РОЛЬ И МЕСТО ИНТЕЛЛЕКТУАЛЬНОГО АНАЛИЗА ДАННЫХ В РАЗВИТИИ ЭМПИРИЧЕСКОЙ СОЦИОЛОГИИ

Кислова Ольга Николаевна – кандидат социологических наук, доцент кафедры методов социологического исследования Харьковского национального университета имени В. Н. Каразина

У статті розглянуті проблеми емпіричної соціології, що виникли у зв'язку з інформатизацією суспільства, та проаналізовані можливості інтелектуального аналізу даних (ІАД) у вирішенні цих проблем. Показано, що потужний арсенал інструментів ІАД дає можливість дослідження великих масивів числової, текстової й візуальної соціологічної інформації. Таким чином, роль і місце ИАД у розвитку емпіричної соціології обумовлені насамперед інтелектуалізацією технологій обробки великих обсягів різнорідних даних, пошуку в них актуальної інформації та апріорно не прогнозованих закономірностей.

The article deals with the problems of empirical sociology, which have appeared due to the informatization of society. Possibilities of Intelligent Data Analysis (IDA) in solving these problems have been done. It has been showed that powerful range of IDA gives the possibility to conduct research of large segments of numerical, text and visual sociological information. So, the role and place of IDA in the process of development of empirical sociology have been motivated by intellectualization of the technologies of processing of large volumes of heterogeneous data, search of actual information and not forecasting tendencies.

Ключевые слова: эмпирическая социология, информатизация общества, интеллектуальный анализ данных, Data Mining, Text Mining, Visual Mining.

В последние годы наметилась тенденция переосмысления достижений в области эмпирической социологии. Появилось много учебников, освещающих историю развития эмпирической социологии [1], публикаций, показывающих направления ее потенциального развития (см., например, [2]), а также актуальность поиска и развития методов социологического анализа, адекватных информационной эпохе [3].

Фундаментальными для эмпирической социологии всегда были и остаются проблемы эмпирического метода, соотношения между социальной и эмпирической реальностью, измерения социальных явлений, стандартизации эмпирического языка (см. [2]). При этом следует отметить, что в условиях становления информационного общества изменяются и перспективы развития эмпирической социологии, появляются новые проблемы. Это *актуализирует* необходимость анализа новых, недавно возникших проблем и поиска способов их разрешения.

Цель данной публикации – исследование возможностей интеллектуального анализа данных (ИАД)[1] в контексте эмпирической социологии. Для достижения поставленной цели мы проанализируем научный дискурс этой области и рассмотрим потенциал ИАД в решении выявленных проблем.

Рассуждая о судьбах и перспективах эмпирической социологии, Б. З. Докторов отмечает ключевые трансформации исследований общественного мнения:

1) распространение «догэллаповских» технологий, основанных на многомиллионной рассылке бюллетеней подписчикам изданий, владельцам телефонов и т.п.;

2) развитие «гэллаповских» технологий опросов по сравнительно небольшим репрезентативным выборкам;

3) становление «постгэллаповских» технологий, основанных на изучении общественного мнения при помощи Интернет [2].

Безусловно, новые опросные методы обладают как достоинствами, так и недостатками (см. [3]), вызывают дискуссии, но, тем не менее, используются, тестируются, постепенно становятся привычными и можно предположить, что в ближайшее десятилетие они найдут широкое применение.

Н. И. Лапин отмечает изменение характера эмпирических исследований, что связано со сравнительно недавно открывшейся возможностью свободно использовать архивы социальных данных [1; 2]. Теперь эмпирические исследования можно проводить, используя вторичную социологическую информацию, что обусловливает необходимость освоения методов ее поиска, а также новых способов ее обработки и анализа. Обратим внимание, что в этом контексте особое значение приобретают современные интеллектуальные технологии, позволяющие среди «мусора» разнородных данных находить актуальную в данный момент информацию, а также предоставляют новые возможности ее обработки.

[1] Под интеллектуальным анализом данных (ИАД) понимают процедуру извлечения новых знаний об исследуемом феномене посредством различных формальных методов обработки эмпирических данных с привлечением компьютерных технологий, в частности, интеллектуальных информационных систем.

Известный российский социолог В. А. Ядов среди наиболее актуальных проблем эмпирической социологии выделяет: 1) отставание в области методов и техник сетевого анализа; 2) поддержание профессиональной культуры анализа данных [2].

