

**И.А. Землянская, Самарский университет;
П.А. Грибенникова, Самарский университет**

ПРИМЕНЕНИЕ МЕТОДА DATA MINING В СОЦИОЛОГИЧЕСКОМ ИССЛЕДОВАНИИ

В данной статье рассматриваются возможности и область использования метода data mining в социологии. Авторы раскрывают специфику как самого метода data mining, так и его разновидности – text-mining, описывают основные виды данных, подлежащих обработке и анализу, классифицируют методы анализа текстовых данных, а также анализируют опыт использования метода data mining в социологии.

Ключевые слова: метод data mining, возможности data mining, опыт использования data mining в социологии, text mining.

**I.A. Zemlyanskay, Samara University;
P.A. Gribennikova, Samara University**

APPLICATION OF THE DATA MINING METHOD IN SOCIOLOGICAL RESEARCH

This article discusses the possibilities and scope of using the data mining method in sociology. The authors reveal the specifics of both the data mining method itself and its variety – text-mining, describe the main types of data to be processed and analyzed, classify text data analysis methods, and analyze the experience of using the data mining method in sociology.

Keywords: data mining method, data mining capabilities, experience of using data mining in sociology, text-mining.

Классическое определение термина data mining дал один из основателей данного направления – Григорий Пятецкий-Шапиро в 1996 г.: data mining – исследование и обнаружение «машиной» (алгоритмами, средствами искусственного интеллекта) в сырых данных скрытых знаний, которые ранее не были известны, нетривиальны, практически полезны, доступны для интерпретации человеком [1, с. 68].

Методология data mining является мультидисциплинарной областью, которая возникла и развивается на базе многих дисциплин, таких как программирование, искусственный интеллект, прикладная статистика, теория баз данных и др.

Знание, извлеченное в результате интеллектуального анализа данных, принято называть «паттерн» [2, с. 144]. Основная цель технологии data mining заключается в выявлении значимых закономерностей в обширных массивах информации.

Английское словосочетание «data-mining» все еще не имеет точного и устоявшегося перевода на русский язык, используют такие словосочетания как: добыча данных, просев информации, извлечение данных и, конечно же, интеллектуальный анализ данных. В англоязычной литературе также можно встретить термин Knowledge Discovery in Databases (KDD) – «обнаружение знаний в больших базах данных», данное определение более точно и полно раскрывает суть самого метода [3, с. 5].

В отличие от традиционных методов анализа, data mining использует автоматизированные алгоритмы, что делает данный метод достаточно сильным инструментом, позволяющим исследователям подвергать анализу большие объемы информации.

Все данные, анализируемые с помощью метода data-mining, делятся на структурированные (обычно хранятся в базах данных) и неструктурированные (текстовые).

Text-mining или извлечение информации из неструктурированных данных – это текстовое направление интеллектуального анализа data-mining. Данный метод включает в себя дополнительный этап – преобразование неструктурированного текста в структурированный массив данных, что позволяет применять к нему стандартный набор методов data mining [4].

Методы анализа текстовых документов подразделяются на две категории: лингвистические – обработка естественного языка (NPL) и статистические – частота встречающихся слов [5, с. 789].

В последнее время социологи все чаще обращаются к методу data mining или его разновидности (text-mining) в своих исследованиях. В частности, text-mining, используют в маркетинговых исследованиях при анализе текстов рекламных сообщений; в социологических исследованиях при анализе социальных сетей, мессенджеров, форумов и др. площадок виртуального пространства; при изучении больших объемов текстов, посвященных определенной тематике [6, с. 163].

В данной связи показателен пример работы, проведенной В.И. Дудиной и Д.И. Юдиной, в которой демонстрируется практическое применение текстового анализа пользовательских комментариев для изучения общественного мнения. В условиях, когда скорость изменения общественной повестки нередко превышает темпы проведения традиционных социологических исследований, автоматизированный сбор и обработка текстов становятся особенно ценными инструментами для социологов. Исследование, проведенное на основе анализа пользовательских реакций на антикоррупционный фильм «Чайка», размещенный на платформе YouTube, позволило авторам применить инструменты тематического моделирования и автоматического извлечения тональности с помощью словарей оценочной лексики для выявления эмоциональной окраски и основных тематических направленностей комментариев. Каждый фрагмент текста был автоматически интерпретирован по шкале эмоциональной окраски при помощи словаря, разработанного на базе НИУ ВШЭ, что позволило сформировать количественные показатели восприятия основных политических фигур, упомянутых в фильме [7, с. 71–72].

