Muzafarova A.I., Minullin D.A., Gafarova V.R.
KFU, Institute of Computational Mathematics and Information Technologies, Institute of Applied Semiotics of the AS RT, Rossia, Tatarstan, Kazan

CLUSTER ANALYSIS OF TEXTS OF LESSONS PLANNING SYSTEM "ELECTRONIC EDUCATION OF THE REPUBLIC OF TATARSTAN"

Abstract. The article is devoted to the application of BigData methods in the school education system, using the example of analyzing the big data on the activities of teachers and student success, collected and continuously updated in the "Electronic Education in the Republic of Tatarstan" system. This work describes the development of a system for clustering texts of lesson planning to determine their belonging to the corresponding teaching materials. Based on the programs developed by the authors, the texts of the lesson planning are divided into 8 clusters, of which 6 clusters are in Russian, 2 clusters are in the Tatar language of instruction. Also, an analysis of the average marks of students was carried out, depending on the teaching materials used by teachers.

Keywords: *big data; data analysis in education; information processing; clustering of texts; comparison of texts.*

Музафарова А.И., Минуллин Д.А., Гафарова В.Р.
Казанский федеральный университет, ИВМиИТ, Институт прикладной семиотики АН РТ, Россия, Татарстан, Казань

КЛАСТЕРНЫЙ АНАЛИЗ ТЕКСТОВ ПОУРОЧНОГО ПЛАНИРОВАНИЯ СИСТЕМЫ «ЭЛЕКТРОННОЕ ОБРАЗОВАНИЕ РЕСПУБЛИКИ ТАТАРСТАН»

Аннотация. Статья посвящена применению методов больших данных (BigData) в системе школьного образования, на примере анализа массив больших данных о деятельности

педагогов и успешности учащихся, собранного и непрерывно обновляемого в системе «Электронное образование в Республике Татарстан». В этой работе описана разработка системы кластеризация текстов поурочного планирования для определения принадлежности их к соответствующему УМК. На основе разработанных авторами программ, тексты поурочного планирования выделены в 8 кластеров, из которых 6 кластеров на русском языке, 2 кластера на татарском языке преподавания. Также проведен анализ средних оценок учеников в зависимости от используемых педагогами УМК.

Ключевые слова: *большие данные, анализ данных в образовании, обработка информации, кластеризация текстов.*

1. Введение

Аналитика больших данных — это процесс изучения больших объёмов данных с целью выявления скрытых закономерностей, рыночных тенденций, предпочтений клиентов и другой полезной информации для принятия правильных решений [1]. Она была принята самыми различными отраслями и стала самостоятельной отраслью [14]. Оперирование большими данными (BigData) в образовании – это технология аналитики образовательной системы, включающей измерение, сбор, анализ и представление структурированных и неструктурированных данных огромных объемов об обучающихся и образовательной среде с целью понимания особенностей функционирования и развития образовательной системы [20]. Работа с объёмными массивами данных требует не только наличия современных аппаратных средств, но также и математических алгоритмов, которые позволили бы сократить необходимое число вычислительных операций для компьютера. Для успешного управления образовательным процессом необходимо оперативно обрабатывать многочисленные разнообразные поступающие данные в онлайн режиме, поэтому применение технологий BigData становится необходимостью [7].

Кластеризация — это задача поиска групп похожих документов в коллекции документов. Алгоритмы кластеризации являются одними из самых популярных методов

интеллектуального анализа данных, который широко применяется для обработки текстовых данных. Они имеют широкий спектр приложений, таких как классификация [2; 3], визуализация [4] и организация документов [6]. Существуют различные методы кластеризации текстов, наиболее популярные из них это LSA/LSI – Latent Semantic Analysis/Indexing [13], Suffix Tree Clustering [15], Scatter/Gather [5]. В последнее время популярность получили методы, основанные на использовании нейронных сетей [10] совместно с классическими методами кластеризации, такими как алгоритм k средних [8]. Алгоритмы, основанные на нейронных сетях, применяются также и в задачах классификация текстов: например, борьба со спамом, распознавание эмоциональной окраски текстов, разделение сайтов по тематическим каталогам, персонализация рекламы [19].

В данной работе решена задача кластеризация большого объёма текстовых данных, накопленных в системе «Электронное образование в Республике Татарстан» с 2014 по 2020 годы. На первом этапе с использованием метода косинусного расстояния тексты поурочного планирования проверены на схожесть. На втором этапе на основе использования метода агломеративной кластеризации тексты сгруппированы в 8 кластеров. Статья устроена следующим образом. В первом разделе описываются возможности гибкой библиотеки Dask для проведения параллельных вычислений в кластерных вычислительных системах, во втором разделе рассматривается оценка схожести текстов с использованием метода косинусного расстояния. В третьем разделе описан алгоритм агломеративной кластеризации, и его применение для кластеризации текстов поурочного планирования на основе матрицы схожести. Далее приводится заключение, подводящее итог статьи.