Как подчеркивает В. Г. Немировский, роль эмпирической социологии в ближайшие годы будет возрастать. Он акцентирует внимание на том, что неклассические и постнеклассические подходы современной социологии порождают новые представления о методах эмпирических исследований, рост внимания социологов к качественным методам, к их взаимосвязи и взаимодействию с методами количественными [2].

Среди наиболее актуальных проблем эмпирической социологии А. В. Тихонов выделяет исследование процессов, устремленных в будущее, поиск способов получения упреждающих знаний об альтернативах и последствиях социальных изменений. Рассуждая о способах решения названных проблем, он пишет: «Этому будут способствовать разработка новых методов и технических средств эмпирических исследований, а, возможно, и пересмотр представлений о социологических методах и процедурах» [2, с. 16]. В данном контексте мы хотим подчеркнуть актуальность исследования социологами эвристических возможностей методов математического и компьютерного прогнозирования, обращение к западным публикациям, где демонстрируются результаты применения этих методов (см., например, [4]).

Анализируя новые проблемы, возникающие в эмпирической социологии в связи с информатизацией общества, А. А. Давыдов подчеркивает необходимость повышения надежности методов сбора информации с использованием Интернет. Он подчеркивает актуальность расширения арсенала методов анализа массивов разнородных данных в Интернете и разработки «интеллектуальных» компьютерных систем, предназначенных для выявления закономерностей в эмпирических данных, для моделирования и прогнозирования [2].

Ю. Н. Толстова считает актуальной проблемой эмпирической социологии использование языка (и аппарата) математики в социологических исследованиях. Она отмечает необходимость корректной постановки содержательных социологических задач, что позволит использовать для их решения адекватные математические методы и модели [6]. При этом она подчеркивает, что следует делать акцент не на «применении метода», а на решении содержательной социологической задачи математическими или компьютерными средствами. «Есть много методов, систем, технологий, созданных для анализа данных, который давно стал «мягким», рассчитанным на человеко-машинный диалог. Но активного использования этих наработок в отечественной социологии не заметно» [2, с. 18].

Мы привели высказывания некоторых известных социологов, касающиеся проблем, возникших в эмпирической социологии в связи с информатизацией общества (оставив вне рассмотрения ее «извечные» проблемы). Некоторые из этих относительно недавно возникших проблем могут быть решены при помощи интеллектуального анализа данных (ИАД), появление которого также обусловлено информатизацией общества и интеллектуализацией средств работы с большими объемами информации.

Возможностям применения ИАД в количественных исследований мы уже посвятили почти два десятка публикаций (в том числе в соавторстве с нашими коллегами [7]). В данной статье мы хотим сконцентрировать внимание на возможностях ИАД в контексте анализа качественной информации, в частности текстовых и графических данных.

В арсенале интеллектуального анализа данных есть множество разнообразных инструментов Text Mining, позволяющих автоматически выделять ключевые понятия, выявлять смысловые структуры в анализируемых текстах и визуализировать их, создавая структурные портреты текстов, что значительно облегчает аналитическую работу социолога и способствует более глубокому анализу сущности исследуемого феномена (см. рисунки 1, 2 и 3).

