

## О КОМПЬЮТЕРНЫХ ТЕХНОЛОГИЯХ ПОИСКА ЭМПИРИЧЕСКИХ ЗАКОНОМЕРНОСТЕЙ В БАЗАХ ДАННЫХ

В статье через призму исследовательской практики анализа массивов социологических данных рассматриваются современные технологии Data Mining, проявившие свою эффективность в поиске эмпирических закономерностей. Очерчивается контур методологических проблем использования компьютерных инструментов Data Mining. Анализируются познавательные возможности Data Mining.

*Ключевые слова:* Data Mining, эмпирическая закономерность, база социологических данных, компьютерная технология, программное обеспечение, прикладная статистика, распознавание образов, алгоритм, инструмент визуализации.

Ускорение темпов накопления информации в базах социологических данных сопряжено с проблемой ее своевременного анализа и осмысления. В этом контексте получает актуальность задача поиска эмпирических закономерностей. Несмотря на высокий приоритет этой задачи до сих пор пути к ее решению не получили окончательного методологического оформления. Тем не менее, в этом направлении проявили свою продуктивность технологии Data Mining [1–4], некоторые из которых имеют пока еще экспериментальный характер. Однако это не помешало получить им, на наш взгляд, заслуженную популярность.

Целью настоящей статьи является рассмотрение аналитического потенциала Data Mining в поиске эмпирических закономерностей на основании личного опыта автора, изучения имеющихся публикаций по данной проблематике и компьютерных инструментов, представленных на рынке программного обеспечения. В этой связи будет уделено внимание: компьютерному инструментарию Data Mining; проблемам использования Data Mining и выбору продуктивной технологии поиска эмпирических закономерностей в базах социологических данных; перспективам развития этой технологии.

### Основные используемые понятия

Понятие «Data Mining» в научной литературе не получило однозначной трактовки и употребляется либо в широком, либо в узком смысле, в зависимости от того, какой аспект

---

\* **Алексей Мичеславович Островский** – кандидат социологических наук, специалист отдела методологии ОАО «Всероссийский центр изучения общественного мнения» (Белгород).

анализа данных высвечивается в большей степени (методологический или инструментальный). *Методологический аспект* очерчивает систему научных представлений Data Mining через взаимосвязь типов данных, методов сбора (в том числе методов измерения), метаметодик и стратегий анализа данных (см. понятие «Методология анализа данных»: [5, с.216]). Эта система научных представлений формируется сейчас на границах различных дисциплин, представленная широким фронтом от теории информации до прикладной статистики. *Инструментальный аспект* включает в орбиту своего рассмотрения специальные алгоритмы Data Mining, методики, вычислительные процедуры, а также разнообразные программно-технические средства для анализа и представления (визуализации) данных. Совокупность перечисленных элементов, так или иначе связанных с анализом данных, предлагается именовать общим термином «*аналитический инструментарий Data Mining*», а ту его часть, которая относится к компьютерным системам, – «*компьютерный инструментарий Data Mining*» или, по-другому, – «*компьютерные инструменты Data Mining*». Систему методологического и инструментального обеспечения процессов *Data Mining* мы будем называть «*технологией Data Mining*» и по аналогии ту ее часть, которая относится к компьютерным системам, – «*компьютерной технологией Data Mining*».

Таким образом, в широком смысле под «Data Mining» понимается особая междисциплинарная область анализа, раскрывающая свой потенциал в процессе поиска эмпирических закономерностей, опираясь на прикладную статистику, распознавание образов, теории баз данных и искусственного «интеллекта» (Artificial Intelligence). Неоднозначность используемого здесь слова «закономерность» вытекает из терминологии английского языка, на котором издается большой объем литературы по рассматриваемой нами проблематике. «Закономерность» в английском языке понимается и как «паттерн» (pattern), и как «правило» (rule), и как «регулярность» (regularity). Каждое из этих определений имеет в контексте Data Mining свое значение. Условно можно считать, что «паттерн» – это некоторый шаблон структуры данных. «Правило» – форма представления взаимосвязи между «паттернами». «Регулярность» – свойство «паттерна» отражать структуру данных. Исторически сложилось так, что в эмпирической социологии за этими терминами закрепились иные трактовки. Такая ситуация привела к понятийным коллизиям. С целью избежать этих коллизий указанные термины используются, как правило, без отрыва от контекста употребления. В статье используется определение «*эмпирическая закономерность*». Под «эмпирической закономерностью» мы предлагаем понимать существенную, статистически-значимую взаимосвязь вероятностной природы между эмпирическими объектами, выраженную в количественной или качественной форме.

