УДК 004.6

Е.В. Малащук, Д.В. Бабин, С.М. Вороной, М.Г. Кочеткова
Донецкий государственный институт искусственного интеллекта, г. Донецк, Украина

Обзор существующих алгоритмов Data Mining для глубинного анализа текстов и методов извлечения знаний

В данной статье рассматриваются современные подходы к получению новых знаний на основе анализа информационного пространства корпоративных сетей и сети Internet и к методам обработки информационных потоков с целью извлечения знаний по технологии Data Mining в отрасли Text Mining. Ставится задача исследования методологий в сфере применения Text Mining, выделяется специфика современных требований к эффективной интеллектуальной переработке данных. Анализируются проблемы, которые неудовлетворительно решаются существующими методами предварительной обработки и доступа к большим объемам информации, рассматриваются особенности различных информационно-поисковых систем и средств извлечения знаний. Большое внимание уделено новому направлению обработки текстовой информации – «глубинному анализу текстов» (Text Mining), объединяющему в себе технологические и методологические подходы контент-анализа, компьютерной лингвистики и искусственного интеллекта.

Введение

Электронная информация играет все большую роль во всех сферах жизни современного общества. В информационных хранилищах, распределенных по всему миру, собраны миллионы терабайт текстовых данных [1]. Однако развитие ресурсов Интернета многократно усугубило проблему информационной перегрузки. Сырые неструктурированные данные составляют не менее 90 % информации, с которой имеют дело пользователи. Очевидно, что найти полезную информацию среди нескольких экзабайт электронных текстов непросто. По этой причине даже весьма ценные для определенных пользователей документы [2] в течение длительного времени остаются невостребованными, поскольку отыскать их очень часто не представляется возможным.

Но вместе с ростом всемирной сети увеличивается и мощность средств поиска в ней необходимой информации. Причем возрастающая эффективность информационно-поисковых систем (ИПС) связана не только с ростом производительности их аппаратной части и увеличением пропускной способности каналов связи, но и с внедрением новых алгоритмов поиска и упорядочивания данных, созданием схем запроса, призванных максимально точно определить информационные потребности каждого конкретного пользователя, внедрением новых поисковых сервисов [3].

Понятно, что без продуктивной переработки потоки сырых данных образуют никому не нужный информационный шум. В связи с этим возникло направление Text Mining, включающее в свой состав технологии Data Mining, что переводится как «добыча» или «раскопка данных». Нередко рядом с Data Mining встречаются слова

«обнаружение знаний в базах данных» (knowledge discovery in databases) и «интеллектуальный анализ данных» [4]. Их можно считать синонимами Data Mining. Возникновение всех указанных терминов связано с новым витком в развитии средств и методов обработки данных.

Постановка задачи

Сфера применения Data Mining ничем не ограничена – она везде, где имеются какие-либо данные. Выделяют пять стандартных типов закономерностей, которые позволяют выявлять методы Data Mining: ассоциация, последовательность, классификация, кластеризация и прогнозирование:

- ассоциация имеет место в том случае, если несколько событий связаны друг с другом;
- если существует цепочка связанных во времени событий, то говорят о последовательности;
- с помощью классификации выявляются признаки, характеризующие группу, к которой принадлежит тот или иной объект. Это делается посредством анализа уже классифицированных объектов и формулирования некоторого набора правил;
- кластеризация отличается от классификации тем, что сами группы заранее не заданы. С помощью кластеризации средства Data Mining самостоятельно выделяют различные однородные группы данных;
- основой для всевозможных систем прогнозирования служит историческая информация, хранящаяся в БД в виде временных рядов. Если удается построить шаблоны, адекватно отражающие динамику поведения целевых показателей, есть вероятность, что с их помощью можно предсказать и поведение системы в будущем.

Традиционная математическая статистика, долгое время претендовавшая на роль основного инструмента анализа данных, откровенно спасовала перед лицом возникших проблем. Главная причина – концепция усреднения по выборке, приводящая к операциям над фиктивными величинами [5]. Методы математической статистики оказались полезными главным образом для проверки заранее сформулированных гипотез (verification-driven data mining) и для «грубого» разведочного анализа, составляющего основу оперативной аналитической обработки данных (online analytical processing, OLAP) [6].