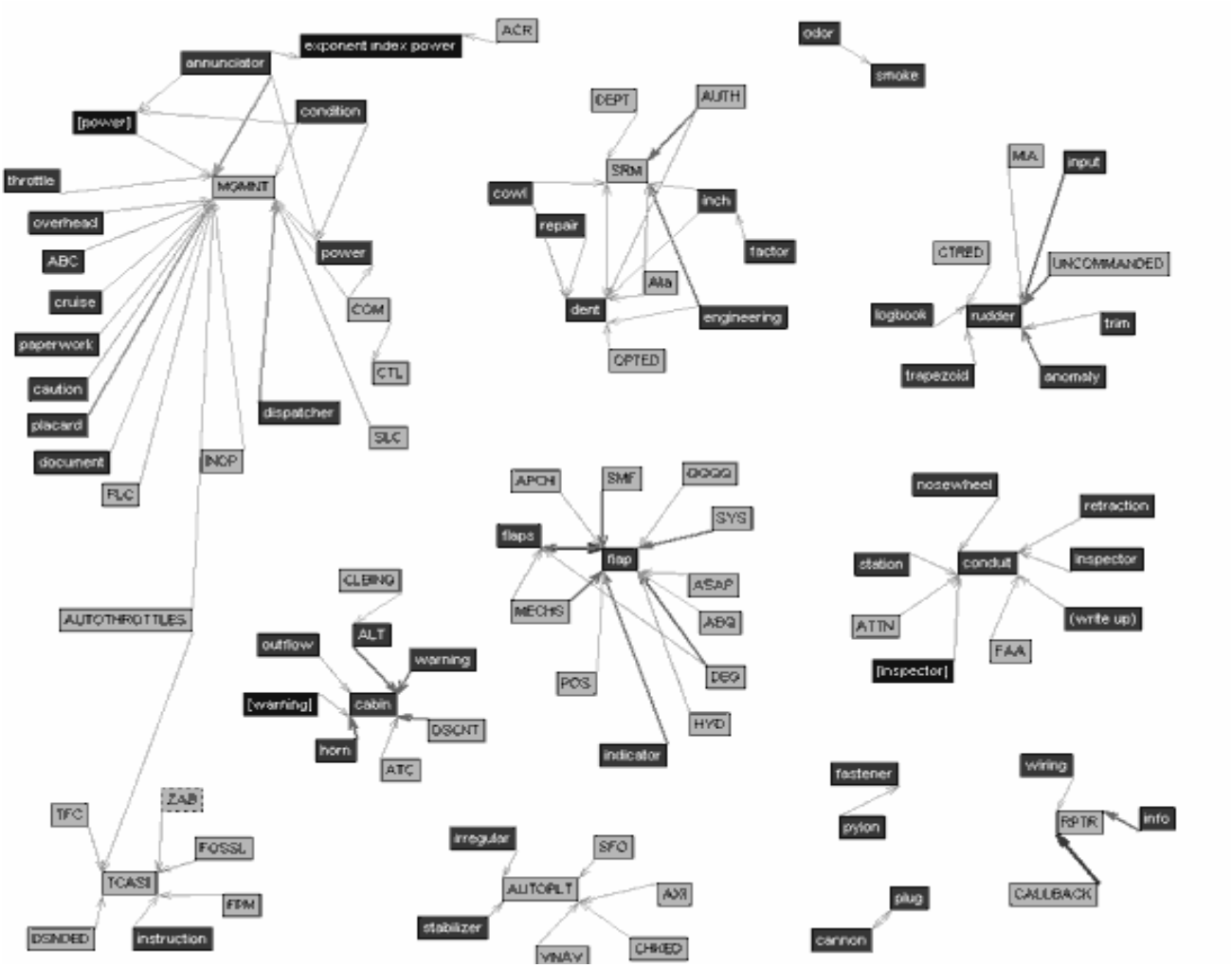

Рисунок 1. **Структурный портрет текста, визуализирующий связи между понятиями методом построения графа взаимосвязей (реализовано при помощи PolyAnalyst 4.6) [8]**

Использование инструментов Text Mining в социологических исследованиях инициирует необходимость тщательного исследования эвристического потенциала разнообразных методов анализа и визуализации текстовых данных, изучения технических и аналитических возможностей отдельных программ, осуществляющих анализ текстов. К сожалению, наши поиски применения технологии Text Mining в социологических исследованиях дали более чем скромные результаты (см. [9]). А ведь в настоящее время существует огромное число разнообразных программ, позволяющих в автоматическом режиме извлекать скрытые закономерности в текстовых данных. Многие из них можно бесплатно скачать в Интернет. Проблема состоит не в доступности инструментальных средств интеллектуального анализа текстов, а в необходимости приложения усилий для освоения этих новых инструментов.

