Преподавание интеллектуального анализа данных студентам-социологам

Кислова О. Н., канд. социол. н., доц.; Харьковский национальный университет имени В.Н. Каразина, Харьков, Украина

> Что может быть заманчивее раскрытия природы талантливого мышления и превращения такого мышления из редких и неустойчивых вспышек в мощный и управляемый огонь познания!
> ***Г.С.Альтшуллер. Творчество как точная наука***

Вступление человечества в информационный век связано прежде всего с колоссальными изменениями в сфере информационной деятельности. Сегодня наша жизнь практически немыслима без компьютера, Интернета и других информационных технологий, которые с каждым днем становятся все более дружественными и удобными благодаря внедрению в них новейших технологических инноваций, в частности, элементов искусственного интеллекта. Интеллектуализация, став императивом развития современных средств коммуникации, поиска информации, вычислений, обработки и анализа данных, значительно повышает доступность информационных технологий для пользователей, имеющих разные уровни компьютерной подготовки.

В контексте развития социологического знания названные процессы постепенно приводят к изменению способов извлечения новых знаний из эмпирических данных. При этом ни в коем случае нельзя говорить об отрицании многолетнего опыта использования традиционных методов анализа социологической информации. Изменение способов обработки и анализа массивов социологических данных, и, как следствие, получение новых знаний об исследуемых социальных феноменах скорее связано с привнесением в практику социологов-аналитиков новых методов и инструментов, которые появились (и продолжают появляться) в процессе становления интеллектуального анализа данных (ИАД). Поэтому перед системой социологического образования встают новые задачи, обусловленные необходимостью внедрения в учебные программы курсов, которые позволят будущим социологам ознакомиться с новейшими достижениями в области обработки и анализа данных. При этом возникает вопрос об актуальности таких нововведений на фоне имеющихся проблем в математическом образовании социологов. Данная работа посвящена прежде всего аргументации целесообразности введения курсов ИАД на социологических факультетах, а также размышлениям о том, в каком объеме и в какой форме лучше преподносить нашим слушателям ИАД, учитывая стойкую «нелюбовь» гуманитариев к любым видам математического формализма.

Прежде всего необходимо сказать несколько слов о сущности ИАД и специфике его применения в социологии. Как известно, интеллектуальный анализ данных (ИАД) представляет собой аналитический процесс исследования человеком большого объема

информации с привлечением компьютерных технологий автоматизированного поиска в исходных данных неочевидных закономерностей, что расширяет возможности аналитика, позволяя не только проверять имеющиеся гипотезы, но и генерировать новые, априорно не прогнозируемые исследователем.

Актуальность рассмотрения новых возможностей исследования социальной реальности, которые открываются перед социологами в связи с развитием ИАД, обусловлена следующими причинами. Во-первых, появляются новые инструменты анализа данных – компьютерные программы, реализующие не только статистические алгоритмы анализа данных, но и методы, основанные на недавних достижениях в области математики: нейронных сетях, генетических алгоритмах, теории нечетких множеств, синергетике, фрактальной математике, теории игр и др. Усовершенствование компьютерной техники («железа»), позволившее интегрировать новую математику и современное программное обеспечение, явилось основой появления разнообразных технологий ИАД.

Первоначально создавались специальные программные продукты для реализации отдельных методов интеллектуального анализа данных, а в последнее время наметилась тенденция включения наиболее распространенных методов ИАД в пакеты статистической обработки данных (примерами тому служат последние версии SPSS и STATISTICA). Процесс все более полной интеллектуализации современных технологий анализа эмпирической информации нацелен на создание максимума удобств аналитику, на устранение необходимости тратить время на рутинные операции (но постановка содержательной задачи и интерпретация результатов остается прерогативой человека). Таким образом, освоение будущими социологами основ работы с современным программным обеспечением позволит им «одним щелчком мыши» выявлять в данных закономерности, исследовать их разные аспекты, используя современные технологии для решения профессиональных задач и не превращаясь при этом в «бесплатное приложение к компьютеру».