Сравнение данных, полученных посредством автоматического анализа, с результатами опросов Левада-Центра показало близкие по характеру оценки, что позволяет говорить о потенциале подобных методов анализа данных в качестве инструмента предварительной оценки общественного мнения [7, с. 68]. Однако, как подчеркивают исследователи, применение подобных алгоритмов требует учета культурно-языковой специфики, а также контекстуальных факторов, включая иронию, сарказм и сленг, которые могут исказить интерпретацию текста [7, с. 73–74].

Другим примером выступает исследование О.Ю. Кольцовой и К.А. Маслинского, которое направлено на выявление тематической структуры российской блогосферы. Авторы демонстрируют, как при помощи алгоритмов машинного обучения и тематического моделирования можно осуществить автоматическую классификацию текстов, что является особенно важным в условиях высокой динамики интернет-контента. Работа авторов наглядно показывает, что современные методы анализа текста позволяют обнаруживать устойчивые тематические кластеры в блогах и форумах, которые формируют ядро общественной дискуссии в интернет-пространстве [8, с. 52–53].

Продолжая рассмотрение прикладных примеров использования методов *data mining* в социологических исследованиях, важно отметить работу М.А. Кашиной и С. Ткача, в которой демонстрируется применение технологий анализа естественного языка (Natural Language Processing, NPL) для систематиза-

ции научного знания. Особенностью данного исследования является то, что оно направлено на изучение не массовых коммуникаций или пользовательских реакций, а на академический дискурс, в частности на тексты аннотаций научных статей, посвященных социологии ценностей [9, с. 48–49].

Анализ более чем сотни аннотаций, которые были отобраны из высокорейтинговых социологических журналов, был проведен при помощи библиотек Python – NLTK, Scikit-learn и SciPy. На первом этапе данные были подвержены машинной кластеризации методом Варда, что позволило выделить 33 тематических кластера. Из них 11 наиболее репрезентативных были отобраны для последующего качественного анализа [9, с. 50–52]. В полученных кластерах были отражены ключевые теоретические и прикладные направления исследований: это гендер, миграция, здоровье, рынок труда, глобализация и другие. Авторы акцентируют внимание на том, что использование подобных подходов позволяет не только идентифицировать наиболее популярные темы, но и выявить малоразвитые или «молчащие» области академического поля, которые не получают достаточного внимания.

Список литературы:

1. Анализ данных и процессов: учеб. пособие / А.А. Барсегян, М.С. Куприянов, И.И. Холод, М.Д. Тесс, С.И. Елизаров. 3-е изд., перераб. и доп. СПб.: БХВ-Петербург, 2009. 512 с.
2. Мосягин А.Б. Использование методологии Data Mining при решении задач обработки социальных данных // Мониторинг. 2015. № 3 (127). С. 143–145.
3. Степанов Р.Г. Технология Data Mining: Интеллектуальный Анализ Данных. Казань. 2008. 57 с.
4. Галина А.В., Есина Е.А. Обзор технологии Text Mining // Аллея Науки: научно-практический электронный журнал. 2018. № 1 (17). С. 393–396.
5. Каримов Р.И. NLP – обработка естественного языка. лингвистический метод обработки текста // Мавлютовские чтения. Уфа: Изд-во УГАТУ, 2021. № 3. С. 250–254.
6. Процедуры и методы социологического исследования. Методы анализа данных в качественном социологическом исследовании: практикум / А.С. Готлиб, В.Л. Лехциер, Ю.В. Епанова [и др.]; под общ. ред. А.С. Готлиб. Самара: Издательство Самарского университета, 2024. 298 с.

7. Дудина В.И., Юдина Д.И. Извлекая мнения из сети Интернет: могут ли методы анализа текстов заменить опросы общественного мнения? // Мониторинг. 2017. № 5 (141). С. 63–77.

8. Кольцова О.Ю., Маслинский К.А. Выявление тематической структуры российской блогосферы: автоматические методы анализа текстов // Социология: методология, методы, математическое моделирование (Социология: 4М). 2016. № 36. С. 113–139.

9. Кашина М.А., Ткач С. Социология ценностей: опыт построения таксономии путем использования технологии анализа естественного языка // Цифровая социология. 2023. № 1. С. 48–58.