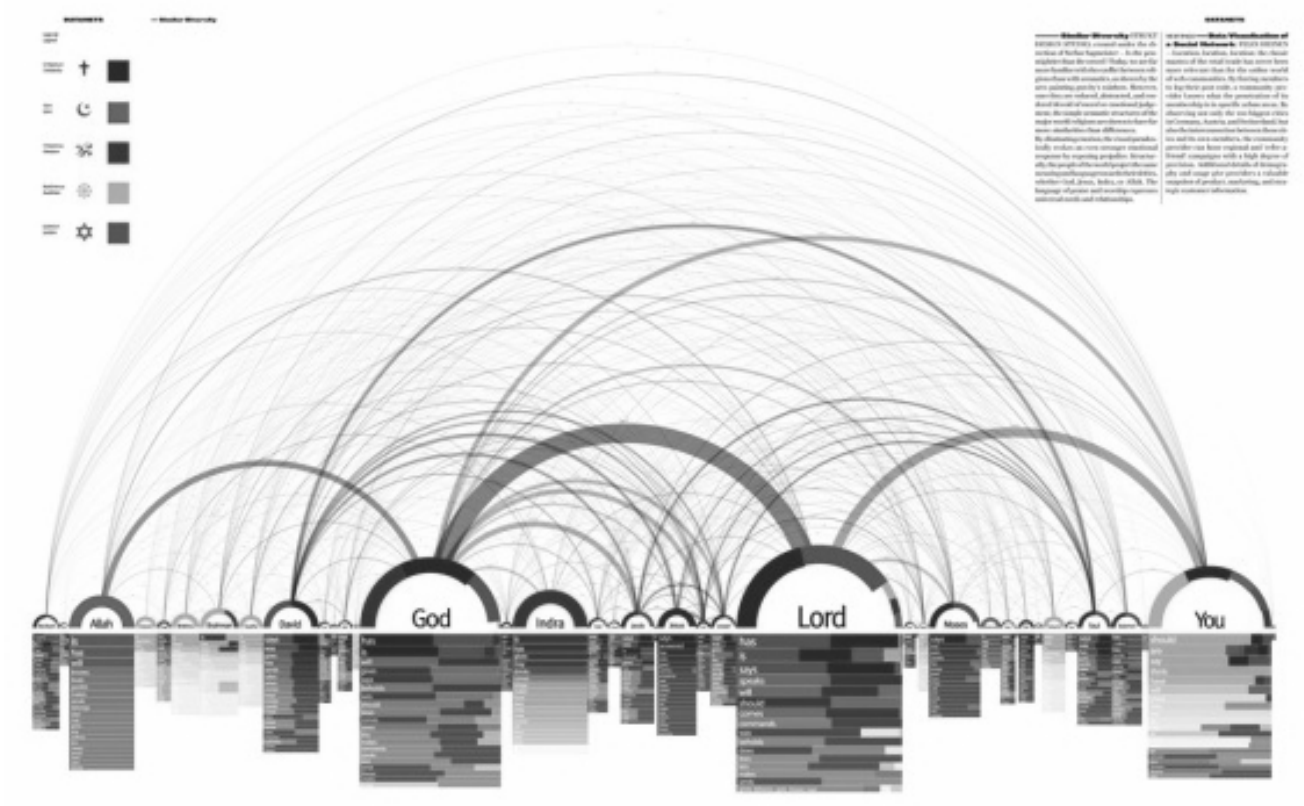

Рисунок 2. **Структурный портрет текста, визуализирующий сходства и различия основных мировых религий (Христианства, Ислама, Индуизма, Буддизма и Иудаизма) в Holy Books [10]**

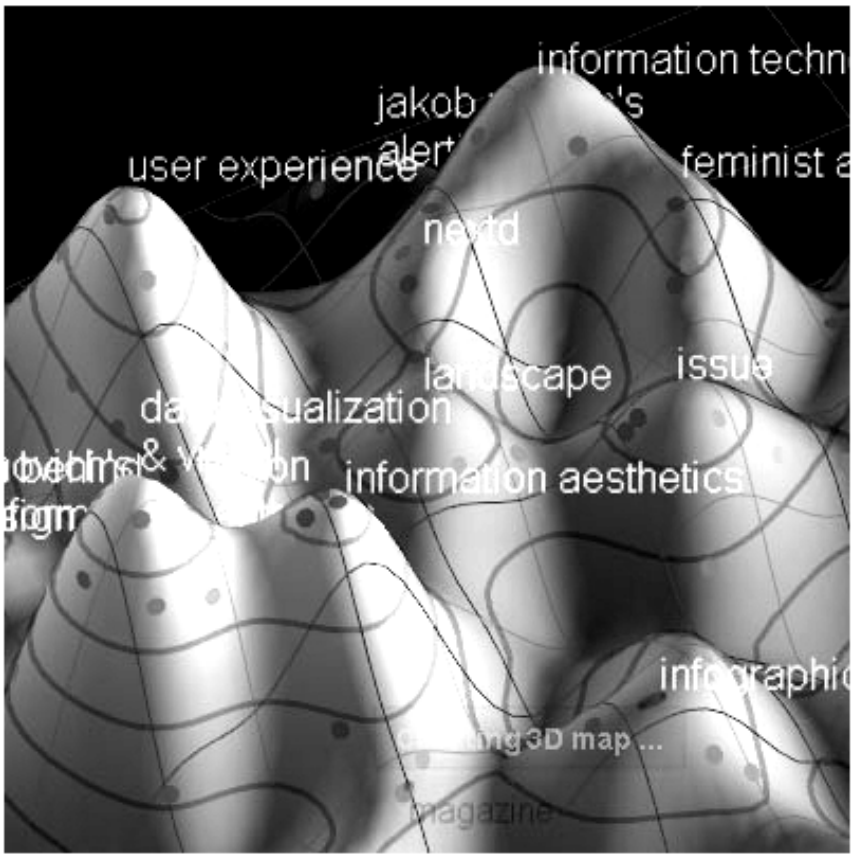

Рисунок 3. **Структурный портрет текстовой информации Интернет, построенный методом «горной» визуализации кластеров, объединяющих тексты по тематическому содержанию (реализовано при помощи Grokker) [11]**

