МАТЕМАТИЧЕСКИЕ И ВЫЧИСЛИТЕЛЬНЫЕ МЕТОДЫ МНОГОМЕРНОГО АНАЛИЗА В ПРОСТРАНСТВЕ СОЦИАЛЬНЫХ ХАРАКТЕРИСТИК НА БАЗЕ БОЛЬШИХ ДАННЫХ

Каменев И.Г.
(*ФИЦ «Информатика и управление» РАН, Москва; НОЧУ ВО «Синергия», Москва*)
Андрианова Д.А.
(ООО «Яндекс», Москва)

В докладе рассмотрена научная проблема автоматизированной обработки больших данных (big data) социального характера. Для ее решения предлагается комплексный Метод метрического анализа данных Method of Metric Data Analizis (MMDA).

В настоящей работе под большими данными понимаются массивы информации естественного происхождения, которые невозможно обработать и представить стандартными методами визуализации из-за большого объема. Требуется, чтобы массив данных имел постоянный объем и количественные характеристики. Данные рассматриваются как элементы многомерного евклидова пространства Rd, т.е. как записи, содержащие d численных компонент x=(x1, ..., xd) из Rd в метрике Чебышевского.

Подготовка таких массивов данных к исследованию их метрических свойств, основных частей и характерных составляющих включает в себя приведение всех характеристик к числовым параметрам: количественным, бинарным, порядковым. Возможно приведение некоторых количественных характеристик к агрегированным показателям для удобства визуализации.

В случае если полученный массив данных X не совпадает с генеральной совокупностью G, мы предполагаем, что множество G ограничено по объему. Тогда помимо задачи метрического анализа данных (метод MMDA) встает задача исследования надежности выводов о свойствах генеральной совокупности G по свойствам совокупности данных X.

Размеры массива big data не позволяют эффективно осуществлять анализ, поэтому решается задача их аппроксимации. Для этого используется стохастическая реализация Метода глубоких ям (МГЯ), позволяющая строить (e,d)-покрытия с с заданными характеристиками надежности. Данная методика разработана на базе (эпсилон, дельта)-сетей и (эпсилон, дельта)-покрытий К. Э. Шеннона, методов аппроксимации Г.К. Каменева и А.В. Лотова.

Скорость сходимости МГЯ определяется e-емкостью аппроксимируемого множества, которая при высокой точности определяется метрической размерностью. Кроме того, на практике МГЯ удобен возможностью приостанавливать аппроксимацию для быстрого получения частичных результатов анализа.

Данные методы позволяют выделить метрическую структуру компонент, изучить их формы и границы, выделить сгущения, провести кластерный анализ на основе выделения топологических компонент e-покрытий.

В качестве метода исследования в докладе используется визуализация массива big data с помощью Диалоговых Карт Решений (ДКР): динамических интерактивных 2D и 3D карт и атласов многомерных e-покрытий, а также алгоритмическое выделение топологических компонент e-покрытий. Все теоретические построения в докладе иллюстрируются примерами реализации методов на массиве big data, доступном в открытых источниках (данные о поездках таксомоторного парка), что обеспечивает потенциальную воспроизводимость метода.

Литература

1. КАМЕНЕВ Г.К. Аппроксимация вполне ограниченных множеств методом Глубоких Ям // Ж. вычисл. матем. и матем. физ. 2001. Т. 41. №11. С. 1751-1760.
2. Taxi Service Trajectory // Taxi-link Company, 2015. – URL: http://geolink.pt/ecmlpkdd2015-challenge/dataset.html (дата обращения: 01.01.2018)