В узком смысле под «Data Mining» мы будем понимать приемы, способы, алгоритмы извлечения («добычи», «раскопки») новых, потенциально полезных свойств данных, процедуры выявления эмпирических закономерностей, взаимосвязей между переменными в больших массивах. Сюда можно отнести инструменты распознавания образов, регрессионного анализа, корреляционного анализа, факторного анализа, поиска ассоциаций, выбросов, аномалий в данных, а также специальные математико-статистические, эвристические алгоритмы и соответствующие им компьютерные программы<sup>1</sup>. Во всем инструментарии социолога подобные средства анализа занимают важное положение, поскольку позволяют эффективно: 1) «обогащать» базы данных (под «обогащением» мы понимаем совокупность технологических процессов первичной обработки данных с целью выделения эмпирических объектов, пригодных для дальнейшего анализа и целесообразного использования<sup>2</sup>); 2) решать задачи моделирования, классификации и прогнозирования.

### **Специфика использования компьютерного инструментария Data Mining**

Спектр компьютерных программных продуктов для поиска эмпирических закономерностей, в той или иной степени реализующих технологии Data Mining, достаточно широк: Clementine, IBM Iminer, Microsoft SQL Server, Oracle Data Mining, Polyanalyst, R, SAS Enterprise Miner, SPSS, Statistica, Weka, WizRule, WizWhy и многие другие. Однако их активное использование сдерживается рядом *субъективных* и *объективных* причин.

К первым из них относятся необходимость владения специальным математическим аппаратом и трудоемкость проведения подготовительных, а также окончательных («шлифовальных») работ, имеющих итерационный характер. На трудоемкость обращают внимание эксперты, которые считают, что предварительные этапы Data Mining занимают не менее 80% временных ресурсов к общему объему работ [6]. Проведение компьютерных расчетов требует тщательной подготовки со стороны исследователя и глубокого понимания реакции алгоритма на структуру исходных для анализа данных. Необходимость многократной рефлексии и объективизации полученных результатов, трансформации их в знания через призму общепризнанных практик, а также мировоззрения исследователя, его практического опыта, интуиции, убеждений, устремлений, мотиваций – создают серьезные сложности применения компьютерных технологий Data Mining.

---

<sup>1</sup> Существуют электронные источники, которые оперативно обновляются и содержат самую свежую информацию о Data mining: [http://www.spss.com/data\\_mining](http://www.spss.com/data_mining); [http://ru.wikipedia.org/wiki/Data\\_mining](http://ru.wikipedia.org/wiki/Data_mining)

<sup>2</sup> Сюда можно отнести «очистку» данных, отсев сфабрикованных наблюдений, ремонт выборки и разнообразные процедуры стандартизации и нормализации данных

К числу *объективных* причин, кроме, сопряженных с известными проблемами математической формализации в социологических исследованиях (например, обсуждаемых в работах [5, 7, 8]), относятся экономические, эргономические и т.д. Объективные причины напрямую не зависят от пользователя, но влияют на используемость программного обеспечения. Прежде всего – это высокая стоимость полнофункциональных лицензионных версий программ, недружественные интерфейсы, функциональная перегрузка аналитических инструментов. В подтверждение последнего тезиса А.Д. Наследов пишет: «Число предлагаемых критериев настолько велико, что даже руководство пользователя программы SPSS объемом около 3000 страниц оказалось не в состоянии вместить описания их всех» [9, с.182]. При этом библиографические источники, на которые ссылаются авторы работ под названием «*Руководство для пользователя*» содержат противоречивые сведения, часто отражают дискуссионные авторские позиции и недостаточно глубоко освещают практические вопросы.

