Подход к извлечению и кластеризации библиографической информации

А.А. Дырночкин
Ульяновский государственный технический университет
Ульяновск, Россия
dyrno4kin@gmail.com

В.С. Мошкин
Ульяновский государственный технический университет
Ульяновск, Россия
postforvadim@ya.ru

***Аннотация*—В данной статье представлена система извлечения библиографической информации для последующего наукометрического анализа публикаций. Описан алгоритм загрузки и предобработки статей. Предложен подход для формирования научных групп по заданной тематике посредством кластеризации текстов аннотаций статей. Также в работе представлены результаты экспериментов с данными по статьям из научной библиотеки elibrary.**

Ключевые слова— анализ текста, парсинг веб-страниц, векторизация текстов, кластеризация, кластеризация k-means.

1. Введение

Анализ наукометрических данных является важным аспектом при формировании рейтингов научной активности или подборе научных групп по определенной тематике.

Данные о научных публикациях хранятся в цифровых библиографических базах данных. Такие базы помогают отслеживать цитируемость статей, опубликованных в научных изданиях. Также они являются одним из источников получения наукометрических данных, для проведения различных оценочных исследования [1-3].

2. Алгоритм формирования научных групп по заданной тематике

В рамках данного исследования был разработан алгоритм извлечения библиографической информации с сайта elibrary, для последующего наукометрического анализа и формирования научных групп по заданной тематике.

Данный алгоритм включает следующие этапы:

1. Загрузка информации по статьям с сайта elibrary

Загрузка статей состоит из нескольких этапов. На первом этапе происходит загрузка всех статей по автору в формате: название статьи и уникальный идентификатор. На втором этапе происходит загрузка всех данных по каждой статье (авторы, аннотация, год издания, ключевые слова и. др.) [4].

2. Предобработка загруженных статей

Предобработка включает в себя перевод иностранных статей на русский язык, токенизацию и стемминг по методу Портера, который не требует дополнительных словарей, а также фильтрацию по стоп словам. На рисунке 1 представлено описание схемы предобработки статей [5].

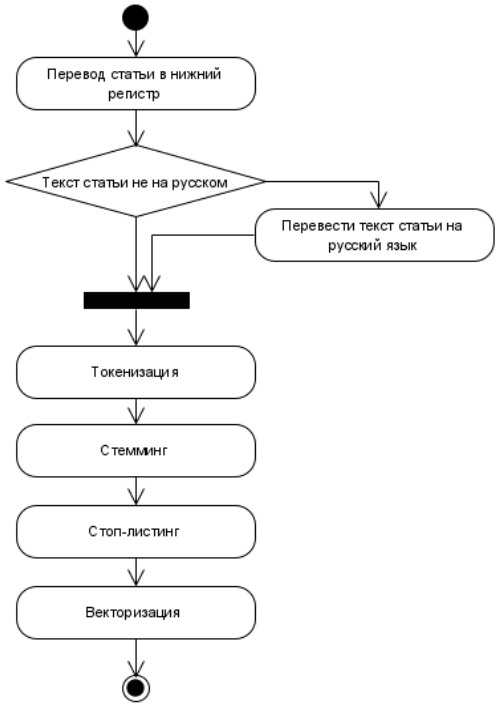

Рис 1. Схема предобработки статей

3. Векторизация предобработанных статей

На данном этапе происходит векторизация предобработанных документов методом TF-IDF.

TF-IDF – мера оценки важности слова в контексте документа, являющегося частью коллекции документов. Вес некоторого слова пропорционален частоте употребления этого слова в документе и обратно пропорционален частоте употребления слова во всех документах коллекции [6].

Частота слова TF рассчитывается по формуле 1.

$$tf(t,d) = \frac{n_i}{\sum_k n_k} \quad (1)$$

где n_i число вхождений слова в документ, а в знаменателе – общее число слов в данном документе.

IDF (обратная частота документа) – инверсия частоты, с которой некоторое слово встречается в документах коллекции. Учёт IDF уменьшает вес часто употребляемых слов [7]. Для каждого слова в пределах одной коллекции документов существует только одно значение IDF. Обратная частота рассчитывается по формуле 2.

$$idf(t,D) = \log \frac{|D|}{|d_i \supset t_i|} \quad (2)$$

где |D| – количество документов в корпусе; $|(d_i \supset t_i)|$ – количество документов, в которых встречается слово

Таким образом, TF-IDF вычисляется по формуле 3

$$tf - idf(t,d,D) = tf(t,d) \times idf(t,D) \quad (3)$$

4. Кластеризация векторизованных текстов методом k-средних.

Результатом данного этапа является разбиение исходного корпуса текстов, на заданное количество кластеров. Метод k-средних разделяет X документов на k кластеров ($k \leq X$) , чтобы минимизировать суммарное квадратичное отклонение точек кластеров от центроидов этих кластеров [8]. Минимальное суммарное отклонение рассчитывается по формуле 4.

$$\min\left[\sum_{i=1}^{k}\sum_{x(j)\in Si}\left\|x^{(j)}-u_i\right\|^2\right] \quad (4)$$

где u_i - центроид для кластера S_i

5. Формирование научных групп по заданной тематике.

