УДК 004.89:002.53

ПРОБЛЕМЫ, ТЕНДЕНЦИИ РАЗВИТИЯ И ПЕРСПЕКТИВНЫЕ НАПРАВЛЕНИЯ ПРИМЕНЕНИЯ ТЕХНОЛОГИЙ DATA MINING

Слесарев Евгений Владимирович, Тесля Валерий Владимирович
ФГБОУ ВПО «Мордовский государственный университет им. Н. П. Огарёва», ООО «Кодер»,
Российская Федерация, г. Саранск
E-mail: slesarevev@gmail.com , valery.tesley@gmail.com , тел. +7 (902) 6686457,
430005, г.Саранск, ул. Богдана Хмельницкого, д. 39, ком. 503.

Аннотация. В статье рассматриваются понятия интеллектуального анализа данных (Data Mining), существующие проблемы данной отрасли знаний и наиболее актуальные направления развития и применения её в практических приложениях.
Ключевые слова: data mining; интеллектуальный анализ данных; обнаружение знаний; масштабируемость; надёжность; распределенные вычисления, облачные среды.

I. ВВЕДЕНИЕ

В последние годы очевиден взрывной характер усложнения всех сфер жизни общества, в особенности экономической, социальной и, более всего, техносферы. Как следствие, непрерывно ускоряющийся рост количества разнообразных данных требует всё более автоматизированной и всё более интеллектуальной их обработки, для нахождения новых, ранее неизвестных знаний, закономерностей и взаимосвязей. Впоследствии найденные знания должны использоваться для выработки управляющих решений, важность которых, в связи с повсеместным усложнением систем, постоянно растёт. Требуется повышение надёжности систем и всё более оперативное реагирование на не всегда вполне очевидные изменения в них. Из-за огромного количества вырабатываемой информации очень малая её часть будет когда-либо увидена и понята человеком. Однако это чаще всего не нужно, достаточно выявить некоторые полезные знания и закономерности с помощью методов Data Mining.

II. DATA MINING – ИНТЕЛЛЕКТУАЛЬНЫЙ АНАЛИЗ ДАННЫХ

Необходимость интеллектуального анализа данных возникла в конце XX века в результате повсеместного распространения информационных технологий, позволяющих детально протоколировать процессы бизнеса и производства [1].

Термин интеллектуальный анализ данных можно понимать двояко. В узком смысле это попытка адекватного русского перевода термина Data Mining (DM), который ввёл в обиход Григорий Пятецкий-Шапиро в 1992 году. Согласно его определению DM (также называемый Knowledge Discovery In Data (KDD) – обнаружение знаний в данных) – это процесс обнаружения в сырых данных ранее неизвестных, нетривиальных, практически полезных, доступных интерпретации знаний, необходимых для принятия решений в различных сферах человеческой деятельности. Дословный перевод – «раскопки (или добыча) данных» [1].

В широком смысле это современная концепция анализа данных, предполагающая, что:

– данные могут быть неточными, неполными (содержать пропуски), противоречивыми, разнородными, косвенными, и при этом иметь гигантские объёмы;

– сами алгоритмы анализа данных могут обладать элементами интеллекта, в частности, способностью обучаться по прецедентам, то есть делать общие выводы на основе частных наблюдений;

– процессы переработки сырых данных в информацию, а информации в знания уже не могут быть выполнены вручную, и требуют автоматизации.

Термин ИАД обозначает не столько конкретную технологию, сколько сам процесс поиска корреляций, тенденций, взаимосвязей и закономерностей посредством различных

математических и статистических алгоритмов: кластеризации, создания субвыборок, регрессионного и корреляционного анализа. Цель этого поиска – представить данные в виде, четко отражающем бизнес-процессы, а также построить модель, при помощи которой можно прогнозировать процессы, критичные для планирования бизнеса [2].

Data Mining может быть использован для решения следующих общих задач:

– распознавание (классификация, диагностика) ситуаций, явлений, объектов или процессов с обоснованием решений;

– прогнозирование ситуаций, явлений, процессов или состояний по выборкам динамических данных;

– кластерный анализ и исследование структуры данных;

– выявление существенных признаков и нахождение простейших описаний;

– нахождение эмпирических закономерностей различного вида;

– построение аналитических описаний множеств (классов) объектов;

– нахождение нестандартных или критических случаев;

– формирование эталонных описаний образов.

Технология DM представляет в сфере принятия решений наибольший интерес [3], поскольку с ее помощью можно провести наиболее глубокий и всесторонний анализ данных и, следовательно, принимать наиболее взвешенные и обоснованные решения.

