Сидоров С. П., доктор физ.-мат. наук, доцент  
Мельничук Д. В., канд. физ.-мат. наук, доцент  
ФГБОУ ВО «Саратовский национальный исследовательский государственный университет имени Н.Г. Чернышевского», г. Саратов

Sidorov S. P., Doctor degree in Physics and Mathematics, Associate Professor  
Melnichuk D. V., PhD in Physics and Mathematics, Associate Professor  
Saratov State University, Saratov

**АНАЛИЗ МЕДИАТЕКСТОВ НА ОСНОВЕ МЕТОДОВ ПОНИЖЕНИЯ РАЗМЕРНОСТИ**

**MEDIA TEXT ANALYSIS BASED ON DIMENSIONALITY REDUCTION METHODS**

**Аннотация**

Мы предлагаем новую методику анализа интересующих тем в медиапространстве. Методика включает в себя несколько этапов. Во-первых, из корпуса исследуемых текстов для каждой публикации извлекаются хэштеги (или ключевые слова), что позволяет построить двусвязную сеть из хэштегов и публикаций. Такое сетевое представление задается матрицей, строки которой соответствуют тегам, а столбцы - статьям. Затем преобразуем это представление в квадратную матрицу совместных упоминаний хэштегов. Далее алгоритм строит проекцию этого пространства на одномерное представление, решая оптимизационную задачу минимизации расстояния пути вокруг всех вершин графа. В результате, получаем ранжированный циклический список хэштегов. Учитывая возможность простой визуализации и интерпретации результатов, данный подход можно использовать и для изучения изменений в структуре медиапространства с течением времени.

**Abstract**

We propose a new methodology for analyzing topics of interest in media space. The method includes several steps. First, hashtags (or keywords) are extracted from the corpus of texts under study for each publication to build a bipartite network of hashtags and publications. Such a network representation is given by a matrix, the rows of which correspond to tags, and the columns to articles. Then, transform this representation to square matrix of hashtag co-mentions. Next, the algorithm builds a projection of this space into a one-dimensional representation, solving the optimization problem of minimizing the distance of the path around all the vertices of the graph. Thus, we get a ranked looped list of hashtags. Given the ability to easily visualize and interpret the results, this approach can also be used to examine changes in the structure of the media space over time.

**Ключевые слова**: Анализ текстов; Разбиение на слова; Анализ главных компонент; Историческая память.

**Keywords**: Text mining; Word Embedding; Traveling Salesman Problem; Principal component analysis; Memory studies.

При поддержке РНФ, 22-18-00153

Supported by the Russian Science Foundation, 22-18-00153

**Введение**

Наш подход включает в себя 4 этапа:

* Извлечение хэштегов (или ключевых слов);
* Построить двусвязную сеть хэштегов и публикаций, представленную в виде матрицы инцидентности и квадратной матрицы совместных упоминаний хэштегов;
* Построить проекцию на одномерное представление, решая оптимизационную задачу минимизации расстояния пути вокруг всех вершин графа;
* Получить ранжированный циклический список хэштегов.

В качестве исходных данных, нами были загружены 917 статей с сайта rg.ru. Эти статьи были опубликованы в период с 01/01/2022 по 01/01/2023. Сбор данных осуществлялся из новостной базы GDELT по запросам "Союза Советских Социалистических Республик" и "СССР".

**Математическая постановка задачи**

: Набор всех сообщений или публикаций в корпусе текстов ( новостей). : Набор хэштегов, используемых в корпусе текстов ( хэштеги). : Множество всех перестановок индексов в множестве .

Далее из множества формируется двухсторонняя сеть. Сеть, состоящая из узлов тегов и узлов статей, является вторым шагом нашего метода. На основе этой сети формируется матрица инцидентности, строки и столбцы которой соответствуют тегам и постам соответственно. Элемент матрицы равен 1, если тег упоминается в посте, и 0 - в противном случае. Каждый тег характеризуется многомерным вектором (соответствующим сырьем матрицы инцидентности).

**Методология**

Подход (Approach ). Подход заключается в простом подсчете количества совпадающих единиц для каждой пары строк  и матрицы и записи полученной суммы в соответствующие ячейки матрицы .

Подход (Approach ). Подход использует матрицу смежности , полученную подходом , суммируя для каждой пары строк и  минимальное значение для каждого столбца и записывая полученную сумму в соответствующую ячейку матрицы . Данный подход предполагает, что если пары тегов имеют общих соседей, то они должны быть более похожими.

Полученные матрицы смежности (близости тегов) были преобразованы в матрицы расстояний и с помощью преобразований и , где и . Матрицы расстояний получены с помощью неметрического многомерного шкалирования Крускала, которое дает двумерную конфигурацию.