2. Обработка больших данных с использованием системы Dask

Основу исследования составили данные, собранные через государственную информационную систему «Электронное

образование в Республике Татарстан». Система включает в себя базы данных образовательной информации по всем учащимся и всем педагогам общеобразовательных организаций РТ. Для исследования наборы данных предоставлялись в формате Comma-separated values (CSV файлов). Общий объём данных составляет более 60 Гб. Такой объём данных невозможно эффективно обработать стандартными средствами. Следовательно, нужны технологии, которые позволяют обработать большой объем неструктурированных данных, систематизировать их, проанализировать и выявлять закономерности.

Для проведения расчетов был развернут вычислительный кластер, состоящий из 4-виртуальных машин, каждая из которых имеет по 1ТБ постоянной памяти, 32 ГБ оперативной памяти, 16 вычислительных ядер. На этом кластере была установлена система для параллельных вычислений - Dask. Dask – гибкая библиотека параллельных вычислений для аналитики, предназначенная главным образом для обеспечения масштабируемости и расширения возможностей существующих пакетов и библиотек [11]. Данная система позволяет производить параллельные вычисления на данных, размер которых превышает доступный объём памяти, на нескольких ядрах или нескольких машинах. Можно даже сконфигурировать Dask для использования ресурсов тысячи машин – каждой с несколькими ядрами (Зятев, 2019).

В обработку поступили данные, характеризующие профессиональную деятельность педагогов, включая темы уроков, задаваемые домашние задания, и оценки учеников. Процесс загрузки данных из CSV-файлов в Dask Dataframe осуществлялся с помощью функции dask.dataframe.read_csv(). Для переименования столбцов использовалась функция dask.dataframe.rename(). Для объединения фреймов данных применялся метод dask.dataframe.merge(). С помощью описанных инструментов все нужные для исследования данные были считаны и объединены в общий dataframe, который использовался для дальнейшей обработки.

Далее для дисциплины «математика» к сгруппированному по классам (1-11 классы) dataframe была применена функция параллельной обработки, которая для каждого учителя выделила темы занятий урока и заданные им домашние задания. После обработки исходных данных на кластере с установленной системой Dask, мы получили тексты поурочных занятий, сгруппированные по педагогам и предметам (Табл. 1)

Таблица 1.

Представление данные после обработки в системе Dask

worker_id	homework	theme
594697	РТ стр. 42 43	Связи между скоростью, временем и расстоянием
594697	РТ стр. 44 45	Письменное умножение двузначного числа на двузначное
594697	РТ стр. 46 47	Письменное умножение двузначного числа на двузначное

3. Определение схожести текстов

Следующей задачей исследования является построение матрицы схожести текстов. В настоящее время поиск сходства между текстами имеет большое практическое применение. Существует много методов и программ для сравнения текстов [16]:

1. Методы сравнения текстов, основанные на коэффициенте подобия Джакарда, косинусное подобие и расстояние Левенштейна.
2. Метод латентно-семантического анализа.
3. Программы обнаружения плагиата, основанные на поиске скрытой семантики текста и т.д.

Тексты поурочного планирования каждого преподавателя, полученные на первом этапе, соединяются в единую строку, которая подвергается следующей обработке:

1. Строка разбивается на токены.

2. Из массива токенов удаляются: знаки препинания, пустые строки и стоп-слова, которые не придают особого значения предложению.

Это необходимо для повышения точности сравнения.

На основе полученных данных осуществляется анализ схожести сравниваемых текстов. В качестве способа сравнения был выбран метод косинусного сходства [12]. Косинусное сходство – это мера сходства между двумя векторами пространства внутренних произведений, которое измеряет косинус угла между ними. Если даны два вектора признаков, A и B, то косинусное сходство, cos(θ), может быть представлено, используя скалярное произведение и норму [17]:

$$\cos(\theta) = \frac{\vec{a}\vec{b}}{\|\vec{a}\|\|\vec{b}\|} = \frac{\sum_1^n a_i b_i}{\sqrt{\sum_1^n a_i}\sqrt{\sum_1^n b_i}}, \quad (1)$$

где a и b вектора, $\|\vec{a}\|$ – координаты первого вектора (количество вхождений для первого текста), $\|\vec{b}\|$ – координаты второго вектора (количество вхождений для второго текста).

Сравнивая попарно вектора строим матрицу схожести текстов.

