

5. Есиков, Д. О. Применение теории оптимальности по Парето для обоснования требуемых значений параметров надёжности функционирования комплексов средств хранения информации / Д. О. Есиков // Тр. Рос. науч.-техн. о-ва радиотехники, электроники и связи им. А. С. Попова, выпуск XXXII. — Тула : ТулГУ, 2014 — 325 с.
6. Кристофидес, Н. Теория графов. Алгоритмический подход / Н. Кристофидес. — М. : Мир, 1978. — 430 с.
7. Есиков, Д. О. Математическая модель обеспечения сохранности информации в распределённых информационных системах за счёт резервирования информационных массивов. / Д. О. Есиков // Тр. Рос. науч.-техн. о-ва радиотехники, электроники и связи им. А. С. Попова, выпуск XXXII. — Тула : ТулГУ, 2014. — 325 с.
8. Алексеев, О. Г. Комплексное применение методов дискретной оптимизации / О. Г. Алексеев. — М. : Наука, 1987. — 248 с.

УДК 004.422.632.001.2

Г. А. Искакова

Евразийский национальный университет имени Л. Н. Гумилева, Астана

МЕТОД КЛАСТЕРИЗАЦИИ АНАЛИЗА И ОБРАБОТКИ БОЛЬШИХ ДАННЫХ, ПРЕДСТАВЛЕННЫХ В МАТРИЧНОЙ ФОРМЕ

Введение. Парадигма больших данных является мировым трендом в области информационных технологий и к настоящему времени мало изучена.

Результаты исследований в области технологии больших данных весьма востребованы как для отраслей мировой экономики, так и для реальных секторов экономики Казахстана.

Сегодня ни одна отрасль и ни одна организация не могут эффективно функционировать без достижений в области информационных технологий. Курс на экономическую индустриализацию страны и максимальное удовлетворение спроса социальной сферы определяют новые вызовы и требования к информационным технологиям и обуславливают необходимость повсеместного внедрения самых современных технологий в социально-экономические системы, к которым, безусловно, относятся технологии анализа и обработки больших данных.

В существующих технологиях и продуктах, разработанных мировыми вендорами, используемые методы являются, как правило, конфиденциальными и вложены в соответствующие коммерческие инструменты.

Одной из основных проблем в использовании больших данных является анализ огромных массивов данных с целью извлечения полезной информации для принятия решения.

В настоящее время для анализа и обработки больших данных необходимы разработка теоретических основ, методов и алгоритмов для создания эффективных инструментов для их применения. В связи с этим указанные проблемы весьма актуальны.

Основная часть. В настоящее время ряд вендоров имеют разработки в области анализа и обработки BigData. Вместе с тем они не получили широкого распространения для эффективного их применения в процессе принятия решений и управления. В связи с этим возникает необходимость разработки новых подходов и методов для анализа и обработки BigData.

Одним из подходов анализа и обработки больших данных является использование матричного представления данных, матричного исчисления и методов оптимизации и на этой основе разработка соответствующих инструментальных средств. Приведём основные задачи анализа и обработки больших данных:

- анализ методов и технологий обработки больших данных;
- анализ особенностей и свойств структурированных и не структурированных данных большого объема и многообразия;
- разработка и исследование методов анализа больших данных, размещённых в хранилищах и представленных в матричной форме;
- разработка методов кластеризации неструктурированных, структурированных и гибридных данных для их эффективного представления;
- разработка новых подходов, методов и средств обработки больших данных;
- разработка инструментальных средств анализа и обработки больших данных;
- разработка базовых методов и моделей анализа и кластеризации на основе матричного исчисления;
- разработка методологических основ и рекомендаций для применения разработанных подходов, методов и инструментальных средств анализа и обработки больших данных в секторах национальной экономики.

Рассмотрим подход, модели и методы кластеризации больших данных, представленных в матричной форме для их анализа и обработки. Отметим, это большие данные во многих случаях удобно представить в матричной форме для последующего анализа. При этом под данными можно понимать отдельные атрибуты, массивы данных, базы данных и данные в структурированных и слабо структурированных форматах.

Постановка задачи. Предположим, что данные большого объёма представлены в матричной форме в виде одной матрицы или совокупности матриц. Сформулируем следующую задачу. Пусть, $Z = \{z_i; i = \overline{1, I}\}$ — множество запросов поиска и извлечения данных, $D = \{d_j; j = \overline{1, J}\}$ — множество атрибутов различных форматов для удовлетворения запросов. Матрица $A = \|a_{ij}\|, i = \overline{1, I}; j = \overline{1, J}$ отражает необходимость поиска и извлечения j -го атрибута и матрицы больших данных для удовлетворения i -го запроса. Необходимо осуществить декомпозицию множество запросов и используемых ими данных таким образом, чтобы минимизировать взаимосвязи по данным между финальными кластерами запросов [1].

Для математической постановки задачи введём следующие переменные и обозначения.

Пусть $x_{mi} = \begin{cases} 1, \text{если } i\text{-й запрос распределяется в } m\text{-й кластер} \\ 0, \text{в противном случае} \end{cases}$,

$y_{jn} = \begin{cases} 1, \text{если } j\text{-запрос распределяется в } n\text{-й кластер} \\ 0, \text{в противном случае} \end{cases}$.

Тогда целевую функцию кластеризации запросов и данных можно сформулировать следующим образом:

$$\sum_{m=1}^M \sum_{i=1}^I \sum_{j=1}^J \sum_{n=1}^N x_{mi} a_{ij} y_{jn} \rightarrow \min.$$

При ограничениях на:

– дублирование запросов в кластерах

$$\sum_{m=1}^M x_{mi} = 1, i = \overline{1, I};$$

– число запросов в каждом кластере запросов

$$\sum_{i=1}^I x_{mi} \leq p_m; \quad m = \overline{1, M};$$

– дублирование данных для удовлетворения запросов

$$\sum_{n=1}^N y_{jn} = 1, \quad j = \overline{1, J};$$

– число данных в каждом кластере данных

$$\sum_{j=1}^J y_{jn} \leq S_n, \quad n = \overline{1, N}.$$

Сформулированная задача относится к блочно-симметричным задачам дискретного программирования и может быть решена с использованием алгоритма итеративных отображений [2].

Заключение. Как показал анализ технологий BigData данное направление в самом начале развития. Результаты заключаются в разработке теоретических основ на основе матричного исчисления и разработки на этой базе множества эффективных инструментов анализа и обработки больших данных.

Конечным результатом будут модели и методы, основанные на матричном исчислении и программные инструменты, которые могут быть реализованы на современной технике и достижениях производителей вычислительных систем.

Список цитируемых источников

1. Казиев, Г. З. Декомпозиция прикладных задач и исходных документов систем обработки данных на этапе технического проектирования / Г. З. Казиев, Г. С. Набиева, Б. М. Даурова // Изданий-Поиск : науч. журн. М-ва образования и науки. — № 1 (1) Сер. Естественно-технических наук. — Алматы : [б. и.], 2012. — С. 285—290.
2. Модели и методы дискретного программирования блочно-симметричные модели-эффективный класс задач дискретного программирования / Г. З. Казиев [и др.] // Вестник КБТУ. — Алматы : [б. и.], 2010. — № 3.

Материал поступил в редакцию 25.02.2015 г.