III. ОТЛИЧИТЕЛЬНЫЕ ОСОБЕННОСТИ И ПРОБЛЕМЫ

Существующие традиционные методы анализа данных (статистические методы) в основном ориентированы на проверку заранее сформулированных гипотез и на «грубый» разведочный анализ, составляющий основу оперативной аналитической обработки данных (OnLine Analytical Processing, OLAP), в то время как одно из основных положений Data Mining, смотрящееся наиболее выигрышно, – поиск неочевидных закономерностей. Инструменты Data Mining могут находить такие закономерности самостоятельно и также самостоятельно строить гипотезы о взаимосвязях. Поскольку именно формулировка гипотезы относительно зависимостей является самой сложной задачей, казавшейся до недавнего времени неразрешимой, преимущество Data Mining по сравнению с другими методами анализа является очевидным. Большинство статистических методов для выявления взаимосвязей в данных используют концепцию усреднения по выборке, приводящую к операциям над несуществующими величинами, тогда как Data Mining оперирует реальными значениями. OLAP больше подходит для понимания ретроспективных данных, Data Mining опирается на ретроспективные данные для получения ответов на вопросы о будущем [3].

Прежде чем использовать технологию Data Mining, необходимо тщательно проанализировать ее проблемы, ограничения и критические вопросы, с ней связанные, а также понять, чего эта технология не может.

Data Mining не может заменить аналитика. Технология не может дать ответы на те вопросы, которые не были заданы. Она не может заменить аналитика, а всего лишь дает ему мощный инструмент для облегчения и улучшения его работы.

Сложность разработки и эксплуатации приложений Data Mining. Поскольку данная технология является мультидисциплинарной областью, для разработки приложения, включающего Data Mining, необходимо задействовать специалистов из разных областей, а также обеспечить их качественное взаимодействие.

Квалификация пользователя. Различные инструменты Data Mining имеют различную степень «дружелюбности» интерфейса и требуют определенной квалификации пользователя. Поэтому программное обеспечение должно соответствовать уровню подготовки пользователя. Использование Data Mining должно быть неразрывно связано с повышением квалификации пользователя.

Извлечение полезных сведений невозможно без хорошего понимания сути данных. Необходим тщательный выбор модели и интерпретация зависимостей или шаблонов,

которые обнаружены. Поэтому работа с такими средствами требует тесного сотрудничества между экспертом в предметной области и специалистом по инструментам Data Mining. Построенные модели должны быть грамотно интегрированы в бизнес-процессы для возможности оценки и обновления моделей. В последнее время системы Data Mining поставляются как часть технологии хранилищ данных.

Сложность подготовки данных. Успешный анализ требует качественной предобработки данных. По утверждению аналитиков и пользователей баз данных, процесс предобработки может занять до 80% процентов всего Data Mining-процесса. Таким образом, чтобы технология работала на себя, потребуется много усилий и времени, которые уходят на предварительный анализ данных, выбор модели и ее корректировку.

Большой процент ложных, недостоверных или бессмысленных результатов. С помощью Data Mining можно отыскивать действительно очень ценную информацию, которая вскоре даст большие дивиденды, в частности в виде финансовой и конкурентной выгоды. Однако Data Mining достаточно часто делает множество ложных и не имеющих смысла открытий. Многие специалисты утверждают, что Data Mining-средства могут выдавать огромное количество статистически недостоверных результатов. Чтобы этого избежать, необходима проверка адекватности полученных моделей на тестовых данных.

Наличие достаточного количества репрезентативных данных. Средства Data Mining, в отличие от статистических, теоретически не требуют наличия строго определенного количества ретроспективных данных. Эта особенность может стать причиной обнаружения недостоверных, ложных моделей и, как результат, принятия на их основе неверных решений. Необходимо осуществлять контроль статистической значимости обнаруженных знаний.

Исследования отмечают, что существуют как успешные решения, использующие Data Mining, так и неудачный опыт применения этой технологии [1]. Области, где применения технологии Data Mining, скорее всего, будут успешными, имеют такие особенности:

- требуют решений, основанных на знаниях;
- имеют изменяющуюся окружающую среду;
- имеют доступные, достаточные и значимые данные;
- обеспечивают высокие дивиденды от правильных решений.

