НАЦІОНАЛЬНИЙ ТЕХНІЧНИЙ УНІВЕРСИТЕТ УКРАЇНИ
«КИЇВСЬКИЙ ПОЛІТЕХНІЧНИЙ ІНСТИТУТ імені ІГОРЯ СІКОРСЬКОГО»

Факультет прикладної математики

Кафедра програмного забезпечення комп'ютерних систем

"На правах рукопису"
УДК 004.032.26

«До захисту допущено»
Науковий керівник кафедри

_____ І. А. Дичка
(підпис)

"___" _____ 2017 р.

Магістерська дисертація

зі спеціальності 8.05010302 "Інженерія програмного забезпечення"

на тему: МЕТОД КЛАСТЕРИЗАЦІЇ КОРОТКИХ ТЕКСТОВИХ ДОКУМЕНТІВ НА ОСНОВІ НЕЙРОННИХ МЕРЕЖ

Виконав: студент 6 курсу, групи КП-52м
Даценко Андрій Сергійович _____
(підпис)

Науковий керівник доц., к.т.н., доц. Заболотня Т.М. _____
(підпис)

Рецензент доц.каф.ММСА ІПСА, доц., к.т.н. Дідковська М.В. _____
(підпис)

Рецензент доц.каф.СПіСКС ФПМ, доц., к.т.н. Марченко О.І. _____
(підпис)

Засвідчую, що у цій магістерській дисертації немає запозичень з праць інших авторів без відповідних посилань.
Студент _____
(підпис)

Київ – 2017

РЕФЕРАТ

Актуальність теми. На сьогоднішній день кластеризація текстових документів переживає бум в сфері аналізу коротких текстових документів. Зазвичай це є аналіз коментарів на форумах або повідомлень в соціальних мережах. Сучасні технології дозволяють передавати в тексті не тільки слова, а й емоції за допомогую спеціальних символів, а також створювати посилання з більшою вагою слова в повідомленні, наприклад хеш-тег. Класичні методи кластеризаціїї в таких випадках вже не видають достатніх результатів, тому в останніх роботах з текстової кластеризації все більше згадуються нейронні мережі – як головний механізм виявлення схожості груп текстів та аналізу текстових документів. Нейронні мережі вже достатньо гарно зарекомендували себе в різних галузях сучасних технологій, тому було вирішено розробити новий метод кластеризації з використанням нейронних мереж.

Об'єктом дослідження є процес аналізу текстових документів та автоматизованої генерації кластерів.

Предметом дослідження є методи та алгоритми автоматизованої кластеризації текстових документів та критерії оцінки ефективності кластеризації.

Мета роботи: створити новий метод кластеризації коротких текстових документів, що буде демонструвати кращі результати кластеризації за критеріями оцінки ефеткивності кластеризації, ніж існуючі методи.

Методи дослідження. В роботі використовуються методи збору даних, методи кластеризації текстових даних та статистичні методи.

Наукова новизна роботи полягає в наступному:

1. Вперше запропоновано нейронну мережу для аналізу текстових докуметів та генеруванню вектору особливостей для кожного слова на основі спільного використання та контексту.

2. Вперше запропоновано метод формування вектору особливостей документу на основі вектору ваги кожного окремого слова у документі.
3. Вперше запропоновано метод автоматичної генерації кластерів використовуючи сумарний вектор особливостей слів у документі.

Практична цінність отриманих в роботі результатів полягає в тому, що запропонований метод кластеризації коротких текстових документів дозволяє адаптуватися до вхідних даних шляхом урахування контексту слів з більшої вибірки, що значно покращує ефективність кластеризації. Таким чином досягаються кращі результати ефективності кластеризації, що вирішує задачу пов’язаності документів один між одним та враховує специфіку мовлення.

Апробація роботи. Основні положення і результати роботи були представлені та обговорювались на IX науковій конференції магістрантів та аспірантів "Прикладна математика та комп’ютинг" ПМК-2017 (Київ, 19–21 квітня 2017 р.) та опубліковані у збірнику тез за результатами конференції.

Структура та обсяг роботи. Магістерська дисертація складається з вступу, п'яти розділів, висновків та додатків.

У вступі надано загальну характеристику роботи, виконано оцінку поточного стану проблеми, обґрунтовано актуальність напрямку досліджень.