Объективные проблемы получают приемлемые решения в современных реалиях. Программное обеспечение, распространяемое на свободных условиях (например, Weka [10]), несмотря на сопротивление со стороны монополистов, постепенно вытесняет коммерческие программные продукты. С недружественностью интерфейсов призваны бороться электронные ассистенты и подсказки различного рода. Редакторы планирования экспериментов (например, такие как в среде Clementine [11]) значительно облегчают выполнение анализа. Раньше аналитик был ограничен узкими рамками таких средств как «командная строка» или «электронная таблица». Сейчас в его распоряжении появились графические средства создания и редактирования схем для потокового анализа данных (в среде Clementine подобные схемы называют «визуальными картами» или «стримами», от английского слова «stream»). Проблема функциональной перегрузки решается уже самим пользователем. Оконные интерфейсы компьютерных программ в принципе позволяют гибко настраивать панели инструментов и меню, скрывать неиспользуемые кнопки (такая функция может работать автоматически) и определять режимы вывода информационных сообщений. Однако доступ к этим возможностям предоставляют не все программы. Нередко можно наблюдать ситуацию, когда изменить порядок кнопок или разделов меню не представляется возможным.

Еще одним фактором, препятствующим полнофункциональному использованию компьютерных инструментов, является преднамеренное сокрытие разработчиками программных продуктов тех или иных алгоритмов извлечения латентной информации, представляющих собой коммерческую тайну. Поэтому в документации к программам могут отсутствовать инструкции и вся необходимая спецификация ассортимента методов Data Mining, которые предлагаются пользователю как «вещь в себе».

Опыт автора статьи в области компьютерного анализа данных позволяет утверждать, что результаты, полученные в процессе запуска «непрозрачных» вычислительных схем, далеко не во всех случаях убедительны. Отдельные программы не предоставляют достаточных сведений для получения доказательных выводов и исчерпывающих уточнений. Бывает очень трудно или совершенно невозможно найти информацию о границах применимости аналитических методов Data Mining. Наши эксперименты показали, что даже родственные, согласно имеющимся описаниям, алгоритмы могут давать не только количественно, но даже качественно различные результаты. Влияет на это обилие всевозможных опций и плохо документированных режимов запуска вычислительных процессов.

Социологи-эмпирики в условиях отсутствия понятных и надежных инструментов вынуждены слепо доверять электронным отчетам. Предполагаем, что в большинстве случаев они не только не в состоянии воспроизвести вычисления посредством карандаша и бумаги, но не могут даже примерно очертить область правдоподобия результатов. Это вполне закономерно. Эволюция информационных технологий сопровождается стремительным усложнением инструментов Data Mining, которые отгораживают социолога от практики анализа данных труднопреодолимым частоколом формул и специальных правил их применения.

Имеет место и полярная проблема, когда у социологов наблюдается «головокружение от успехов» в области компьютерного анализа данных. В этом случае не стоит забывать, что компьютерные программы оперируют абстракциями, в контексте которых исследуемые социологические феномены могут терять содержание и смысл, превращаясь в умозрительные конструкции.

Качество результатов работы не только экзотических методов Data Mining, но и кластерного, факторного, регрессионного анализа в многомерном случае не всегда легко проверить. Математико-статистические оценки качества не гарантируют успешной интерпретации результатов. Поэтому существует сильный соблазн попросту подогнать расчеты под свои гипотезы. Эта ситуация усугубляется, когда приходится решать задачи не имеющие числового решения.