Во-вторых, появление и стремительное развитие технологий интеллектуального анализа web-контента, изображений, аудио и видеоматериалов, мультимедиа и т.п. инициирует процесс исследования возможностей использования новых типов данных в качестве носителей социологической информации[1] (см., например, [2, 3]). Получается, что

[1] «Социологической информацией называются любые эмпирические данные, которые содержат информацию о социальной реальности: социальных явлениях, социальных процессах, социальных общностях, социальных институтах, социальных системах, социальных группах и других социальных феноменах» [1]. До последнего времени социологи использовали в качестве эмпирических данных только числовые или текстовые массивы. Числовые массивы (квантифицированные мнения респондентов) – в количественных исследованиях, текстовые – в качественных.

сегодня технологическое развитие опережает методологическую рефлексию, ведь известно, что данные и методы их анализа тесно взаимосвязаны: методология исследования определяет тип получаемых данных, который, в свою очередь, обусловливает выбор методов их анализа. Технологии ИАД предоставляют разнообразные инструменты, позволяющие исследовать эмпирические данные любого типа: алгоритмы data mining позволяют эффективно обрабатывать и анализировать числовые данные; text mining – текстовые; web mining – web-контент; visual mining – фотографии и другие визуальные объекты; multimedia mining – мультимедийные данные и др. Большинство из перечисленных технологий «майнинга» находится сейчас в процессе становления, а практическое их применение – на самой начальной стадии. Однако, учитывая темпы технологического развития, можно предположить, что в ближайшие годы эти технологии станут обыденными инструментами анализа данных. Готовы ли социологи разработать методологические обоснования, которые позволят включить новые технологии в арсенал методов социологического анализа, является проблемой, требующей самостоятельного рассмотрения.

Подчеркнем, что интеллектуальный анализ данных невозможно реализовать без специализированных пакетов программ, реализующих эвристические алгоритмы выявления закономерностей, релевантных как накопленным данным, так и целям их обработки. Таким образом, ИАД предполагает наличие ***данных*** (числовых, текстовых или других), ***цели*** (определяющей вид искомой закономерности: ассоциации, классификации, кластеризации или др.), ***математического аппарата***, способного решить задачу поиска определенного вида закономерности, и ***программного инструмента***, реализующего соответствующий математический метод (см. рис. 1).