В основу современной технологии Data Mining (discovery-driven data mining) положена концепция шаблонов (паттернов) [4], отражающих фрагменты многоаспектных взаимоотношений в данных. Эти шаблоны представляют собой закономерности, свойственные подвыборкам данных, которые могут быть компактно выражены в понятной человеку форме. Поиск шаблонов производится методами, не ограниченными рамками априорных предположений о структуре выборки и виде распределений значений анализируемых показателей. Примеры заданий на такой поиск при использовании Data Mining приведены в табл. 1.

Можно привести еще много примеров различных областей знания, где решение задач методами Data Mining играет или будет играть ведущую роль. Особенность этих областей заключается в их сложной системной организации. Они относятся главным образом к надкибернетическому уровню организации систем [6], закономерности которого не могут быть достаточно точно описаны на языке статистических или иных аналитических математических моделей [7]. Данные в указанных областях неоднородны, гетерогенны, нестационарны и часто отличаются высокой размерностью.

Таблица 1 – Примеры формулировок задач по методикам OLAP и Data Mining

OLAP	Data Mining
Каковы средние показатели травматизма для курящих и некурящих?	Какие факторы лучше всего предсказывают несчастные случаи?
Каковы средние размеры телефонных счетов существующих клиентов в сравнении со счетами бывших клиентов (отказавшихся от услуг телефонной компании)?	Какие характеристики отличают клиентов, которые, по всей вероятности, собираются отказаться от услуг телефонной компании?
Какова средняя величина ежедневных покупок по украденной и не украденной кредитной карточке?	Какие схемы покупок характерны для мошенничества с кредитными карточками?

Важное положение Data Mining – нетривиальность разыскиваемых шаблонов. Это означает, что найденные шаблоны должны отражать неочевидные, неожиданные (unexpected) регулярности в данных, составляющие так называемые скрытые знания (hidden knowledge). К обществу пришло понимание, что сырые данные (raw data) содержат глубинный пласт знаний, при грамотной раскопке которого могут быть обнаружены настоящие самородки (рис. 1).

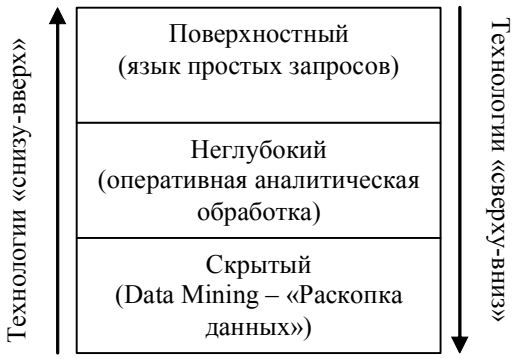

Рисунок 1 – Уровни извлекаемых из данных знаний

Специфика современных требований к эффективной интеллектуальной переработке данных следующая:
- данные имеют неограниченный объем;
- данные являются разнородными (количественными, качественными, текстовыми);
- результаты должны быть конкретны и понятны;
- инструменты для обработки сырых данных должны быть просты в использовании.

Классы систем Data Mining

Data Mining является мультидисциплинарной областью [8], возникшей и развивающейся на базе достижений прикладной статистики, распознавания образов, методов искусственного интеллекта, теории баз данных и др. (рис. 2). Отсюда обилие методов и алгоритмов, реализованных в различных действующих системах Data Mining. Многие из таких систем интегрируют в себе сразу несколько подходов.

Тем не менее, как правило, в каждой системе имеется какой-то ключевой компонент, на который делается главная ставка.