IV. ТЕНДЕНЦИИ РАЗВИТИЯ И ПЕРСПЕКТИВНЫЕ НАПРАВЛЕНИЯ ПРИМЕНЕНИЯ

Разнообразие данных, задач анализа данных и подходов к ним ставит множество сложных вопросов перед исследователями в данной области. Ключевые направления исследований: развитие эффективных методов анализа данных, систем и сервисов, интерактивных и интегрированных сред для интеллектуального анализа данных. Использование методов ИАД для решения сложных прикладных проблем является важной задачей для исследователей и разработчиков DM-систем и приложений. Вот некоторые тенденции развития области, которые отражают пути достижения этих задач.

Исследование областей применения. Приложения интеллектуального анализа данных уже сильно помогают предпринимательской деятельности получать конкурентные преимущества. Область применимости ИАД для бизнеса продолжает расширяться, так как в розничной торговле стали широко применяться электронная коммерция и электронный маркетинг. Data Mining начинает использоваться в других областях, таких как анализ текстов и веб-ресурсов, финансовый анализ, промышленность, государственный сектор, биология, медицина и естественные науки. Новейшие области применения включают в себя интеллектуальный анализ данных для борьбы с терроризмом и мобильный (беспроводной) DM. Поскольку универсальные Data Mining-системы могут иметь ограничения в решении конкретных прикладных задач, очевидна тенденция развития более проблемно-ориентированных систем и инструментов, а также скрытых функций ИАД, встроенных в различные сервисы.

Масштабируемые и интерактивные методы интеллектуального анализа данных. В отличие от традиционных методов анализа данных, DM должен быть в состоянии обрабатывать огромные объемы данных эффективно и, если это возможно, в интерактивном режиме. Поскольку объемы собираемых данных продолжают неумолимо расти, становятся необходимы хорошо масштабируемые алгоритмы. Важным направлением в сторону улучшения общей эффективности процесса анализа данных, при одновременном повышении уровня взаимодействия с пользователем, является constraint-based mining, интеллектуальный анализ данных с комплексной увязкой параметров. Это предоставляет пользователям дополнительные возможности управления, позволяя устанавливать и использовать некоторые ограничения, чтобы руководить системами ИАД в поисках интересных паттернов и знаний.

Интеграция ИАД с системами поиска, баз данных, хранилищ данных и облачных вычислений. Вышеуказанные системы являются господствующими тенденциями в области систем обработки информации и вычислительных систем. Важно гарантировать, чтобы Data Mining служил неотъемлемым компонентом анализа данных, который можно было бы легко интегрировать в такие среды обработки информации. Подсистема интеллектуального анализа данных должна быть реализована в виде легко интегрируемого унифицированного фреймворка или в виде скрытой функция. Это позволит обеспечить доступность данных, переносимость методов ИАД, масштабируемость, высокую производительность и единую среду обработки информации для многомерного анализа и исследований данных.

Интеллектуальный анализ социальных и информационных сетей. Данная задача является критичной, потому что такие сети уже используются повсеместно и продолжают бурно эволюционировать. Остро стоит задача мониторинга сетей и поиска скрытых зависимостей в собранных данных. Развитие масштабируемых и эффективных методов и приложений обнаружения знаний для большого количества сетевых данных имеет очень важное значение. Наиболее активные разработки ведутся в следующих сферах: очистка, интеграция и валидация данных с помощью анализа информационных сетей; контролируемая кластеризация и классификация гомогенных и гетерогенных сетей; выявление ролей, предсказание связей, установление сходств в информационных сетях; установление эволюционных алгоритмов, поиск аномалий.

Интеллектуальный анализ пространственно-временных данных и движущихся объектов. Данное направление быстро развивается из-за широкого использования сотовых телефонов, GPS-устройств, датчиков, сенсоров и прочего оборудования беспроводной связи. Естественно, что существует множество сложных вопросов реализации эффективного и своевременного обнаружения знаний в таких данных, поскольку очевидно широкое внедрение сенсорных сетей в ближайшем будущем во многие аспекты жизнедеятельности человека.

Интеллектуальный анализ биологических и биомедицинских данных. Уникальное сочетание сложности, богатства, размера и значения биологических и биомедицинских данных требует особого внимания. Анализ белковых и ДНК-последовательностей, биологических путей метаболизма и биологических сетей – это только некоторые перспективные области. Другие области биологических применений ИАД включают в себя добычу знаний из гор специализированной литературы, анализ связей в гетерогенных биологических данных, а также интеграция и консолидация биологической информации.