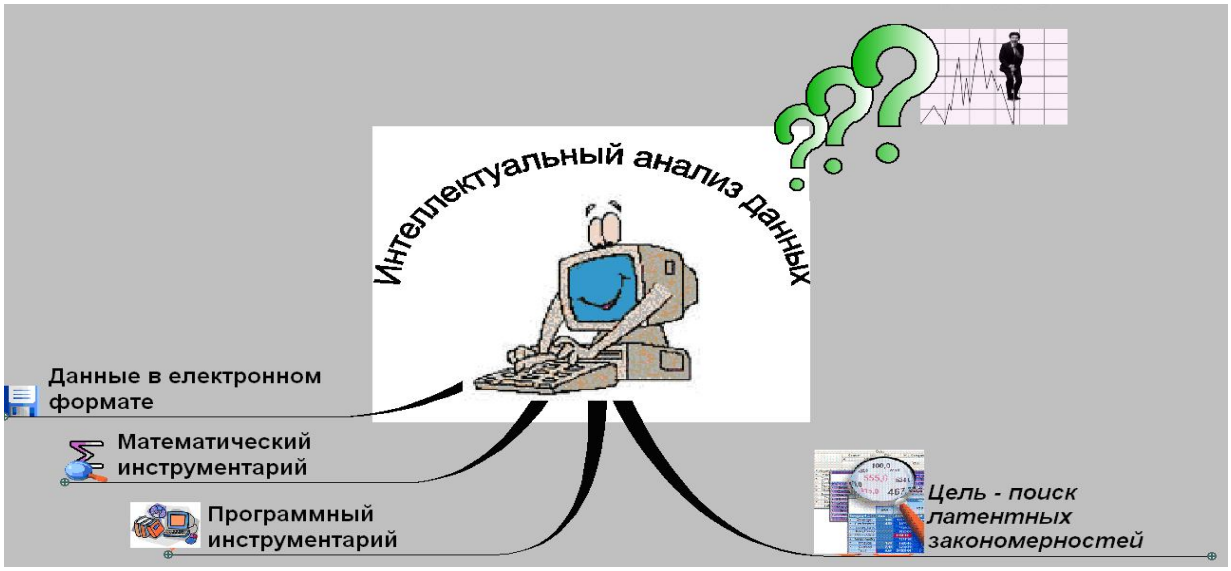

Рис. 1. Основные компоненты технологии интеллектуального анализа данных

Интеллектуальный анализ данных является синтетической областью, цель которой часто ограничивается нахождением в автоматическом режиме ***закономерностей***[2] (моделей и отношений), скрытых в базе (массиве) данных, которые не всегда могут быть найдены общеизвестными статистическими методами, хотя исторически сложилось так, что мощный арсенал методов прикладной статистики стал первым направлением развития средств интеллектуального анализа данных. Значимость статистических методов для интеллектуального анализа данных достаточно велика. Несмотря на то, что они разработаны в рамках статистической парадигмы, эти методы все же способны решать часть задач интеллектуального анализа данных, но они не позволяют генерировать новые гипотезы в автоматизированном режиме. Таким образом, формулировка гипотез остается прерогативой исследователя и производится «вручную», без помощи информационных систем.

Классическим примером применения статистических методов в процессе ИАД является проведение кластерного анализа, когда исследователь только задает классифицирующие признаки, но заранее не может предположить ни состав, ни объем кластеров. Таким образом, априорные предположения исследователя не полны, рабочая гипотеза не может быть четко сформулирована. Более того, при проведении кластерного анализа приходится перебирать « в ручную» гипотезы о количестве кластеров.

Хотя прикладная статистика и входит в математический инструментарий интеллектуального анализа данных, но она является только частью этого инструментария. В основу большинства методов ИАД положена концепция шаблонов (паттернов) и зависимостей, отражающих многоаспектные взаимоотношения в данных. Поиск паттернов может производиться автоматическими методами, не ограниченными рамками априорных предположений о структуре выборки и виде распределений значений анализируемых показателей (что является обязательным в рамках статистической парадигмы).

Важной особенностью интеллектуального анализа данных является то, что он позволяет более полно использовать способности человека, освобождая его не только от рутинных вычислений, но даже от «рутинной» формулировки гипотез (естественно, при наличии «сильной» интеллектуальной системы, оснащенной «хорошим» математическим аппаратом, позволяющим реализовать методологию генерации и отбора наиболее интересных гипотез). Однако ИАД не решает задачи за аналитика, а всего лишь служит инструментом, который способствует поиску нетривиальных решений содержательных

[2] Далее мы покажем, что поиск эмпирических закономерностей – это промежуточный этап в процессе выявления нового знания из массивов социологической информации.

задач. Для того, чтобы не профанировать методы интеллектуального анализа данных, их следует понимать, знать достоинства и недостатки, границы применения каждого из них. Кроме того, ИАД предполагает совместное использование различных методов и алгоритмов при исследовании одного и того же социального феномена, реализуя таким способом принцип триангуляции в процессе анализа эмпирических данных. В данном контексте особое значение приобретают математические знания, навыки компьютерной реализации различных методов анализа данных и корректной интерпретации полученных результатов. Соответственно, если мы хотим, чтобы отечественная социология в ближайшем будущем не осталась на «задворках» информационной эпохи, необходимо включать в учебные программы сегодняшних студентов-социологов курсы, дающие возможность приобрести такие знания и умения.

Как мы уже отмечали, целью применения отдельных методов ИАД является выявление закономерностей, скрытых в массивах социологической информации. Однако если рассматривать интеллектуальный анализ не как простую сумму методов, алгоритмов и технологических решений, а перейти на методологический уровень, то можно говорить о том, что ИАД – это не только современная концепция анализа данных, но и методология познания, предназначенная для **поиска нового знания** в обширных массивах разнородной информации.

В. К. Финн утверждает: «При широком толковании термина “интеллектуальный анализ данных” (ИАД) “data mining” и “knowledge discovery” являются видом ИАД. Это широкое толкование ИАД состоит в том, что из неупорядоченных и неформализованных данных посредством различных формальных методов, могущих перерабатывать эти данные посредством некоторых алгоритмов в интерпретируемые результаты так, что из них можно извлечь некоторые знания в явном виде такие, что до применения этих методов эти знания были скрыты в массиве данных» [4, с. 4].

Основная идея ИАД имеет глубокие философские корни и может быть выражена крылатой фразой «ничто не ново под луною», интерпретация которой в современном компьютеризированном контексте состоит в следующем: данные помимо явной информации содержат «знания» (т.е. актуальную информацию, но в неявном виде); новые знания об исследуемом объекте, породившем анализируемые данные, можно извлекать непосредственно из этих данных, применяя современные интеллектуальные технологии.