Для реализации предложенного алгоритма был разработан модуль загрузки статей с сайта elibrary и модуль предобработки и кластеризации на языке программирования Java.

Общая архитектура системы представлена на рисунке 2.

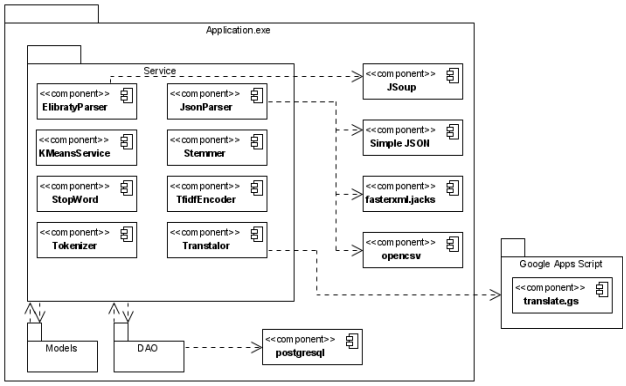

Рис. 2 Общая схема системы

Для проведения кластеризации была загружена библиографическая информация по сотрудникам УлГТУ, всего было загружено порядка 14 тысяч статей в формате: название статьи и уникальный идентификатор, и в дальнейшем было загружено полное описание 1000 статей для проведения экспериментов.

В результате кластеризации корпус статей был разделен на 12 кластеров. На рисунке 3 представлен результат кластеризации и вывода рекомендаций по формированию научной группы.

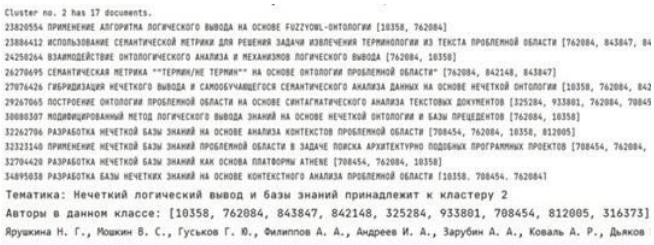

Рис. 3 Результаты кластеризации

3. Заключение

Разработанная в рамках исследования система позволяет извлекать библиографическую информацию по авторам из системы elibrary. Также представлен механизм кластеризации научных статей и формирования рекомендаций при составлении научных групп по заданной тематике.

Благодарности

Работа выполнена при поддержке Минобрнауки России в рамках проекта № 075-00233-20-05 от 03.11.2020 «Исследование интеллектуального предиктивного мультимодального анализа больших данных и извлечения знаний из различных источников».

Литература

[1] Николаев, А.В. Критическая кластеризации научной литературы / А.В. Николаев, В.В. Жуков // Информационно-телекоммуникационные технологии и математическое моделирование высокотехнологичных систем. – 2021. – С. 268-273.

[2] Низомутдинов, Б.А. Автоматизированный сбор данных для наукометрического анализа / Б.А. Низомутдинов, А.С. Тропников // Научный сервис в сети Интернет. – Федеральный исследовательский центр Институт прикладной математики им. МВ Келдыша Российской академии наук. – 2019. – Т. 21. – С. 523-531.

[3] Пархоменко, П.А. Обзор и экспериментальное сравнение методов кластеризации текстов / П.А. Пархоменко, А.А. Григорьев, Н.А. Астраханцев //Труды Института системного программирования РАН. – 2017. – Т. 29, № 2. – С. 161-200.

[4] Мусаев, А.А. Обзор современных технологий извлечения знаний из текстовых сообщений / А.А. Мусаев //Computer. – 2021. – Т. 13, № 6. – С. 1291-1315.

[5] Юферев, В.И. Векторизация текстов на основе word-embedding моделей с использованием кластеризации / В.И. Юферев, Н.А. Разин // Моделирование и анализ информационных систем. – 2021. – Т. 28, № 3. – С. 292-311.

[6] Кравченко, Ю.А. Векторизация текста с использованием методов интеллектуального анализа данных / Ю.А. Кравченко, А.М. Мансур, М.Ж. Хуссайн // Известия Южного федерального университета. Технические науки. – 2021. – № 2(219). – С. 154-167.

[7] Alam, M. A Review on Clustering of Web Search Result / M. Alam, K. Sadaf // Advances in Computing and Information Technology. Advances in Intelligent Systems and Computing. – Springer, Berlin, Heidelberg, 2013. – Vol. 177.

[8] Трубников, В.С. Проектирование системы сбора, анализа и визуализации наукометрических данных / В.С. Трубников, К.А. Туральчук // Проблемы современной науки и образования. – 2015. – № 6(36).

[9] Yarushkina, N. Development of a knowledge base based on context analysis of external information resources / N. Yarushkina, V. Moshkin, A. Filippov // Proceedings of the International conference Information Technology and Nanotechnology. Session Data Science. – Samara, Russia, 2018. – P. 328-337.