Знания обуславливаются интеллектом

Интеллектуальный анализ данных как важнейший инструмент формирования интеллектуального капитала организаций

Аннотация

Рассматриваются возможности использования данных, накопленных на предприятии, для увеличения его интеллектуального капитала (ИК) за счет поиска в них полезных нетривиальных знаний, а также разработки соответствующих алгоритмов и методов. Кратко анализируются существующие подходы к поиску знаний в базах данных (БД). Показывается перспективность применения для этих целей технологий интеллектуального анализа данных (ИАД), комбинирующих методы прикладной статистики и эвристического поиска на основе самообучающихся моделей для поиска знаний в больших БД.

Ключевые слова: интеллектуальный капитал, знания, базы данных, интеллектуальный анализ данных, прикладная статистика, экспертные системы, эвристический поиск, машинное обучение, аналитическая платформа

Орешков Вячеслав Игоревич
соискатель, учетно-финансовый факультет, Рязанский государственный агротехнологический университет им. П.А. Костычева; инженер, кафедра радиоуправления и связи, Рязанский государственный радиотехнический университет
vyacheslav.oreshkov@yandex.ru

В постиндустриальном обществе возрастает роль интеллектуальных и информационных ресурсов. Знания и информация становятся критически важными элементами экономических систем, поскольку экономические и бизнес-процессы всегда основаны на определенном уровне знаний и обработке соответствующей информации. Главным источником конкурентных преимуществ стало обладание технологиями генерирования знаний, лежащих в основе формирования интеллектуального капитала (ИК) организаций [1].

Капитал прирастает знаниями

ИК образован активами двух видов: человеческими – знаниями и опытом людей, и нематериальными – программным обеспечением, базами данных (БД), патентами, публикациями, научными методами, алгоритмами и т.д. К человеческим активам в первую очередь относят знания. И это неслучайно.

Действительно, с общенаучной точки зрения знания есть «... адекватное отражение действительности в сознании человека в виде представлений, выводов, понятий, суждений, правил». Если предметной областью является экономика, то такой «действительностью» оказываются экономические и бизнес-процессы, явления, которые, в свою очередь, описываются различными данными. Следовательно, реализовав процесс преобразования данных в знания, можно обеспечить приращение ИК организации как за счет увеличения индивидуальных знаний людей, так и за счет нематериальных активов (методов и алгоритмов, реализующих поиск знаний, хранилищ данных, соответствующего программного обеспечения (ПО).

Обнаруженные в БД знания можно разделить на *тривиальные* и *нетривиальные*. К первым относятся те, которые можно получить путем простого визуального анализа данных либо путем их преобразования с помощью математических расчетов. Нетривиальные знания, называемые также *глубинными* или *скрытыми*, требуют более сложного подхода к обработке данных, использования алгоритмов поиска неочевидных закономерностей и связей между элементами данных, интерпретация которых человеком позволяет ему генерировать новые знания. Очевидно, что нетривиальные знания, а также методы и алгоритмы их поиска, являются наиболее ценными активами ИК, поэтому в обладании ими заинтересована любая организация.

Поиск ведется по направлениям

Большинство предприятий ведут сбор информации, отражающей различные аспекты деятельности, а в корпоративных БД накапливаются огромные массивы данных, которые могут служить источником новых нетривиальных знаний. Однако их поиск сопряжен с рядом проблем технического и методологического плана, основными из которых являются [2]:

- разнообразие типов и форматов данных требует решения сложной задачи организации доступа к ним;
- данные обычно содержат факторы, снижающие их качество – шумы, дубликаты, противоречия, пропуски, аномальные значения и т.д., что требует их очистки;
- представление данных не всегда соответствует методам преобразований, выполняемых в процес-

к человеческим активам в первую очередь относят знания

се поиска знаний, поэтому необходима процедура их предобработки (изменение форматов и типов данных, квантование, кодирование, группировка, сокращение размерности, обогащение и т.д.);

• собственно обработка данных с целью обнаружения в них скрытых закономерностей, зависимостей и структур, построение моделей;

• визуализация – представление обнаруженных особенностей в данных в наиболее удобной для восприятия и интерпретации форме;

• тиражирование знаний – процесс передачи найденных знаний всем заинтересованным потребителям.