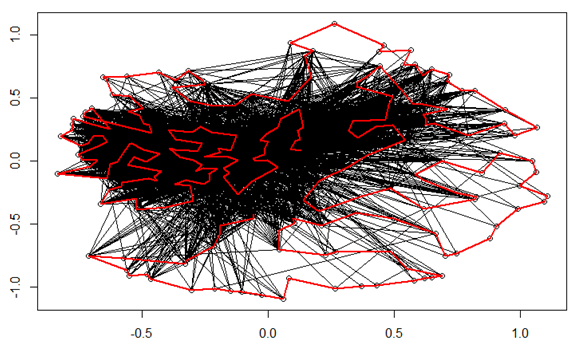
Мы формулируем оптимизацию разбиения и группировки слов как задачу о коммивояжере и решаем ее для нахождения оптимального одномерного вложения:

..

Мы используем подход, который реализует процедуру улучшения обмена между двумя ребрами. Это процедура уточнения тура, которая систематически меняет местами два ребра в графе, представленном матрицей расстояний. Далее мы рассматривали результат на метриках и в пространствах и в задаче .

**Эмпирические результаты**

На рисунке показан гамильтонов путь (Kendall’s horse shoe) с оптимальным обходом (красная линия) для подхода SoS в пространстве . Каждая точка представляет собой хэштег с его координатами в виде вектора вложения.



Для оценки качества полученных маршрутов в графе, нами были использованы следующие метрики:

* Cor R. Взвешенный коэффициент корреляции как мера эффективности для упорядочивания.
* ME. Мера эффективности, чем она больше, тем лучше выбран маршрут: .
* Stress. Измеряет краткость представления матрицы и может рассматриваться как функция чистоты, которая сравнивает значения в матрице с ее соседями .

Результат проверки методов:

| **Methods** | **Cor R** | **ME** | **Moore stress** | **Neumann stress** |
| --- | --- | --- | --- | --- |
| **SoS TSP L2** | 0.03 | 342 | **35534** | 17486 |
| **S TSP L2** | 0.07 | 333 | 35848 | 17492 |
| **SoS TSP L1** | 0.02 | **373** | 35584 | **17408** |
| **S TSP L1** | **0.10** | 353 | 35704 | 17412 |
| **SoS PCA** | 0.04 | 201 | 36430 | 17982 |
| **S PCA** | 0.06 | 211 | 36754 | 18090 |

Примеры сегментов:

| **Methods** | **Segments around “Judiciary.”** | **Segments around “NATO Expansion.”** |
| --- | --- | --- |
| **SoS TSP L2** | Право; Правовые; Постановления и разъяснения Верховного Суда; Судебная власть; Верховный Суд; Конституция России; Министерство образования и науки | Ядерная безопасность; Великобритания; Евразийский экономический союз (ЕАЭС); Расширение НАТО; Россия и НАТО; Казахстан; "Единая Россия" |

Как видно из верхней таблицы, в соответствии с критериями оценки качества полученных упорядоченных списков, трудно сделать выбор в пользу того или иного подхода.

Списки некоторых кольцевых сегментов мы приводим в нижней части таблицы. Отметим, что подход SoS TSP L2 обеспечивает наиболее естественный порядок, поскольку последовательные слова семантически схожи.

**Выводы**

Предложенный метод, позволяет разделить корпус текстов на множества, извлечь из каждого множества хэштеги (ключевые слова), создать сеть хэштегов и публикаций, и преобразовать ее в матрицу совместных упоминаний. Затем алгоритм ранжирует хэштеги и представляет пространство в одном измерении.

В заключение можно сказать, что определить предпочтительный подход достаточно сложно, однако подход SoS TSP L2 считается естественным в силу семантического сходства между последовательными словами.

**Список цитируемой литературы:**

1. Sato, R. Word Tour: One-dimensional Word Embeddings via the Traveling Salesman Problem // Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. – 2022. – P. 2166-2171. – DOI: 10.18653/v1/2022.naacl-main.157.
2. Farouk, M. Measuring text similarity based on structure and word embedding // Cognitive Systems Research. – 2020. – Vol. 63. – P. 1-10. – ISSN 1389-0417. – DOI: https://doi.org/10.1016/j.cogsys.2020.04.002.
3. Kenter, T., & de Rijke, M. Short Text Similarity with Word Embeddings // Proceedings of the 24th ACM International on Conference on Information and Knowledge Management, CIKM '15. – New York, NY, USA: Association for Computing Machinery, 2015. – P. 1411-1420. – ISBN 9781450337946. – DOI: 10.1145/2806416.2806475.