Вероятно, следует задуматься, когда выводы рождаются «на скорую руку» непосредственно в рамках расчетных моделей или когда рассуждения социолога подгоняются под определенный компьютерный стандарт. Отмеченному выше «головокружению» потворствуют «интеллектуальные» методы Data Mining из серии «для ленивых». В рекламных проспектах к программному обеспечению встречаются утверждения, что пользователь может применять подобные методы, даже не имея какой-либо начальной математической подготовки. Ведь достаточно просто нажимать кнопки в

нужной последовательности, а программа сама выдаст результат. Однако более глубокое изучение подобных программ приводит к пониманию того, что пользователь как минимум должен достаточно хорошо ориентироваться в выборе управляющих параметров расчета, которые приходится корректировать всякий раз для нового эксперимента. Опции, выставленные «по умолчанию», дают адекватные результаты только для тривиальных случаев. Более подробно эти вопросы рассматривает в своей работе А.П. Кулаичев [12].

### **Ограничения на применимость технологий Data Mining**

«Методологическая сила» технологий Data Mining имеет известные ограничения. Не все методы Data Mining подпадают под определение алгоритма, под которым понимается конечная последовательность вполне определенных действий (шагов), приводящая к желаемому результату. Это связано с тем, что вычислительные стратегии оказываются зависимыми от потока эмпирических данных. В этом случае вычисления не могут быть точно и детально определены априори. Только апостериорная оценка данных при начальных итерациях позволяет определить дальнейший ход решения задачи.

Алгоритмами, например, не являются достаточно популярные в последнее время искусственные нейронные сети (введены в базовый пакет SPSS версии 16.0), представляющие собой одно из перспективных направлений по распознаванию образов. Используемые в пакете эвристики основаны на статистической очевидности и не гарантируют получения достоверных результатов для всего множества исходных для анализа значений. При обучении нейронной сети возможны такие проблемы как «... паралич или попадание сети в локальный минимум поверхности ошибок. Невозможно заранее предсказать появление той или иной проблемы, равно как и дать однозначные рекомендации к их разрешению» [13]. Структуру нейронной сети после успешного обучения не всегда просто описать математически. Кроме того, эти описания при решении практических задач оказываются чрезмерно сложными и почти непригодны для содержательно анализа.

Положение не спасает наличие развитых инструментов визуализации эмпирических закономерностей [3, с. 94 – 116], так как в исследовании многомерных связей они малоэффективны. Структуры размерности 4 и выше не поддаются пространственному восприятию человека, а как следствие – целостному анализу и осмыслению. Проекция же этих структур на двумерную плоскость или в трехмерное пространство, как камни из японского сада Рёандзи, всегда имеют «слепые пятна», порождаемые необходимостью «заморозки» «лишних» размерностей. Кроме того, эффективное использование различных математических метрик при сокращении размерности анализируемой структуры требует

глубокого абстрагирования и невозможно без развитого пространственного воображения, которым владеют даже не все профессиональные исследователи. В целях усиления наглядности приходится жертвовать определенной информацией, что непременно сказывается на увеличении абстрактности получаемых результатов. Сказанное почти без изменений можно отнести практически ко всем многомерным методам. Отмеченные проблемы обуславливают невысокое доверие пользователей к используемым инструментам Data Mining и вызывают значительные затруднения при интерпретации полученной информации.

Область исследований Data Mining находится в процессе институционализации. Поэтому не существует четких инструкций по использованию технологий Data Mining и не очерчены границы их эффективности. Однако усилиями консорциума компаний SPSS, NCR, DaimlerChrysler и OHRA разработан специальный межотраслевой стандарт CRISP-DM<sup>1</sup> (Cross-Industry Standard Process for Data Mining), представляющий собой пошаговое руководство по использованию методологического потенциала Data Mining. На наш взгляд, слабой стороной этого стандарта является его ориентированность на бизнес-задачи и коммерческие приложения в ущерб научным исследовательским задачам. Кроме того, многие положения этого стандарта, как собственно и ряда других, аналогичных ему (например, SEMMA от SAS [14]) лишь приблизительно, в общих чертах описывают сущность методологической стратегии и поэтому допускают свободную интерпретацию. Консультированием по широкому кругу вопросов Data Mining на профессиональной основе занимаются консалтинговые компании, среди которых следует выделить Two Crows<sup>2</sup>. Консультации носят коммерческий характер и поэтому не всегда доступны обычным пользователям Data Mining.