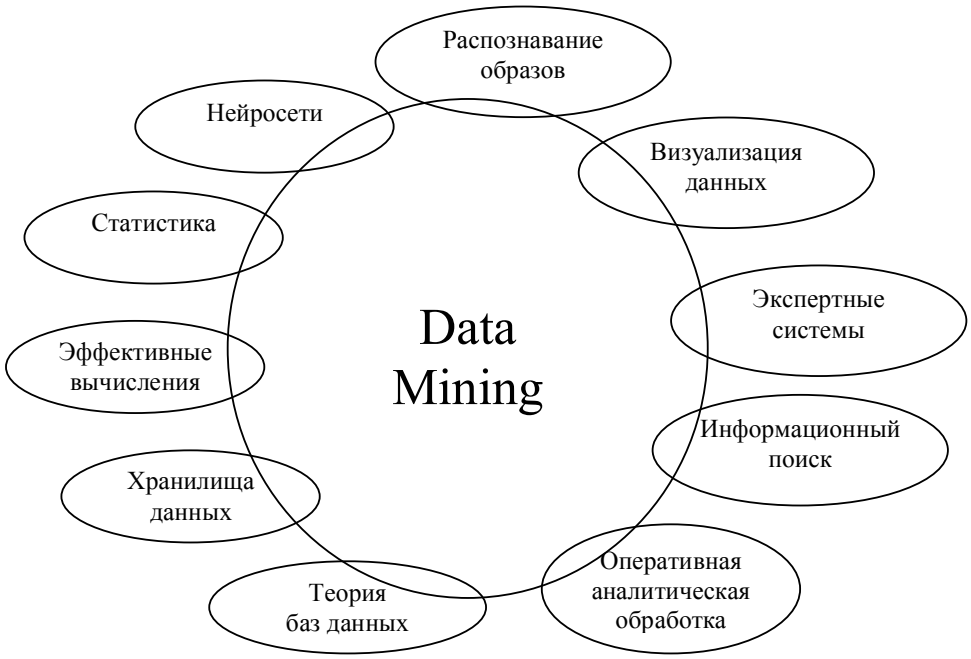

Рисунок 2 – Data Mining как мультидисциплинарная область

Ниже приводится классификация указанных ключевых компонентов на основе работы [5]. Выделенным классам дается краткая характеристика.

Предметно-ориентированные аналитические системы

Предметно-ориентированные аналитические системы очень разнообразны. Наиболее широкий подкласс таких систем, получивший распространение в области исследования финансовых рынков, носит название «технический анализ». Он представляет собой совокупность нескольких десятков методов прогноза динамики цен и выбора оптимальной структуры инвестиционного портфеля, основанных на различных эмпирических моделях динамики рынка [9]. Эти методы часто используют несложный статистический аппарат, но максимально учитывают сложившуюся в своей области специфику (профессиональный язык, системы различных индексов и пр.).

Статистические пакеты

Последние версии почти всех известных статистических пакетов включают наряду с традиционными статистическими методами также элементы Data Mining. Но основное внимание в них уделяется все же классическим методикам – корреляционному, регрессионному, факторному анализу и другим [10]. Недостатком систем этого класса считают требование к специальной подготовке пользователя. Также отмечают, что мощные современные статистические пакеты являются слишком «тяжеловесными» для массового применения в финансах и бизнесе.

Есть еще более серьезный принципиальный недостаток статистических пакетов, ограничивающий их применение в Data Mining. Большинство методов, входящих в состав пакетов, опираются на статистическую парадигму, в которой главными фигурантами служат усредненные характеристики выборки. А эти характеристики, как указывалось выше, при исследовании реальных сложных жизненных феноменов часто являются фиктивными величинами.

Нейронные сети

Это большой класс систем, архитектура которых имеет аналогию (как теперь известно, довольно слабую) с построением нервной ткани из нейронов [11]. В одной из наиболее распространенных архитектур, многослойном персептроне с обратным распространением ошибки [12], имитируется работа нейронов в составе иерархической сети, где каждый нейрон более высокого уровня соединен своими входами с выходами нейронов нижележащего слоя. На нейроны самого нижнего слоя подаются значения входных параметров, на основе которых нужно принимать какие-то решения, прогнозировать развитие ситуации и т.д. Эти значения рассматриваются как сигналы, передающиеся в следующий слой, ослабляясь или усиливаясь в зависимости от числовых значений (весов), приписываемых межнейронным связям. В результате на выходе нейрона самого верхнего слоя вырабатывается некоторое значение, которое рассматривается как ответ – реакция всей сети на введенные значения входных параметров. Для того чтобы сеть можно было применять в дальнейшем, ее прежде надо «натренировать» на полученных ранее данных, для которых известны и значения входных параметров, и правильные ответы. Тренировка состоит в подборе весов межнейронных связей, обеспечивающих наибольшую близость ответов сети к известным правильным ответам [13].

Основным недостатком нейросетевой парадигмы является необходимость иметь очень большой объем обучающей выборки. Другой существенный недостаток заключается в том, что даже натренированная нейронная сеть представляет собой черный ящик. Знания, зафиксированные как веса нескольких сотен межнейронных связей, совершенно не поддаются анализу и интерпретации человеком [12] (известные попытки дать интерпретацию структуре настроенной нейросети выглядят неубедительными).

