными и несвязанными. Они есть часть одного потока: у истока его находятся данные, в процессе передачи которых возникает *информация*, и в результате использования *информации*, при определенных условиях, возникают *знания*.

В лекции уже отмечалось, что в процессе движения вверх *по информационной пирамиде* объемы данных переходят в ценность знаний. Однако большие объемы данных вовсе не означают и, тем более, не гарантируют получение знаний. Существует определенная зависимость ценности полученных знаний от качества и мощности процедур обработки данных. Типичным примером *информации*, которую нельзя превратить в *знание*, является текст на иностранном языке. При отсутствии словаря и переводчика эта *информация* вообще не имеет ценности, она не может перейти в *знание*. При наличии словаря процесс перехода от *информации* к знанию возможен, но длителен и трудоемок. При наличии переводчика *информация* действительно переходит в *знания*.

Таким образом, для получения ценных знаний необходимы качественные процедуры обработки. Процесс перехода от данных к *знаниям* занимает много времени и стоит дорого. Поэтому очевидно, что технология *Data* Mining с ее мощными и разнообразными алгоритмами является инструментом, при помощи которого, продвигаясь вверх *по информационной пирамиде*, мы можем получать действительно качественные и ценные *знания*.