Компьютерный анализ социологических данных, аккумулировавших социальный опыт в условиях неполноты, неточности, противоречивости информации, требует привлечения дополнительных сведений и использования специальных вычислительных (главным образом, комбинаторных) схем, достаточно трудоемких во временном отношении. Последнее обстоятельство приводит к необходимости введения определенных упрощений, большинство из которых связано с вероятностными допущениями. Такие допущения, принятые без достаточных оснований, могут существенно сказаться на результативности компьютерных расчетов. Так, популярный алгоритм распознавания образов k-средних [15] при выборе начального местоположения центров тяжести классов использует случайный выбор. Это не всегда оправдано и может приводить в ряде случаев к неоднозначной классификации. На пользователя возлагается важная обязанность

---

<sup>1</sup> Официальный сайт: <http://www.crisp-dm.org>

<sup>2</sup> Официальный сайт: <http://www.twocrows.com>

грамотного выбора числа классов и подбора режимов асимптотической сходимости метода. Но в общем случае возникает необходимость в радикальных средствах – в привлечении априорной информации через введение дополнительных моделей качества решения и опорных координат центров тяжести. Далек не каждый программный продукт предоставляет пользователю такие возможности.

Аналогичные рассуждения можно отнести и к так называемым «генетическим алгоритмам», которые активно используют в решении задач поиска вероятностные процедуры с датчиками случайных чисел [3, с.166–172]. Работа генетических алгоритмов заключается в последовательном подборе, комбинировании и вариации искомых параметров данных с использованием вычислительных механизмов, напоминающих дарвинские механизмы биологической эволюции [16]. Подобную искусственную «эволюцию» может «заклинить» на непродуктивном пути поиска решения в зоне локального максимума, а оптимальное решение, если оно существует, окажется недостижимым. В этом случае также требуется привлечение дополнительной информации. В специальной литературе подобным проблемам уделяется достаточно много внимания, однако эффективных методик для конечного пользователя пока не выработано. Следует подчеркнуть, что нейронные сети и генетические алгоритмы относятся к разведочному инструментарию, поэтому спектр возможных решений при тех или иных параметрах расчета нужно рассматривать не более как вспомогательный материал.

### **Выбор продуктивной технологии поиска эмпирических закономерностей**

Data Mining выгодно отличается от традиционных технологий, которые не справились в современных условиях с обработкой и анализом нахлынувших потоков разнородной информации. Существуют многочисленные научные труды, учебные пособия, специальные руководства и электронные публикации, в которых достаточно глубоко рассматриваются проблемы использования технологий Data Mining. Например, одна из самых популярных поисковых систем Интернет Google<sup>1</sup> выдает на конец марта 2008 года по запросу «Data Mining» более 15 миллионов электронных источников. Это число позволяет судить о значимости и востребованности Data Mining.

В социологической литературе достаточно много теоретико-методологических статей, восхищающихся достоинствами различных математических методов Data Mining, но крайне мало работ, в которых описываются содержательные схемы анализа с соответствующими им расчетами и доказательствами продуктивности. Во многих статьях высвечиваются главные, вполне ясные и даже очевидные аспекты, в ущерб методическим

---

<sup>1</sup> Режим доступа: <http://www.google.com>

нюансам и вычислительным тонкостям, влияющих на валидность и состоятельность конечных интерпретаций. Статьи в российских социологических журналах по смежной с Data Mining проблематике носят пока еще обзорный, отчасти переводной характер, например [17, 18]. На наш взгляд, электронные источники на русском языке поверхностны по содержанию, разобщены и зачастую имеют коммерческий характер.

Каждый метод Data Mining требует специальной адаптации и тщательного социологического апробирования. Без этого невозможно его полномасштабное использование. В условиях сложившейся ситуации исследователи вновь и вновь «набивают себе шишки», «наступая на грабли» и каждый раз по-новому интерпретируя пробелы в документации к программному обеспечению.