Системы рассуждений на основе аналогичных случаев

Идея систем case based reasoning (CBR) на первый взгляд крайне проста. Для того чтобы сделать прогноз на будущее или выбрать правильное решение, эти системы находят в прошлом близкие аналоги наличной ситуации и выбирают тот же ответ, который был для них правильным. Поэтому этот метод еще называют методом «ближайшего соседа» (nearest neighbour). В последнее время распространение получил также термин memory based reasoning, который акцентирует внимание, что решение принимается на основании всей информации, накопленной в памяти [14].

Системы CBR показывают неплохие результаты в самых разнообразных задачах. Главным их минусом считают то, что они вообще не создают каких-либо моделей или правил, обобщающих предыдущий опыт: в выборе решения они основываются на всем массиве доступных исторических данных, поэтому невозможно сказать, на основе каких конкретно факторов CBR-системы строят свои ответы.

Другой минус заключается в произволе, который допускают системы CBR при выборе меры «близости». От этой меры самым решительным образом зависит объем множества прецедентов, которые нужно хранить в памяти для достижения удовлетворительной классификации или прогноза [15].

Деревья решений (decision trees)

Деревья решения являются одним из наиболее популярных подходов к решению задач Data Mining. Они создают иерархическую структуру классифицирующих правил типа «ЕСЛИ... ТО...» (if – then), имеющую вид дерева [16]. Для принятия решения, к какому классу отнести некоторый объект или ситуацию, требуется ответить на вопросы, стоящие в узлах этого дерева, начиная с его корня. Вопросы имеют вид «значение параметра А больше х?». Если ответ положительный, осуществляется переход к правому узлу следующего уровня, если отрицательный – то к левому узлу; затем снова следует вопрос, связанный с соответствующим узлом.

Популярность подхода связана как бы с наглядностью и понятностью. Но деревья решений принципиально не способны находить лучшие (наиболее полные и точные) правила в данных. Они реализуют наивный принцип последовательного просмотра признаков и «цепляют» фактически осколки настоящих закономерностей, создавая лишь иллюзию логического вывода [17].

Эволюционное программирование

В данной системе гипотезы о виде зависимости целевой переменной от других переменных формулируются в виде программ на некотором внутреннем языке программирования [18]. Процесс построения программ строится как эволюция в мире программ (этим подход немного похож на генетические алгоритмы). Когда система находит программу, более или менее удовлетворительно выражающую искомую зависимость, она начинает вносить в нее небольшие модификации и отбирает среди построенных дочерних программ те, которые повышают точность. Таким образом система «выращивает» несколько генетических линий программ, которые конкурируют между собой в точности выражения искомой зависимости.

Другое направление эволюционного программирования связано с поиском зависимости целевых переменных от остальных в форме функций какого-то определенного вида [7]. Например, в одном из наиболее удачных алгоритмов этого типа – методе группового учета аргументов (МГУА) – зависимость ищут в форме полиномов.

Генетические алгоритмы

Data Mining не основная область применения генетических алгоритмов. Их нужно рассматривать скорее как мощное средство решения разнообразных комбинаторных задач и задач оптимизации. Тем не менее генетические алгоритмы вошли сейчас в стандартный инструментарий методов Data Mining, поэтому они и включены в данный обзор.

Первый шаг при построении генетических алгоритмов – это кодировка в базе данных исходных логических закономерностей, которые именуют хромосомами, а весь набор таких закономерностей называют популяцией хромосом. Далее для

реализации концепции отбора вводится способ сопоставления различных хромосом. Популяция обрабатывается с помощью процедур репродукции, изменчивости (мутаций), генетической композиции. Эти процедуры имитируют биологические процессы. Наиболее важные среди них: случайные мутации данных в индивидуальных хромосомах, переходы (кроссинговер) и рекомбинация генетического материала, содержащегося в индивидуальных родительских хромосомах (аналогично гетеросексуальной репродукции) и миграции генов [19]. В ходе работы процедур на каждой стадии эволюции получаются популяции со все более совершенными индивидуумами.

Генетические алгоритмы удобны тем, что их легко распараллеливать. Например, можно разбить поколение на несколько групп и работать с каждой из них независимо, обмениваясь время от времени несколькими хромосомами. Существуют также и другие методы распараллеливания генетических алгоритмов.