У першому розділі розглянуто теоретичні відомості, існуючі методи кластеризації текстових даних. Розглянуто особливості реалізації існуючих методів кластеризації текстових документів.

У другому розділі обґрунтовано вибір методів, що надають можливості для кластеризації коротких текстових документів; розглянуто модифікації до існуючух методів; запропоновано новий метод кластеризації коротких текстових документів.

У третьому розділі запропоновано засоби реалізації; наведено огляд архітектурних підходів до організації програмного забезпечення; запропоновано структуру та особливості реалізації методу; наведено відповідні графічні матеріали, що ілюструють взаємодію елементів системи.

У четвертому розділі наведено результати роботи алгоритму, підтверджено на практиці гіпотезу про те, що застосування розробленого алгоритму надає кращі результати кластеризації; здійснено порівняння ефективності кластеризації з існуючими методами; зроблено висновок щодо можливості застосування даного підходу для використання з різними вхідними даними для вирішення задачі кластеризації коротких текстових документів; запропоновано шляхи покращення та вектори розвитку для подальшої роботи.

У п'ятому розділі подано аналіз програмного продукту, його оцінку та перспективи для виходу на ринок. Наведені слабкі та сильні сторони проекту, порівняння з аналогами та конкурентоспроможність.

У висновках проаналізовано отримані результати роботи.

У додатках наведено фрагменти програмної реалізації запропонованого способу та копії графічних матеріалів.

Ключові слова: кластеризація, нейронні мережі, короткі текстові документи, машинне навчання.

ABSTRACT

Topic relevance. Novadays document clustering is having boom in sphere of analysis of short text documents clustering. Regulary it is related to analysis of comments on forums or social networks posts. Modern technologies allows to expose not ony textual data but also emotions using special symbols or add weight to message or word by using hash tag symbol. Classical methods of clustering analysis in such cases does not give proper results, due that reason in latest researches of document clustering a lot of neural networks approaches has been seen as main instrument of documents similarity detection and document analysis. Neural networks already recommended themselves in different fields of modern technologies, thus it was reasonable to implement new approach of document clustering using neural networks.

Object of reasearch is a process of text document analysis and automated cluster generation.

Subject of research are methods and algorithms for automated clusterign of text and criteria for evaluating the efficiency of clustering.

Purpose of research: is to create a new method of short text documents clustering, which will demonstrate the best clustering results according to criteria for assessing the efficiency of clustering than existing methods.

Research methods. Methods of data collection, methods of clustering of text data and statistical methods are used in this work.

The scientific novelty of the work is as follows:

1. For the first time, a neural network was proposed for the analysis of text documents and the generation of a features vector for each word based on general use and context.
2. For the first time, the method of forming a features document vector based on the weights vector of each individual word in a document is proposed.

3. For the first time, the method of automatic generation of clusters is proposed, which uses the generalized vector of features of words in a document.

The practical value of the results obtained in the work is that the proposed method of short text documents clustering allows to adapt to the input data by taking into account the context of words from a larger sample, which greatly improves the efficiency of clustering. In this way, the best results of clustering efficiency are achieved, which solves the problem of the interconnection of documents between each other and takes into account the specificity of speech.

Approbation of reasearch. The main provisions and results of the work were presented and discussed at the IX scientific conference of masters and postgraduates "Applied Mathematics and Computing", AMC-2017 (Kyiv, April 19-21, 2017) and published in the abstracts on the results of the conference.

Structure and scope of work. The master's dissertation consists of an introduction, five sections, conclusions and appendices.

The introduction gives a general description of the work, an assessment of the current state of the problem is performed, the relevance of the research direction is substantiated.

The first section deals with theoretical information, existing methods of clustering of textual data. The peculiarities of realization of existing methods of clustering of text documents are considered.

The second chapter justifies the choice of methods that provide opportunities for clustering of short text documents; modifications to existing methods are considered; a new method for short text documents clustering is proposed.

The third section proposes the means of implementation; an overview of architectural approaches to software organization is presented; the structure and features of the method’s implementation are proposed; the corresponding graphic materials illustrating the interaction of the system elements are given.

In the fourth section, the results of the algorithm are presented, and the hypothesis is confirmed in practice that the application of the developed algorithm provides better clustering results; comparison of the accuracy of work with existing methods is performed; the conclusion is drawn about the possibility of using this approach with different input data to solve the problem of short text documents clustering; the ways of improvement and development vectors for further work are suggested.