Data Mining, связанный с разработкой программного обеспечения и проектированием систем. Компьютерные программы и большие информационные системы становятся всё более сложными и громоздкими, а также, как правило, всё больше стремятся к архитектуре интеграции множества компонентов, разработанных разными командами. Эта тенденция сделала намного более сложной задачу обеспечения надёжности и отказоустойчивости ПО. Анализ программ, содержащих ошибки, по существу, является

процессом интеллектуального анализа данных. Программы во время своего исполнения генерируют большие объемы данных о своём состоянии и происходящих процессах (логи). Эти данные, сопровождающие работу программ, необходимо всесторонне анализировать и отслеживать, проводить тесты – это может открыть важные скрытые паттерны и аномалии, обнаружение которых обеспечит повышение надёжности и отказоустойчивости, и, возможно, приблизит полностью автоматическое обнаружение ошибок (багов) в ПО. Дальнейшее развитие методологии интеллектуального анализа данных для отладки ПО и информационных систем безусловно необходимо для улучшения их эксплуатационных характеристик.

Визуальный и звуковой анализ данных. Данная сфера представляется эффективным способом интеграции с визуальными и слуховыми системами людей для обнаружения знаний в огромных объемах данных. Планомерное развитие подобных технологий будет способствовать участию человека в повышении эффективности анализа данных.

Распределенный ИАД, и интеллектуальный анализ потоков данных в реальном времени. Традиционные методы ИАД, предназначенные для работы в централизованном размещении, плохо работают в существующих на данный момент распределенных вычислительных средах (интернет, локальные сети, высокоскоростные беспроводные сети, сенсорные сети и облачные среды). Ожидаются серьёзные подвижки в области распределенных методов Data Mining. Кроме того, многие области приложений, использующие потоки данных (электронная коммерция, анализ веб-ресурсов, биржевой анализ, системы обнаружения и предотвращения вторжений, мобильный Data Mining), требуют, чтобы модели анализа данных строились в реальном времени. В этом направлении также необходимы дополнительные исследования.

V. ВЫВОДЫ И ЗАКЛЮЧЕНИЕ

Непрерывно ускоряющийся технический прогресс, вырабатывая огромное количество данных дал сильный толчок развитию сферы интеллектуального анализа данных (Data Mining). Несмотря на некоторые проблемы использования, данные технологии невероятно активно развиваются и постепенно входят в различные сферы повседневной жизни. Наиболее значительными тенденциями развития являются: масштабируемые и интерактивные, распределенные методы ИАД, интеграция DM-технологий с существующими информационными системами. Очень актуальны приложения Data Mining в следующих областях: системы поиска, облачные вычисления, социальные и информационные сети, биология и медицина, разработка ПО, мобильные и беспроводные технологии. В статье приведено краткое описание наиболее актуальных проблем, тенденций и приложений технологий интеллектуального анализа данных.

VI. ЛИТЕРАТУРА

1. Frawley W., Piatetsky-Shapiro G., Matheus C. Knowledge Discovery in Databases: An Overview. – AI Magazine. – 1992. – p. 213-228.

2. Макарычев П. П., Афонин А. Ю. Оперативный и интеллектуальный анализ данных: учеб. пособие. – Пенза : Изд-во ПГУ, 2010. – 156 с.

3. Чубукова И. А. Data Mining. Учебное пособие. – М.: Интернет-университет информационных технологий - ИНТУИТ.ру, БИНОМ. Лаборатория знаний, 2008. – 384 с.

4. Барсегян А. А., Куприянов М. С., Степаненко В. В. Технологии анализа данных: Data Mining, Visual Mining, Text Mining, OLAP. – 2-е изд., перераб. и доп. – СПб.: БХВ-Петербург, 2007. – 384 с.

5. Han J., Kamber M., Pei J. Data mining: concepts and techniques. – 3rd ed. – Morgan Kaufmann / Elsevier, 2012. – 744 p.

CHALLENGES, TRENDS AND PROMISING APPLICATIONS OF DATA MINING TECHNOLOGIES

Evgeny V. Slesarev, Valery V. Tesley
N. P. Ogarev's Mordovian State University, Coder OOO
Russian Federation, Saransk city
E-mail: slesarevev@gmail.com , valery.tesley@gmail.com , tel +7 (902) 6686457
39, B. Hmelnitskogo str., room 503, 430005, Saransk, Mordovia, Russia

Annotation. The article describes data mining as a knowledge discovery technology and points its challenges, emerging trends and promising applications.
Keywords: data mining; knowledge discovery; scalability; robustness; distributed computing, cloud computing.