Генетические алгоритмы имеют ряд недостатков. Критерий отбора хромосом и используемые процедуры являются эвристическими и вовсе не гарантируют нахождения лучшего решения. Как и в реальной жизни, эволюцию может «заклинить» на какой-либо непродуктивной ветви. И наоборот, можно привести примеры, как два неперспективных родителя, которые будут исключены из эволюции генетическим алгоритмом, оказываются способными произвести высокоэффективного потомка. Это особенно становится заметно при решении задач высокой размерности со сложными внутренними связями.

Алгоритмы ограниченного перебора

Алгоритмы ограниченного перебора были предложены в середине 60-х годов М.М. Бонгардом для поиска логических закономерностей в данных [20]. С тех пор они продемонстрировали свою эффективность при решении множества задач из самых различных областей.

Эти алгоритмы вычисляют частоты комбинаций простых логических событий в подгруппах данных. Примеры простых логических событий: $X = a$; $X < a$; $X > a$; $a < X < b$ и др., где X – какой-либо параметр, a и b – константы. Ограничением служит длина комбинации простых логических событий (у М. Бонгарда она была равна 3). На основании анализа вычисленных частот делается заключение о полезности той или иной комбинации для установления ассоциации в данных, для классификации, прогнозирования и т.п.

Но проблемы заключаются в том, что алгоритмы ограниченного перебора не могут обнаруживать все логические правила в данных и с самого начала работы алгоритма производится эвристический поиск простых логических событий, на которых потом строится весь дальнейший анализ. Другой момент – такие системы выдают решение за приемлемое время только для сравнительно небольшой размерности данных.

Системы для визуализации многомерных данных

В той или иной мере средства для графического отображения данных поддерживаются всеми системами Data Mining. Вместе с тем весьма внушительную долю составлют системы, специализирующиеся исключительно на этой функции [21].

В подобных системах основное внимание сконцентрировано на дружественности пользовательского интерфейса, позволяющего ассоциировать с анализируемыми показателями различные параметры диаграммы рассеивания объектов (записей) базы данных. К таким параметрам относятся цвет, форма, ориентация относительно собственной оси, размеры и другие свойства графических элементов изображения. Кроме того, системы визуализации данных снабжены удобными средствами для масштабирования и вращения изображений.

Выводы

Системы Data Mining экспоненциально развиваются. В этом развитии принимают участие практически все крупнейшие корпорации[1]. В частности, Microsoft [22] непосредственно руководит большим сектором данного рынка (издает специальный журнал, проводит конференции, разрабатывает собственные продукты).

Системы Data Mining применяются по двум основным направлениям [23]:

– массовый продукт для бизнес-приложений;

– как инструменты для проведения уникальных исследований (генетика, химия, медицина и пр.).

Количество инсталляций массовых продуктов, судя по имеющимся сведениям, сегодня достигает десятков тысяч. Лидеры Data Mining связывают будущее этих систем с использованием их в качестве интеллектуальных приложений, встроенных в корпоративные хранилища данных [24].

Несмотря на обилие методов Data Mining, приоритет постепенно все более смещается в сторону логических алгоритмов поиска в данных if-then правил (Text Mining). С их помощью решаются задачи прогнозирования, классификации, распознавания образов, сегментации баз данных, извлечения из данных «скрытых» знаний, интерпретации данных, установления ассоциаций в БД и др. Результаты таких алгоритмов эффективны и легко интерпретируются.

Вместе с тем главной проблемой логических методов обнаружения закономерностей является проблема перебора вариантов за приемлемое время. Известные методы либо искусственно ограничивают такой перебор [20] (алгоритмы KOPA, WizWhy), либо строят деревья решений (алгоритмы CART, CHAID, ID3, See5, Sipina и др.), имеющих принципиальные ограничения эффективности поиска if-then правил. Другие проблемы связаны с тем, что известные методы поиска логических правил не поддерживают функцию обобщения найденных правил и функцию поиска оптимальной композиции таких правил. Удачное решение указанных проблем может составить предмет новых конкурентоспособных разработок.

Литература

1. Базы данных. Интеллектуальная обработка информации / В.В. Корнеев, А.Ф. Гареев, С.В. Васютин, В.В. Райх. – М.: Нолидж, 2001 – 653 с.
2. Спирли Э. Корпоративные хранилища данных. Планирование, разработка, реализация Пер. с англ. – М.: Вильямс, 2001 – Т. 1. – 521 с.
3. Салливан. Данных – больше, доступ – лучше // Computerworld Россия. – 2001. – № 38. – С. 54-95.