В информационных технологиях можно выделить три основных направления, в которых реализуется поиск знаний в БД.

1. Прикладная статистика [3] имеет хорошо разработанную теорию, ее методы и алгоритмы анализа являются математически корректными и обоснованными во всем их разнообразии. На рынке представлено большое количество программных продуктов; статистические пакеты Statistica, SPSS, PASW, Minitab, STADIA, Statgraphics и др. При этом они являются весьма сложными, требуют математической подготовки, имеют низкую интерпретируемость результатов, плохо отражают причинно-следственные связи. Кроме того, статистические пакеты ориентированы на работу с относительно небольшими, локальными БД и не имеют развитых средств их очистки и предобработки.

2. Экспертные системы (системы, основанные на знаниях) [4] используют формализованные знания экспертов, помещенные в базу знаний, откуда они извлекаются с помощью средств логического вывода. Преимуществом таких систем является диалог на естественном языке, а имитация рассуждений человека, реализуемая с помощью интеллектуального интерфейса, обеспечивает высокий уровень понимания как поставленной задачи, так и результатов. К недостаткам можно отнести высокую трудоемкость разработки, узкоспециальную направленность, слабые возможности актуализации. Но главное – пользователь фактически работает не с данными, а с мнением экспертов, принимавших участие в формировании базы знаний, т.е. задействует ранее открытые знания, что резко снижает их ценность.

3. Интеллектуальный анализ данных (Data Mining – DM). Сравнительно новое направление, возникшее

на стыке теории БД, искусственного интеллекта, машинного обучения и прикладной статистики. Сам термин *data mining* дословно означает «раскопка данных», «разработка данных», впервые был введен В.Г. Пятецким-Шапиро как «совокупность методов обнаружения в данных ранее неизвестных, нетривиальных, практически полезных и доступных интерпретации знаний, необходимых для принятия решений в различных сферах человеческой деятельности» [5].

В основе технологии DM лежит использование компьютерных моделей, решающих основные задачи анализа данных – классификацию, численное предсказание, кластеризацию, ассоциацию, прогнозирование [2]. Математический аппарат DM включает как статистические методы (регрессию, корреляционный анализ, байесовскую классификацию и др. [2]), так и самообучающиеся алгоритмы (нейронные сети, деревья решений, карты Кохонена, ассоциативные правила, нечеткую логику и др. [2]). Именно последние считаются «ядром» DM, поскольку способны автономно, с минимальным вмешательством человека, извлекать скрытые (которые нельзя определить визуально или рассчитать) закономерности в данных, интерпретация которых позволяет генерировать нетривиальные знания.

Для комплексного решения задач DM в последнее десятилетие появился специальный класс ПО – аналитические платформы (АП). Популярность данного направления оказалась настолько велика, что в разработке АП «отметилось» большинство крупнейших компаний на рынке информационных технологий – Microsoft, IBM, Oracle, SAS Institute, Silicon Graphics, StatSoft, и множество менее известных (Angross Software, Neuro Solution и др.) [6]. Значительное количество свободных продуктов разрабатывается и поддерживается крупными университетами (Weka, RapidMiner, Orange). Ведущими отечественными разработками являются PolyAnalyst (Megaputer) и Deductor (ООО «Аналитические технологии»).

нетривиальные знания, а также методы и алгоритмы их поиска являются наиболее ценными активами ИК, поэтому в обладании ими заинтересована любая организация

Симбиоз формирует образ

Аппарат DM сформировался на основе двух направлений – прикладной статистики и машинного обучения (МО) [2]. Чтобы эффективно применять соответствующие методы и правильно интерпретировать

результаты, нужно представлять их ограничения и особенности. Статистические методы и алгоритмы являются корректными математически (для них сформулированы критерии точности и оптимальности, доказаны соответствующие теоремы), но не всегда корректно отражают закономерности реальных процессов и явлений, сложны для понимания и интерпретации, требуют значительной априорной информации об исследуемых процессах и явлениях.