В настоящее время возникает потребность в социологическом анализе визуальной информации. П. Штомпка предполагает, что фотографии могут выступать в качестве эмпирических данных при исследовании разнообразных социальных феноменов [12]. В этом контексте весьма полезными окажутся интеллектуальные инструменты Visual Mining, которые в автоматическом режиме способны выявлять структуры (закономерности) в анализируемых графических образах (естественно, предварительно фотографии необходимо представить в формате, пригодном для компьютерной обработки). Обращая внимание на возможности Visual Mining, нам представляется необходимым отметить, что терминология в сфере визуального анализа эмпирических данных еще не устоялась. В литературе термины Visual Mining и Graph Mining часто используются как синонимы и могут обозначать две совершенно разные процедуры:

- анализ визуальной информации (т.е. анализ произвольных изображений, например, фотографий);
- визуализацию числовых данных в виде графиков, диаграмм или когнитивных образов.

Мы предлагаем использовать понятие Visual Mining для анализа графического эмпирического материала, а термином Graph Mining обозначать процедуру визуализации эмпирических данных, осуществляемую для их аналитического исследования. При этом следует учитывать, что средства Graph Mining могут использоваться как в количественных исследованиях (см., например, [13]), так и в качественных [9].

Таким образом, можно увидеть, что ИАД дает возможность осуществлять обработку ***любых данных***: числовых, текстовых, графических. Для каждого типа социологической информации в арсенале ИАД есть свои инструменты:

- ✓ алгоритмы Data Mining позволяют извлекать закономерности из числовых данных и применяются в количественных исследованиях;
- ✓ процедуры Text Mining могут быть очень полезны для обработки результатов качественных исследований;
- ✓ средства Visual Mining, созданные для анализа графической информации, могут использоваться в ходе социологического анализа фотографических данных.

Безусловно, при анализе текстовой и визуальной информации возникает масса вопросов, на которые ИАД не дает ответа (например, вопросы, связанные с формированием выборки). Однако его назначение иное – выявление новых знаний в массивах эмпирических данных. В этом направлении ИАД предоставляет массу интересных возможностей, использование которых расширяет возможности социологического анализа.

По нашему мнению, ИАД может сыграть определенную роль в становлении «постгэллаповских» технологий изучения общественного мнения при помощи Интернет. Примером тому служит использование американскими аналитиками интеллектуальных средств когнитивной визуализации при изучении политической блогосферы [14] с целью определения рейтинга основных политических сил США перед президентскими выборами 2008 года (см. рисунок 4).

Таким образом, отдельные методы ИАД уже показали свою полезность в сфере электоральных исследований, что, как мы надеемся, явится стимулом к использованию интеллектуального анализа данных и украинскими социологами.