Новое знание является ключевым понятием ИАД. Что же представляет собой новое знание, ради которого социологу предлагается освоить непривычные информационные технологии? Н. Г. Загоруйко считает, что «знания представляют собой обобщенное описание основного содержания информации, представленной в данных» [5, с. 6].

Г. Пиатецкий-Шапиро, один из основоположников ИАД, утверждает, что ***новое знание в контексте ИАД*** имеет формальный характер и представляет собой интересную закономерность, конкретизируя термин «интересность закономерности» как интегральную меру ценности закономерности, с точки зрения обоснованности, новизны, полезности и понятности [6].

Мы полагаем, что ***новым знанием в контексте социологии*** (естественно речь идет только о новом знании, выявляемом при помощи ИАД, исключая другие способы получения социологического знания) является концептуальная модель, которая строится в ходе интерпретации закономерностей, найденных методами ИАД в массиве социологической информации, а используемая социологом теория выполняет функцию базы знаний, на основе которой и оценивается степень «интересности» интерпретируемой закономерности.

Специфика ИАД в социологии имеет две стороны. С одной стороны, она обусловлена теми же факторами, что и специфика анализа социологических данных, основой которого является комплексное использование разнообразных методов (как математико-статистических, так и эвристических), опирающееся на методологические принципы, подробно описанные Ю. Н. Толстовой [7].

С другой стороны, идеология ИАД предполагает ***полезность*** знаний, выявленных методами ИАД. В социологии это приводит к необходимости разделения этапа интерпретации результатов на две взаимосвязанные части:

1) Формальная интерпретация закономерностей, которая состоит в проверке степени «интересности» закономерностей теми способами, которые заложены в математический аппарат конкретного метода ИАД, что дает возможность найти новое знание, которое ***возможно*** является новым социологическим знанием.

2) Содержательная интерпретация в конкретной предметной области, которая состоит в проверке ***полезности*** знаний, найденных на предыдущем этапе, для построения и/или уточнения концептуальных моделей исследуемого социального феномена.

ИАД можно рассматривать как переход от данных к знаниям, который осуществляется благодаря выявлению закономерностей в исходных данных с использованием специфических методов «майнинга» («раскопки данных») (см. рис. 2).

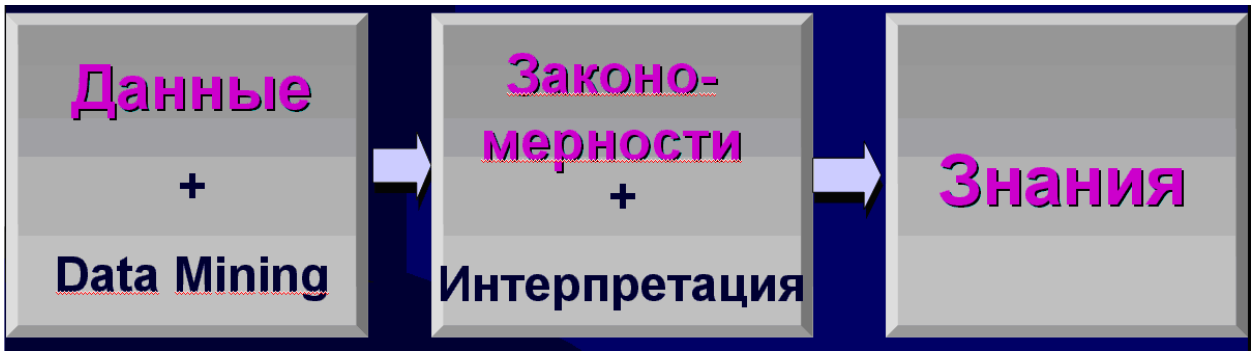

Рис. 2. ИАД как процесс извлечения знаний из имеющихся данных.

При этом необходимо отметить, что не переводимый на русский язык термин «data mining» первоначально использовался для обозначения поиска закономерностей в данных любого типа. Однако развитие технологий интеллектуального анализа текстовых массивов (text mining), а затем содержания сети Интернет (web mining), графических изображений (visual mining) и других, показало необходимость терминологического различения этих видов анализа данных, поскольку их математический аппарат имеет значительные отличия. Даже применение одних и тех же математических методов (например, нейронных сетей) в исследовании числовых и текстовых данных приобретает значительные отличия.