На практике часто используется не какой-нибудь один, а целая группа инструментов. Методы запускаются в пошаговом режиме, в определенной последовательности. Задача социолога – организовать их согласованную работу, ведь компьютерной программе нельзя сообщить конечную цель анализа в явном виде. В многообразии всех инструментов и вычислительных стратегий легко запутаться. Поэтому существуют пакеты шаблонов построения вычислительных экспериментов (для Clementine это CAT – Clementine Application Template [19]) и специальные электронные «путеводители» по дереву математико-статистических методов. Подобные «путеводители» позволяют в каждом конкретном случае выбрать продуктивную технологию анализа и подобрать адекватные компьютерные инструменты. Среди отечественных разработок в этом направлении следует отметить раздел «Какой метод анализа выбрать?» справочной службы программы STADIA, разработанной А.П. Кулаичевым<sup>1</sup>. Автору статьи неизвестны другие подобные этому справочники для Data Mining, ориентированные на пользователя.

На переднем крае развития технологий Data Mining следует выделить методы поиска эмпирических закономерностей, которые эффективны для анализа нечисловой информации. Развернутое представление о методах изучения объектов нечисловой природы в контексте социологического анализа можно найти в работе Ю.Н. Толстой [20].

Проблемы автоматизации поиска эмпирических закономерностей в данных имеют давнюю историю. Алгоритм случайного поиска с адаптацией был разработан Г.С. Лбовым [3, с. 197 – 198]. Первые комбинаторные алгоритмы ограниченного перебора были предложены в 1960-х годах. Здесь особо следует выделить пионерскую работу отечественного кибернетика М.М. Бонгарда [21]. Среди современных программных инструментов неполного перебора лидируют продукты WizWhy, WizRule корпорации WizSoft<sup>2</sup>. Программы имеют добротные справочные руководства пользователя,

---

<sup>1</sup> Демо-версию STADIA можно скачать по ссылке: <http://statsoft.msu.ru/products.htm>

<sup>2</sup> Официальный сайт, где можно скачать демо-версии продуктов WizWhy и WizRule: <http://www.wizsoft.com>

нелишенные, впрочем, некоторых недостатков. Несмотря на отмеченное лидерство можно подобрать простые примеры, когда указанные продукты демонстрируют свои ограничения, о которых недостаточно ясно сказано в документации (см., например, тесты В.А.Дюка «Умение решать очевидные задачи», «Умение находить наиболее полные и точные правила», «Ложные закономерности» [22]).

Отдельного внимания заслуживают методы построения деревьев решений, например, реализованные в модуле SPSS AnswerTree: CHAID, Exhaustive CHAID, CRT, QUEST [23]. Несмотря на очевидные преимущества, заключающиеся в простоте, наглядности, структурности представления решения, подобные методы имеют довольно узкую область эффективности. В.А. Дюк считает, что они плохо приспособлены для решения многомерных задач, когда однозначный выбор корня дерева построения решения невозможен. В процессе сегментации признак (корень дерева) «вырывается» из целостной системы описания многомерного объекта [22]. Безусловно, такая ситуация наносит серьезный вред качеству интерпретации.

В начале 1990-х годов активизировались исследования по разработке методов поиска ассоциативных правил [24]. Эти методы востребованы в маркетинговых исследованиях, когда, например, нужно в динамике проанализировать потребительскую корзину. Появились алгоритмы AIS, SETM, затем – Apriori [25]. Последний алгоритм распространяется сейчас в виде программной реализации на свободных условиях. В связи с необходимостью повышения эффективности работы алгоритма, разработаны его модификации AprioriHybrid, AprioriTid, которые позволяют получить решение за меньшее число обращений к записям базы данных.

Мы считаем, что выявление ассоциативных правил в формате логического выражения может способствовать информативному раскрытию исследуемых эмпирических закономерностей, имеющих нечисловую природу и дополнить результаты, полученные в рамках традиционных подходов. Самой простой формой представления правил является импликация вида «если ..., то...» (if-then) [22].