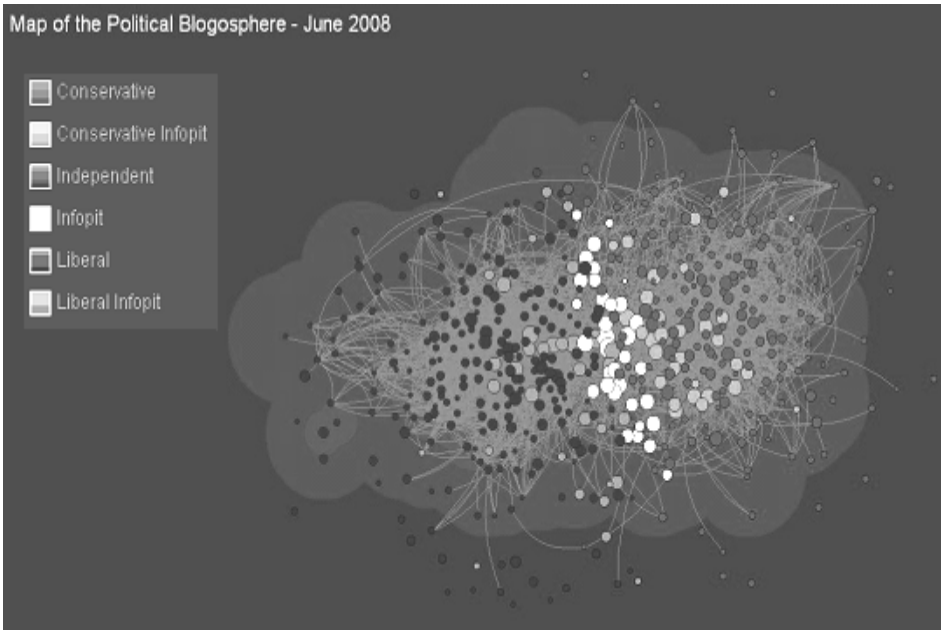

Рисунок 4. **Визуализация текстовой информации в политической блогосфере США (Выборы Президента США 2008 г.). Источник: [14].**

Проблема отставания эмпирической социологии в области методов и техник сетевого анализа, отмеченная В. А. Ядовым, безусловно, не может быть решена только технологическими средствами ИАД, поскольку необходимо теоретическое осмысление потенциала сетевого анализа в социологии. Работа в этом направлении идет, получены очень интересные результаты [15]. Однако мы считаем, что необходима также и популяризация инструментальных средств, позволяющих реализовать сетевой анализ на массиве социологических данных. Это даст возможность социологам активнее использовать данный метод. Мы хотим обратить внимание на некоторые публикации, знакомящие с эвристическим потенциалом Data Mining при проведении анализа социальных сетей (см., например, [16]).

Необходимость повышения уровня профессионализма социологов-аналитиков, отмечается многими социологами. При этом предполагается именно профессионализм в проведении социологических исследований, включающий в себя навыки и умения в таких областях как операционализация понятий, разработка инструментария, расчет выборки, анализ эмпирических данных и др. Приобретение таких навыков неразрывно связано с мотивацией изучения и последующего использования языка формализации (как в количественных, так и в качественных исследованиях) и языка математики (в количественных исследованиях). Как мотивировать социологов-гуманитариев изучить математический язык и методы (иногда достаточно сложные) анализа данных? Нам представляется, что именно ИАД может быть решением данной проблемы, поскольку он значительно облегчает использование сложного математического инструментария.

Современные системы компьютерной обработки данных освобождают социолога от рутинных вычислений. Но это не в полной мере решает проблему «страха» гуманитариев перед математическими методами, поскольку все равно в ходе интерпретации приходится оперировать различными, не всегда понятными статистическими показателями. Поэтому в настоящее время в сфере искусственного интеллекта особо акцентируется значимость развития методов визуализации. Интеллектуальные системы анализа информации создаются с ориентацией на дружественный интерфейс, максимальное упрощение диалога между человеком и компьютером, а результаты выводятся не только в виде объемных таблиц, содержащих разнообразные «страшные» статистические показатели, но в графическом виде, наглядно демонстрирующем результаты проведенного анализа.

Когда результат представлен наглядно, так, что его можно увидеть, бросив всего один взгляд, возникает ощущение легкости применения сложных методов, но при этом встает вопрос об интерпретации. Вот здесь и понадобятся знания о том, что граф взаимосвязи понятий (рис. 1) является визуальным аналогом корреляционного анализа. Интерпретируя картину «горной» местности легче понять сущность кластерного анализа, значимость понятия «внутрикластерная дисперсия»[2] для описания полученных результатов. Изучить все эти «сложные» термины гораздо проще, когда есть понимание того, где они применяются. Таким образом, визуализация результатов анализа данных способствует не только быстрому решению поставленных социологом содержательных задач, но и мотивирует изучить те показатели, которые способствуют интерпретации результатов, описывая сущность исследуемых социальных явлений, позволяют оценить качество полученных моделей.

В этом контексте следует отметить, что не существует одного универсального метода, который давал бы решение всех задач, возникающих в процессе анализа социологических данных. В каждом конкретном случае следует выбирать методы, адекватные как исследуемому социальному феномену, так и эмпирической информации, характеризующей его. Именно поэтому мы считаем, что необходимо популяризировать эвристические возможности различных (особенно возникших недавно) методов анализа данных, области их применения, достоинства и недостатки, а особенно примеры применения в конкретных социологических исследованиях.