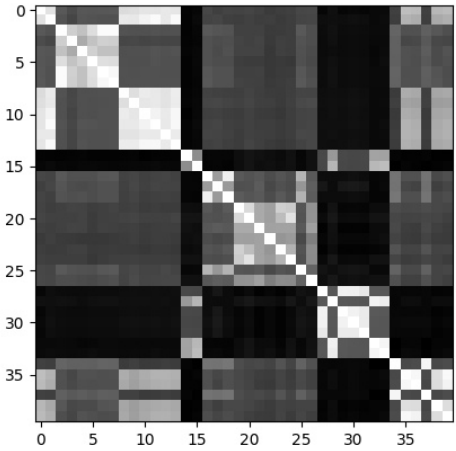

Рис. 1. Матрица похожести для 40 векторов

На рисунке 1 представлено изображение матрицы схожести текстов тематического планирования занятий для 40 преподавателей. По диагонали 100% совпадение, так как сравниваются одинаковые тексты. Чем ярче цвет, тем выше степень похожести между двумя текстовыми данными.

4. Кластерный анализ

Для обучения учеников начальной школы используется 11 различный УМК. Такие как: "Школа России", "Перспектива", "Начальная школа XXI века", "Гармония", "Перспективная начальная школа", "Планета знаний", "РИТМ", "Начальная инновационная школа", система Л.В. Занкова, система Д.Б. Эльконина-В.В. Давыдова, **"Учусь учиться (математика Л.В. Петерсон)" [18].** Однако в Республике Татарстан о основном используется только около 8-ми УМК, 2 из которых на татарском языке. Следовательно, было принято решение полученные данные по тематическому планированию разделить на 8 различных кластеров. Для осуществления этой задачи был выбран алгоритм, который относится к классу *агломеративных:* основной операцией является слияние нескольких уже имеющихся кластеров в один более крупный кластер.

Суть алгоритма заключается в следующем. Изначально все объекты считаются отдельными кластерами. Предполагается, что кластеризация образует покрытие исходного множества. При необходимости можно считать, что изолированные (не принадлежащие ни одному из кластеров) элементы образуют тривиальные кластеры из одного элемента. Затем начинается процесс последовательного слияния кластеров, на каждой итерации которого выбираются два наиболее близких кластера и объединяются в один новый [9]. Алгоритм заканчивает работу, когда остается только один кластер, совпадающий с исходным множеством. Таким образом создается дерево от листьев к стволу.

Расстояния между объектами (таблица 2) исходной выборки рассчитываются на основе матрицы схожести (таблица 1).

В итоге получается дерево кластеров, из которого можно выбрать кластеризацию с требуемой степенью точности.

Таблица 2.

Матрица расстояний для 6 векторов

0	0.1157	0.6469	0.6954	0.6555	0.6590
0.1157	0	0.6693	0.185	0.6624	0.6703
0.6469	0.6693	0	0.1880	0.2390	0.1294
0.6954	0.7185	0.1880	0	0.3266	0.2314
0.6555	0.6624	0.2390	0.3266	0	0.1807
0.6590	0.6703	0.1294	0.2314	0.1807	0

В идеальном раскладе уже на двух кластерах получилось бы разбиение на кластеры содержащий тексты русском языке и на татарском языке. Но этого не произошло, так как в заданиях, которые указывают учителя содержатся не только татарские слова, но и русские. Следовательно, однозначного разбиения на русский и татарский звенья на двух кластерах получить невозможно. Увеличивая количество кластеров уже на трёх кластерах, мы получаем один кластер на татарском языке, и два кластера на русском языке. При разделении на восемь кластеров отчетливо выделились два кластера на татарском языке и 6 кластеров на русском языке. Таким образом, мы предположительно получаем 6 различных групп УМК на русском языке и 2 на татарском языке (таблица 3). Принадлежность кластеров к определенным УМК было

определено на основе сравнения текстов поурочного планирования, введенных в систему педагогами с текстами из учебников по соответствующим УМК.

Таблица 3.

Результаты кластерного анализа

№ кластера	Количество учителей в кластере	Язык кластера	Средняя оценка учеников
0	314	Русский	4.01
1	503	Русский	3.95
2	154	Русский	3.97
3	502	Русский	3.90
4	223	Русский	3.97
5	117	Русский	4.03
6	124	Татарский	3.93
7	63	Татарский	4.00

Анализ результатов кластеризации показал, что разработанные программные средства позволяют корректно сгруппировать отдельные группы тексты поурочного планирования в соответствующие УМК. Также для каждой группы учителей (рис. 2) была рассчитана средняя оценка учеников, обучающихся по соответствующему УМК.