Следует подчеркнуть, что комбинаторные методы, лежащие в основе большинства технологий поиска ассоциативных правил Data Mining, в указанном формате, избегают полного перебора вариантов решения, так как этот перебор даже для работы с относительно небольшими массивами данных требует привлечения значительных вычислительных затрат, несопоставимых с вычислительными возможностями ни современных компьютеров, ни суперкомпьютеров ближайшего будущего. Поэтому в компьютерных технологиях Data Mining используются специальные эвристики – различные вычислительные приемы, позволяющие обойти полный перебор и ускорить процесс решения задачи. Однако при этом решение существенно теряет в полноте и/или точности. Иначе говоря, вместо «истинных»

закономерностей, обнаруживаются их отдельные «фрагменты», так называемые «осколки знаний» [26]. На пути решения этой проблемы предлагаются усовершенствованные технологии Deep Data Mining, существенно не уступающие по эффективности алгоритмам полного перебора при решении поисковых задач, однако обладающие приемлемой вычислительной сложностью [27].

К основным достоинствам этой, на наш взгляд, продуктивной технологии Deep Data Mining для поиска эмпирических закономерностей в формате «если ..., то...» следует отнести следующие положения:

1) совместно могут быть проанализированы переменные, измеренные по разнотипным шкалам, что крайне важно в социологических исследованиях;

2) могут быть проверены описательные гипотезы, а также найдены неожиданные, аномальные (ad-hoc) эмпирические закономерности, «открывающие» новые стороны исследуемых феноменов;

3) имеется обоснование валидности найденных эмпирических закономерностей через введение системы оценок точности и полноты правил в формате «если ..., то...» (пусть  $R$  – это правило вида  $R = \{\text{если } A, \text{ то } B\}$ , тогда *точность*  $R$  – это доля  $B$  среди случаев  $A$ , выраженная, например, в относительных единицах, *полнота*  $R$  – это доля случаев  $A$  среди случаев  $B$  [3, с. 192]);

4) технологию можно использовать для выявления сфальсифицированных наблюдений, когда, например, нарушаются или, наоборот, «подозрительно» полно и точно подтверждаются эмпирические закономерности;

5) технология сохраняет свою продуктивность в других перспективных направлениях исследовательского поиска – в «интеллектуальном» анализе текстов (Text Mining) и анализе сетевого контента (Web Mining).

### **О перспективах развития поисковой технологии Deep Data Mining**

Направление дальнейшего совершенствования технологии поиска эмпирических закономерностей в базах социологических данных, по нашему мнению, находится в русле стратегии «смыслового усиления» и оптимизации описания обнаруженных закономерностей в формате «если ..., то...». Существующие технологии Deep Data Mining не позволяют окончательно уйти от избыточных, либо, наоборот, неполных решений. Разработанная система математико-статистических оценок качества решения не может снять эту проблему, ибо принципиальное ограничение заключается в том, что технология не в состоянии при выборе решения «разумно» оперировать такими понятиями как «смысл» или, скажем, «оптимальность». В лучшем случае инструменты Deep Data Mining могут

потребовать уточнений со стороны пользователя, описав ему проблемную ситуацию на специальном языке. Поэтому проводить «глубокий» (deep) анализ данных могут только специально подготовленные аналитики или эксперты, способные сообщить компьютерной системе необходимую информацию.

В завершении остается отметить, что какими бы совершенными и изощренными ни были бы компьютерные инструменты Deep Data Mining, они никогда полностью не освободят социолога от необходимости думать и активно участвовать в процессе анализа данных. В лучшем случае можно добиться проведения анализа в полуавтоматическом режиме под управлением и контролем человека. При этом компьютер возьмет на себя выполнение трудоемких вычислительных и других рутинных операций, например, связанных с визуализацией промежуточных и конечных результатов, а социолог – всю творческую работу.