Наша собственная практика применения ИАД в социологических исследованиях дает основания с уверенностью утверждать, что ИАД является очень полезным инструментом познания социальной действительности. Освоение методов ИАД будущими социологами, с нашей точки зрения, является важным шагом в повышении уровня их профессиональной культуры.

Преподавание ИАД на социологических факультетах связано с разрешением большого числа вопросов. Насколько глубокие знания ИАД нужны социологу? Какие методы ИАД прежде всего следует преподавать? Нужно ли социологу в полном объеме знать математический аппарат тех методов, которые он применяет? Какие пакеты программ, реализующих ИАД, в первую очередь следует осваивать? Есть ли необходимость введения курса ИАД в учебную программу?

Учитывая «сложные взаимоотношения» большинства наших студентов с математическими дисциплинами, приходится избирать тактику преподнесения материала без строгого математического обоснования. При этом у нас теплится надежда, что те студенты, которые желают корректно применять ИАД, самостоятельно изучат рекомендуемую литературу. Мотивацией к изучению ИАД, как нам представляется, может служить ознакомление с результатами его применения.

На социологическом факультете Харьковского национального университета имени В. Н. Каразина внедрение ИАД в профессиональное образование социологов происходит постепенно. В настоящее время пока еще нет отдельного курса, посвященного ИАД. Теоретические аспекты ИАД излагаются очень кратко в пределах курса "Методы анализа социологической информации". Рассмотрение возможностей практической реализации некоторых методов ИАД входит в курс "Методы компьютерной обработки социологической информации: ОСА, SPSS". Почти две трети курса "Методы многомерного анализа социологической информации" посвящено Data Mining. Многие темы ИАД предлагаются для самостоятельного изучения наиболее заинтересованными слушателями (преподаватель при этом дает консультации, помогая ориентироваться в литературе и компьютерных программах, реализующих выбранный метод). Например, применение нейронных сетей в социологических исследованиях, возможности обработки социологических данных в Clementina, методы когнитивной визуализации в STATISTICA, инструменты Web Mining и возможности их использования в социологических исследованиях, графическое представление сложных знаний (MindManager, ConceptDraw MindMap и другой «софт для мозгов»).

Опыт преподавания названных курсов показывает, что большинство студентов-социологов, ориентированных после окончания университета работать по специальности, интересуются возможностями технологии ИАД и согласны прилагать усилия, чтобы разобраться в ее тонкостях, несмотря на сложность теоретических и прикладных аспектов интеллектуальных вычислений.

В заключение отметим, что в современных реалиях культура использования информационных технологий и компьютеров становится частью общей культуры человека, а культура применения ИАД для анализа социологической информации, как нам кажется, в ближайшее время станет одним из показателей профессионализма социолога.

Литература

1. Татарова Г.Г. Методология анализа данных в социологии **(**введение**)**: учебник для вузов. – М.: NOTA BENE, 1999. – 224 с.

2. Давыдов А.А. Системная социология: анализ мультимедийной информации в Интернете [Электронный ресурс]. – Режим доступа: http://www.isras.ru/files/File/Publication/Multimedia_Information_DavydovA.pdf

3. Кислова О. Н. Интеллектуализация информационных технологий как фактор развития интеллектуального анализа социологических данных // Методологія, теорія та

практика соціологічного аналізу сучасного суспільства. Збірник наукових праць. – Харків: видавничий центр ХНУ імені В. Н. Каразіна, 2009. – С. 318-324.

4. Финн В.К. Об интеллектуальном анализе данных // Новости искусственного интеллекта. — 2004. — № 3. — С. 3-18.

5. Загоруйко Н.Г. Прикладные методы анализа данных и знаний. – Новосибирск: Изд. ИМ СО РАН, 1999. – 270 с.

6. Piatetsky-Shapiro G. Knowledge Discovery in Real Databases: A Report on the IJCAI-89 Workshop // AI Magazine. – 1991. – №11(5). – P. 68–70.

7. Толстова Ю.Н. Анализ социологических данных. Методология, дескриптивная статистика, изучение связей между номинальными признаками. – М.: Научный мир, 2000. – 352с.