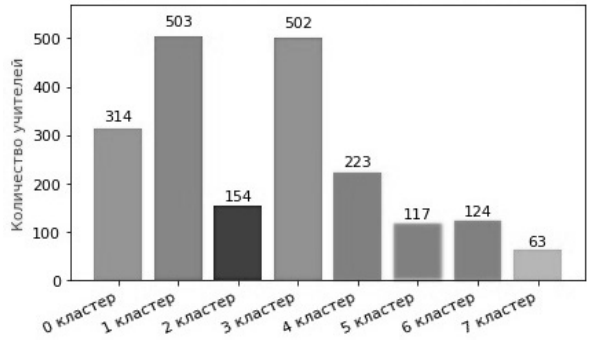

Рис. 2. Диаграмма распределения педагогов по кластерам

5. Заключение

В результате проведенного исследования, посвященного классификации текстов для определения принадлежности к соответствующему УМК, в 4 классе по дисциплине математика были выделены 8 кластеров, два из которых на татарском языке. Для каждого кластера был подобран предполагаемый учебно-методический комплекс и проведён сравнительный анализ успеваемости учеников. Разработанный программный комплекс можно использовать для автоматической обработки текстов поурочного планирования в целях определения УМК для всех предметов и классов.

Благодарности

Исследование выполнено при финансовой поддержке РФФИ в рамках научного проекта «Цифровая модель формирования индивидуальной траектории профессионального развития учителя на основе больших данных и нейросетей (на примере Республики Татарстан)», № 19-29-14082.

ЛИТЕРАТУРА

1. Ajah, I.A. (2019) Nweke, H.F. Big Data and Business Analytics: Trends, Platforms, Success Factors and Applications. Big Data Cogn. Comput. 3, 32.

2. AnickP., Vaithyanathan S. (1997) Exploiting Clustering and Phrases for Context-Based Information Retrieval. ACM SIGIR Conference

3. Angelova R., Siersdorfer S. (2006) A neighborhood-based approach forclustering of linked document collections. CIKM Conference.

4. Chakrabarti K., Mehrotra S. (2000) Local Dimension reduction: A newApproach to Indexing High Dimensional Spaces, VLDB Conference.

5. Douglass R. Cutting, David R. Karger, Jan O. Pedersen, and John W. Tukey. (1992). Scatter/Gather: a cluster-based approach to browsing large document collections. In Proceedings of the 15th annual international ACM SIGIR conference on Research and development in information retrieval (SIGIR '92). Association for Computing Machinery, New York, NY, USA, 318–329.

6. Fisher D. (1987) Knowledge Acquisition via incremental conceptual clustering. Machine Learning, 2: pp. 139–172.

7. Javidi G., Rajabion L. and Sheybani E.(2017) "Educational Data Mining and Learning Analytics: Overview of Benefits and Challenges," 2017 International Conference on Computational Science and Computational Intelligence (CSCI), pp. 1102-1107.

8. Manning et al. (2008) Christopher D Manning, Prabhakar Raghavan, Hinrich Schütze, et al. Introduction to information retrieval, volume 1. Cambridge university press Cambridge.

9. Mullner D. (2011). Modern hierarchical, agglomerative clustering algorithms, ArXiv, abs/1109.2378

10. Mikolov et al. (2013) Tomas Mikolov, Greg Corrado, Kai Chen, and Jeffrey Dean. 2013. Efficient Estimation of Word Representations in Vector Space. Proceedings of ICLR 2013, pages 1–12.

11. Rocklin M. Dask: Parallel Computation with Blocked algorithms and Task Scheduling, Proceedings of the 14th Python in Science Conference, 130 - 136 ,2015.

12. Singhal, A. (2001). "Modern Information Retrieval: A Brief Overview". Bulletin of the IEEE Computer Society Technical Committee on Data Engineering 24 (4): 35–43.

13. Susan T. (2005). "Latent Semantic Analysis". Annual Review of Information Science and Technology. 38: 188–230.

14. Youssra Riahi (2018). Big Data and Big Data Analytics: Concepts, Tyoes and Technologies – Moroko.

15. Zamir O., Etzioni O., Web Document Clustering: AFeasibility Demonstration, Proc. ACM SIGIRconference on Research and development in informationretrieval, New York, USA, pp. 46-54, 1998.

16. Бермудес С.Х.Г., Керимова С.У. (2016). О методе определения текстовой близости, основанном на семантических классах // Инженерный вестник Дона. № 4(43)

17. Гришин В.Д. (2018). Метода анализа и поиска заимствований в тексте // Проблемы науки. №7 (31)

18. Школьный гид (2018). Программы начальной школы [Электронный ресур]. URL: https://schoolguide.ru/index.php/progs.html

19. Попова Е.С., Спицын В.Г., Иванова Ю.А. (2019). Использование искусственных нейронных сетей для решения задачи классификации текста. Труды международной конференции по компьютерной графики и зрению "Графикон", - С. 270-273

20. Утёмов В.В., Горев П.М. (2018). Развитие образовательных систем на основе технологии BigData // Научно-методический электронный журнал «Концепт». - №6 (июнь). – С. 449-461.