## ЛИТЕРАТУРА

1. Анализ данных // Энциклопедия социологии. Режим доступа: <http://slovari.yandex.ru/art.xml?art=sociology/soc/soc-0032.htm>
2. *Витяев Е.Е.* Извлечение знаний из данных. Компьютерное познание. Модели когнитивных процессов: Монография. Новосибирск: Изд-во Новосибирского гос. ун-та, 2006.
3. *Дюк В., Самойленко, А.* Data Mining. СПб.: Питер, 2001.
4. *Чубукова И.А.* Data Mining. М.: Интернет-университет информационных технологий: БИНОМ: Лаборатория знаний, 2006.
5. *Татарова Г.Г.* Основы типологического анализа в социологических исследованиях. Учебное пособие. М.: Издательский Дом «Высшее Образование и Наука», 2007.
6. Технологии баз данных Data Mining. Мнение экспертов о Data Mining. Режим доступа: <http://www.technology-db.ru/about/clause/412/241171/>
7. *Татарова Г.Г.* Методология анализа данных в социологии (введение). Учебник для вузов. М.: NOTA BENE, 1999.
8. *Толстова Ю.Н.* Может ли социология «разговаривать» на языке математики? // Социологические исследования. 2000. №5. С. 107 – 116.
9. *Наследов А.Д.* SPSS: Компьютерный анализ данных в психологии и социальных науках. 2- изд. СПб: Питер, 2007.
10. Weka. Режим доступа: [http://en.wikipedia.org/wiki/Weka\\_\(machine\\_learning\)](http://en.wikipedia.org/wiki/Weka_(machine_learning))
11. Clementine. Режим доступа: <http://www.spss.com/clementine/>
12. *Кулаичев А.П.* Методы и средства анализа данных в среде Windows. Изд. 3-е. М.: Информатика и компьютеры, 1999.

13. Нейронные сети. Режим доступа: [http://ru.wikipedia.org/wiki/Нейронные\\_сети](http://ru.wikipedia.org/wiki/Нейронные_сети)
14. SEMMA. Режим доступа:  
<http://www.sas.com/technologies/analytics/datamining/miner/semma.html>
15. K-means algorithm. Режим доступа: [http://en.wikipedia.org/wiki/K-means\\_algorithm](http://en.wikipedia.org/wiki/K-means_algorithm)
16. Генетический алгоритм. Режим доступа:  
[http://ru.wikipedia.org/wiki/Генетический\\_алгоритм](http://ru.wikipedia.org/wiki/Генетический_алгоритм)
17. *Давыдов А.А.* О компьютерной теории социальных агентов // Социологические исследования. 2006. №2(262) С.19 – 28
18. *Давыдов А.А.* Компьютерная теория социальных систем // Социологические исследования. 2005. №6(254) С.14 – 24.
19. Get better results with application best-practice templates. Режим доступа:  
<http://www.spss.com/clementine/cats.htm>
20. *Толстова Ю.Н.* Анализ социологических данных. Методология, дескриптивная статистика, изучение связей между номинальными признаками. М.: Научный мир, 2000.
21. *Бонгард М.М.* Проблема узнавания. М.: Наука, 1967.
22. *Дюк В.А., Асеев М.Г.* Поиск if-then правил в данных: проблемы и перспективы. Режим доступа: <http://www.datadiver.nw.ru/Articles/Problems.htm>
23. AnswerTree. Режим доступа: [www.spss.com/answertree/](http://www.spss.com/answertree/)
24. Association rule. Режим доступа: [http://en.wikipedia.org/wiki/Association\\_rule](http://en.wikipedia.org/wiki/Association_rule)
25. Apriori algorithm. Режим доступа: [http://en.wikipedia.org/wiki/Apriori\\_algorithm](http://en.wikipedia.org/wiki/Apriori_algorithm)
26. *Дюк В.А.* Осколки знаний. Режим доступа:  
[http://www.inftech.webservis.ru/it/articles/v\\_a\\_duk/ar1.html](http://www.inftech.webservis.ru/it/articles/v_a_duk/ar1.html)
27. *Дюк В.А.* Data Mining – состояние проблемы, новые решения. Режим доступа:  
<http://www.inftech.webservis.ru/it/database/datamining/ar1.html>