[1] http://www.kdnuggets.com

4. Knowledge Discovery Through Data Mining: What Is Knowledge Discovery? – USA, IL: Tandem Computers Inc., 1996. – 785 p.
5. Киселев М., Соломатин Е. Средства добычи знаний в бизнесе и финансах // Открытые системы. – 1997. – № 4. – С. 41-44.
6. Boulding K.E. General Systems Theory – The Skeleton of Science // Management Science. – 1956. – № 2. – С. 87-115.
7. Гик Дж. Прикладная общая теория систем. – М.: Мир, 1981. – 543 с.
8. Кречетов Н. Продукты для интеллектуального анализа данных // Рынок программных средств. – 1997. – № 15. – С. 32-39.
9. Дрезнер Х., Хостманн Б., Байтендийк Ф. Вниманию руководства: Обновленные Волшебные Квадраты Gartner для систем интеллектуальной поддержки бизнеса. – М.: Inside Gartner Group, 2003. – 689 с.
10. Прикладная статистика: Классификация и снижение размерности / С.А. Айвазян, В.М. Бухштабер, И.С. Юнюков, Л.Д. Мешалкин. – М.: Финансы и статистика, 1989. – 483 с.
11. Artificial Neural Networks: Concepts and Theory // IEEE Computer Society Press. – 1992. – № 8. – С. 23-39.
12. Короткий С. Нейронные сети: алгоритм обратного распространения. – СПб.: Питер, 2000. – 435 с.
13. Короткий С. Нейронные сети: обучение без учителя. – СПб.: Питер, 2001. – 545 с.
14. Kimbal R. The Data Warehouse Toolkit: Practical Techniques for Building Dimensional Data Warehouses. – John Willey&Sons, 1996. – 793 p.
15. Дюк В.А. Обработка данных на ПК в примерах. – СПб.: Питер, 1997. – 235 с.
16. Самойленко А. Data mining: учебный курс. – СПб: Питер, 2001. – 343 с.
17. Thomsen E. OLAP Solutions: Building Multidimensional Information Systems. – Wiley Computer Publishing, 1997. – 852 p.
18. Дюк В.А., Самойленко А.П. Data Mining: Учебный курс. – СПб.: Питер, 2001. – 312 с.
19. Уоссермен Ф. Нейрокомпьютерная техника. – М.: Мир, 1992. – 468 с.
20. Архипенков С., Голубев Д., Максименко О. Хранилища данных. От концепции до внедрения / Под общ. ред. С.Я. Архипенковаю – М.: ДИАЛОГ-МИФИ, 2002. – № 7. – С. 78-104.
21. Том Салливан. Это надо рисовать: Программное обеспечение анализа данных становится более выразительным // ComputerWorld Россия. – 2000. – № 42. – С. 53-64.
22. B. de Ville. Microsoft Data Mining. – Digital Press, 2001.
23. Liautaud B., Hammond M. e-Business Intelligence: Turning Information into Knoledge into Profit. – McGraw-Hill, 2001.
24. Кристин Комафорд. Корпоративная отчетность: Серверная архитектура для распределенного доступа к информации // Открытые системы. – 1999. – № 2. – С. 68-97.

Є.В. Малащук, Д.В. Бабін, С.М. Вороной, М.Г. Кочеткова

Огляд існуючих алгоритмів Data Mining для глибинного аналізу текстів і методів здобуття знань

У даній статті розглядаються сучасні підходи до одержання нових знань на основі аналізу інформаційного простору корпоративних мереж і мережі Internet і до методів обробки інформаційних потоків з метою здобуття знань за технологією Data Mining у галузі Text Mining. Ставиться задача дослідження методологій у сфері застосування Text Mining, виділяється специфіка сучасних вимог до ефективної інтелектуальної переробки даних. Аналізуються проблеми, що незадовільно вирішуються існуючими методами попередньої обробки і доступу до великих обсягів інформації, розглядаються особливості різних інформаційно-пошукових систем і засобів здобуття знань. Велику увагу приділено новому напрямку обробки текстової інформації – «глибинному аналізу текстів» (Text Mining), що поєднує в собі технологічні і методологічні підходи контент-анализу, комп'ютерної лінгвістики і штучного інтелекту.

Статья поступила в редакцию 18.07.2005.