В современном технизированном мире развитие науки тесно связано с использованием новых информационных технологий. Успехи в области искусственного интеллекта определили вектор дальнейшего технологического развития, инициировали процесс интеллектуализации технических средств, что, в частности, привело к становлению новой парадигмы познания – интеллектуального анализа данных (ИАД). Извлечение новых знаний об исследуемом объекте, породившем данные, непосредственно из этих данных – это процесс, который невозможно осуществить без идей и средств искусственного интеллекта. Основная идея ИАД состоит в том, что данные хранят информацию, невидимую под привычным углом зрения. Чтобы увидеть нечто «новое» в массиве эмпирических данных необходимо изменить способ «видения». ИАД для этого предлагает принцип вариативного моделирования, реализующийся в построении и совместном применении не менее двух разных моделей исследуемого объекта.

Вариативное (от англ. variety – разнообразие, многосторонность) моделирование – это метод исследования, основанный на замене изучаемого объекта-оригинала набором разнообразных моделей,

[2] Визуальный аналог этого показателя – высота горы. Чем ниже внутрикластерная дисперсия (т.е. меньше различия между объектами одного кластера), тем выше «гора», визуализирующая данный кластер.

исследование которых дает многогранную информацию об исходном объекте. В контексте анализа социологических данных применение вариативного моделирования означает рассмотрение данных сквозь призму, например, факторной и кластерной моделей. Интерпретация результатов совместного применения этих методов дает значительно больше информации об исследуемом феномене, чем применение каждого метода отдельно. Кроме того, использование нескольких методов решения одной и той же задачи позволяет осуществить триангуляцию. Так, например, использование дискриминантного анализа для проверки результатов кластерного анализа представляет данные в виде совершенно разных моделей. Если результаты использования этих моделей дают похожие результаты, то, соответственно, доверие к результатам возрастает. Таким образом, осуществляется проверка результатов непосредственно в ходе анализа данных. Чем больше методов (и соответствующих им моделей) применяется, тем лучше становятся «видны» скрытые в данных закономерности.

Метафорой, приблизительно поясняющей, что происходит в процессе ИАД, могут служить стереограммы, которые, на первый взгляд, выглядят как орнамент, набор повторяющихся немного искаженных изображений. Однако расслабленное созерцание стереограмм приводит к одновременному включению нескольких углов зрения, благодаря чему сретеограмма начинает видеться как объемная картинка, показывающая совершенно неожиданные изображения. Так и данные в процессе ИАД дают неожиданные результаты, показывают априорно не прогнозируемые закономерности.

Итак, мы очертили те возможности ИАД, которые способствуют решению некоторых проблем эмпирической социологии, возникших в процессе информатизации. В то же время мы хотим отметить, перечисленное не является исчерпывающим описанием всего потенциала ИАД в социологии.

ИАД предоставляет социологу новые возможности, которые полезны для решения не только задач эмпирической социологии, но также могут способствовать развитию теоретического социологического знания. Так, например, методы computational sociology (в частности, имитационное моделирование) позволяют осуществлять верификацию классических социологических теорий и конструировать новые [17]. При этом мы с большим сожалением вынуждены констатировать, что в нашей стране работы такого формата пока не проводятся. Изучение опыта зарубежных коллег, возможно, мотивирует молодых социологов освоить современные технологии научного познания, адекватные информационной эпохе, а внедрение ИАД в практику социологического анализа даст новые интересные результаты.

Литература:

1. Беляева Л.А. Эмпирическая социология в России и Восточной Европе. – М.: ГУ ВШЭ, 2004. – 406 с.; Ионин Л.Г. Философия и методология эмпирической социологии. М.: ГУ ВШЭ, 2004. – 267 с.; Лапин Н.И. Эмпирическая социология в Западной Европе. – М.: ГУ ВШЭ, 2004. – 381 с.; Анурин В.Ф. Эмпирическая социология. – М.: Академический проект, 2003. – 288 с.
2. Беляева Л.А., Давыдов А.А., Данилов А.Н., Докторов Б.З., Лапин Н.И., Левашов В.К., Немировский В.Г., Тихонов А.В., Толстова Ю.Н., Тощенко Ж.Т., Ядов В.А. Судьбы и перспективы эмпирической социологии // Социологические исследования. 2005. – № 10. – С. 3-21.
3. Чураков А.Н. Информационное общество и эмпирическая социология // Социологические исследования. – 1998. – № 1. – С. 35-44; Филиппова Т. Эмпирическая социология в информационном обществе. "Web-опросы в России - "за" и "против". – Доступно на: http://www.isn.ru/info/seminar-doc/soc.doc; Хитров А. Блог как феномен культуры // Журнал социологии и социальной антропологии. – 2007. – Т. 10. – Спецвыпуск. – С. 66-76; Давыдов А.А. Социология изучает блогосферу // Социологические исследования. – 2008. – № 11. – С. 92-101; Качанов Ю.Л. Теоретические предпосылки эмпирического исследования социологической теории // Социологические исследования. – 2000. – № 10. – С. 3-10; Докторов Б.З. К попытке определения пространства американских методических исследований опросных технологий // Социология. – 4М. – 2005. – № 20. – С. 10-31;
4. Simulating Societies: The Computer Simulation of Social Phenomena/ editors N. Gilbert, J. Doran. – London: UCL Press, 1994; Gilbert N., Troitzsch K. Simulation for the Social Scientist. Buckingham. – UK: Open Univ Press, 1999.
5. Давыдов А.А. Системный подход в социологии: новые направления, теории и методы анализа социальных систем. – М.: КомКнига, 2005. – 324 с.
6. Толстова Ю.Н. Математико-статистические модели в социологии. – М.: ИД ГУ-ВШЭ, 2007. – 243 с.
7. Кислова О.Н. Интеллектуальный анализ данных: возможности и перспективы применения в социологических исследованиях // Методологія, теорія та практика соціологічного аналізу сучасного суспільства. Збірник наукових праць. – Харків: Видавничий центр Харківського національного університету імені В.Н. Каразіна, 2005. – С. 237-243; Сокурянська Л.Г., Кислова О.М. Використання інтелектуального аналізу даних у дослідженнях феномену знехтування у студентському середовищі // Український соціум. – № 3-4 (14-15). – 2006. – С.65-76; Кислова О.М., Ніколаєвська А.М. Досвід застосування технології інтелектуального аналізу даних (ІАД) при вивченні моральних феноменів // Вісник Одеського національного університету. – Том. 12. – Випуск 6. – Серія «Соціологія і політичні науки». – 2007. – С. 621-628; Сокурянская Л.Г., Кислова О.Н. Ценностный мир постсоветского студенчества: результаты применения методов интеллектуального анализа данных // Социология. – №3. – 2008. – С. 88-100.
8. http://www.bitconsulting.ru/article/pa/pa_alg.shtml

9. Винокурова И.С. Применение компьютера для контент-анализа социологической информации // Методы социологических исследований. – М.: ТЕИС, 2006. – С. 83-109; Давыдов А.А. Информационный дизайн в Visual Text Analytics – инструмент системного социолога. – доступно на: http://www.isras.ru/index.php?page_id=968
10. http://similardiversity.net
11. http://www.grokker.com
12. Штомпка П. Визуальная социология. Фотография как метод исследования. – М.: Логос, 2007. – 168 с.
13. Кислова О.Н. Когнитивная визуализация как инструмент интеллектуального анализа социологических данных // Вісник Харківського національного університету імені В.Н. Каразіна "Соціологічні дослідження сучасного суспільства: методологія, теорія та методи". – № 795. – 2008. – С. 78-84; Кислова О.Н. Визуализация социологических данных как альтернатива традиционным методам дескриптивного анализа // Методологія, теорія та практика соціологічного аналізу сучасного суспільства. Збірник наукових праць. – Харків: Видавничий центр Харківського національного університету імені В.Н. Каразіна, 2008. – С. 142-152.
14. http://presidentialwatch08.com/index.php/map
15. Горбачик А., Жулькевська О. Мережевий підхід до вивчення структури українського парламенту // Соціологія: теорія, методи, маркетинг. – 2006. – №3. – С. 161-181.
16. Jensen D, Neville J. Data Mining in Social Networks. – доступно на: http://kdl.cs.umass.edu/prople/jensen/papers/nas02/pdf; Wil M., van der Aalst, Song M. Mining Social Networks: Uncovering Interaction Patterns in Business Processes. – доступно на: http://psim.tm.tue.nl/staff/wvdaalst/
17. Hanneman R. Simulation Modeling and Theoretical Analysis in Sociology// Sociological Perspectives. – 1995. – Vol 38. – №4. – P. 457-462; Sawyer R. Artificial Societies: Multi agent systems and the micro-macro link in sociological theory//Sociological Methods and Research. – 2003. – V31. – №3. – P. 325-363; Moss S., Sawyer, R. K., Conte, R., Edmonds B. Sociology and social theory in agent based social simulation: A symposium// Computational and Mathematical Organization Theory. – 2001. – Vol 7. – №3. – P.183-205.