Методы МО, напротив, позволяют строить модели зависимостей в данных с минимальным вмешательством пользователя, их результаты наглядны и хорошо интерпретируемы. Но большинство алгоритмов МО являются эвристическими, т.е. корректность их для всех возможных случаев не доказана, но известно, что они дают приемлемый результат в большинстве практически значимых случаев. Эвристические алгоритмы широко применяются в задачах, для которых отсутствует общая постановка и решение, что делает их незаменимыми при работе с большими БД, содержащими скрытые закономерности и структуры.

Использование статистических методов совместно с алгоритмами эвристического поиска позволяет перейти к формированию и восприятию сложных образов, описывающих особенности исследуемых данных. Они могут включать в себя многомерные диаграммы и графы, карты, правила различных типов, формируемые на естественном языке, иерархические древовидные структуры и т.д. Это позволяет перенести поиск знаний на когнитивный уровень, когда пользователь распознает «ситуацию» как своего рода образ, что более соответствует механизму восприятия человеком окружающей действительности, чем обычные таблицы, графики или статистические показатели.

Актуальность внедрения технологий интеллектуального анализа данных в процесс управления знаниями обусловлена тем, что все больше отраслей народного хозяйства становятся отраслями, основанными на знаниях. Типичным примером может служить современный агропромышленный комплекс, где предприятия не только производят сельхозпродукцию, но и накапливают собственные знания и опыт, направленные на разработку стратегий увеличения выхода продукции и минимизации издержек, продвижения своей продукции на рынке и повышения конкурентоспособности [6].

Литература

1. Брукинг Э. Интеллектуальный капитал [Текст]. – СПб: Питер, 2001.
2. Паклин Н.Б., Орешков В.И. Бизнес-аналитика: от данных к знаниям (+CD) [Текст]. – СПб.: Питер, 2010.
3. Орлов А. И. Прикладная статистика: учебник для вузов [Текст].— М.: Экзамен, 2006. – 672 с.
4. Джарратано Дж., Райли Г. Экспертные системы: принципы разработки и программирование [Текст] / Пер. с англ. – М.: Вильямс, 2006.

Вывод

Таким образом, важнейшим направлением формирования интеллектуального капитала организации является поиск новых нетривиальных знаний в базах данных. Наиболее перспективной технологией такого поиска является интеллектуальный анализ данных, в рамках которого могут быть реализованы все этапы трансформации данных в знания. Главной составляющей эффективности DM является стратегия интеллектуального поиска по нескольким альтернативным направлениям – численного предсказания, классификации, кластеризации, ассоциации и др., которая обеспечивает качественный переход на когнитивный уровень восприятия сложных явлений и процессов, способствуя обнаружению глубинных знаний и росту ИК предприятий и организаций.

КЭ

Литература

5. Advances in Knowledge Discovery and Data Mining. Edited by Usama M. Fayyad, Gregory Piatetsky-Shapiro. MIT Press, 1996.

6. Васильев Е.П. Современные аналитические платформы для задач АПК [Текст] / Е.П. Васильев, В.И. Орешков // Вестник Рязанского гос. агротехнологического университета имени П.А. Костычева. – Рязань: РГАТУ. 2011 г. Вып. 1.

Vyacheslav I. Oreshkov
Applicant, Accounting Financial Faculty, Kostychev Ryazan State Agro-Technological University, Engineer, Chair of Radiocontrol and Communication, Ryazan State Radio-Technical University

Intelligent Data Analysis as Most Important Implement of Company Intellectual Capital Development

Abstract

Article considers opportunities of data utilization, accumulated at enterprise, for growth of its intellectual capital (IC) by means of searching of their useful substantial knowledge, and by development of relevant algorithms and methods as well. Article analyses briefly existed approaches to searching of knowledge in data base (DB). Article presents perceptiveness of utilization of intellectual data analysis technologies (IAA) for these purposes, which combines practical statistics methods and heuristic searching based on self-correcting models for searching of knowledge in large DB.

Keywords: intellectual capital, knowledge, data base, intellectual data analysis, practical statistics, expert systems, heuristic searching, computer-based education, analytic platform