The fifth section provides analysis of the software product, its evaluation and prospects for market entry. The weak and strong sides of the project, comparisons with the analogues and competitiveness are presented.

The conclusion consists of analyzis of the results of the work.

The annexes show fragments of the software implementation of the proposed method and copies of graphic materials.

Key words: clustering, neural networks, short text documents, machine learning.

РЕФЕРАТ

Актуальность темы. На сегодняшний день кластеризация текстовых документов переживает бум в сфере анализа коротких текстовых документов. Обычно это анализ комментариев на форумах или сообщений в социальных сетях. Современные технологии позволяют передавать в тексте не только слова, но и эмоции с помощью специальных символов, а также создавать выражения с большим весом слова в сообщении, например хэш-теги. Классические методы кластеризациии в таких случаях уже не выдают достаточных результатов, поэтому в последних работах с текстовой кластеризации все больше упоминаются нейронные сети – как главный механизм нахождения сходства групп текстов и анализа текстовых документов. Нейронные сети уже достаточно хорошо зарекомендовали себя в различных отраслях современных технологий, поэтому было решено разработать новый метод кластеризации с использованием нейронных сетей.

Объектом исследования является процесс анализа текстовых документов и автоматизированной генерации кластеров.

Предметом исследования являются методы и алгоритмы автоматизированной кластеризации текстовых документов и критерии оценки эффективности кластеризации.

Цель работы: создать новый метод кластеризации коротких текстовых документов, будет демонстрировать лучшие результаты кластеризации по критериям оценки ефеткивности кластеризации, чем существующие методы.

Методы исследования. В работе используются методы сбора данных, методы кластеризации текстовых данных и статистические методы.

Научная новизна работы заключается в следующем:

1. Впервые предложено нейронную сеть для анализа текстовых докуметов и генерированию вектора особенностей для каждого слова на основе совместного использования и контекста.

2. Впервые предложен метод формирования вектора особенностей документа на основе вектора веса каждого отдельного слова в документе.
3. Впервые предложен метод автоматической генерации кластеров используя суммарный вектор особенностей слов в документе.

Практическая ценность полученных в работе результатов заключается в том, что предложенный метод кластеризации коротких текстовых документов позволяет адаптироваться к входным данным путем учета контекста слов с большей выборки, значительно улучшает эффективность кластеризации. Таким образом достигаются лучшие результаты эффективности кластеризации, который решает задачу связанности документов друг между другом и учитывает специфику речи.

Апробация работы. Основные положения и результаты работы были представлены и обсуждались на IX научной конференции магистрантов и аспирантов "Прикладная математика и компьютинг" ПМК-2017 (Киев, 19-21 апреля 2017) и опубликованы в сборнике тезисов по итогам конференции.

Структура и объем работы. Магистерская диссертация состоит из введения, пяти глав, заключения и приложений.

Во введении дана общая характеристика работы, выполнена оценка текущего состояния проблемы, обоснована актуальность направления исследований.

В первой главе рассмотрены теоретические сведения, существующие методы кластеризации текстовых данных. Рассмотрены особенности реализации существующих методов кластеризации текстовых документов.

Во втором разделе обоснован выбор методов, предоставляющих возможности для кластеризации коротких текстовых документов; рассмотрены модификации к иснуючух методов; предложен новый метод кластеризации коротких текстовых документов.

В третьем разделе предложены средства реализации; приведен обзор архитектурных подходов к организации программного обеспечения; предложена структура и особенности реализации метода; приведены соответствующие графические материалы, иллюстрирующие взаимодействие элементов системы.

В четвертом разделе приведены результаты работы алгоритма, подтверждено на практике гипотезу о том, что применение разработанного алгоритма предоставляет лучшие результаты кластеризации; проведено сравнение эффективности кластеризации с существующими методами; сделан вывод о возможности применения данного подхода для использования с различными входными данными для решения задачи кластеризации коротких текстовых документов; предложены пути улучшения и векторы развития для дальнейшей работы.

В пятом разделе представлен анализ программного продукта, его оценку и перспективы для выхода на рынок. Приведенные слабые и сильные стороны проекта, сравнение с аналогами и конкурентоспособность.

В выводах проанализированы полученные результаты работы.

В приложениях приведены фрагменты программной реализации предложенного способа и копии графических материалов.

Ключевые слова: кластеризация, нейронные сети, короткие текстовые документы, машинное обучение.