УДК 004.75+004.91

# РАСПРЕДЕЛЕННАЯ СИСТЕМА СТАТИСТИЧЕСКОГО АНАЛИЗА ДОКУМЕНТОВ

*Северин К. М., студент*

*Белорусский государственный университет информатики и радиоэлектроники*
*г. Минск, Республика Беларусь*

*Парамонов А.И. – канд. техн. наук, доцент, зав. каф. ИСиТ*

**Аннотация.** Рассмотрены методы анализа текста. Выполнен анализ задач, решаемых статистическим методом. Изучена проблема автоматической классификации текстовых документов Рассмотрен алгоритм TextRank для классификации документа. Предложено проектное решение информационно-аналитической системы классификации документов.

**Ключевые слова.** Анализ документов, статистический анализ, TextRank, алгоритм, классификация текста, информационно-аналитическая система, граф, распределенная система, сервер, клиент.

***Введение.*** Значительный рост объемов цифровой информации, в том числе цифровых документов, определяет потребность в новых инструментах для их обработки. Зачастую это массивы неструктурированных текстовых документов, по которым необходимо выполнять поиск, осуществлять их систематизацию, проводить оценку и многие другие операции. Очевидно, что для всех этих задач необходимо выполнить анализ цифрового текста.

В настоящее время разработано и используется множество разных подходов и методов анализа текстовых документов, которые направлены на решение различных прикладных задач: интент-анализ, контент-анализ, фоносемантический, графематический, морфологический, синтаксический, семантический и другие виды анализа [1].

***Основная часть.*** Зачастую для достижения лучших показателей при обработке текста любому качественному анализу должен предшествовать количественный. Стоит отметить, что применение статистики для анализа текстов – традиционная задача. На сегодняшний день существует несколько методов статистического анализа текста. Среди них выделяют процедуры количественных исследований, частотный анализ, контент-анализ, ранжирование данных и другие. С помощью статистического анализа можно решить множество задач:

— классификация текста (статистическая стилистика);

— установление авторства (проведения атрибуции текстов на основании неповторимого сочетания статистических параметров авторского текста);

— описание поведения языковых единиц (букв, морфем, слов) в тексте, их распределение, сочетаемость, частота употребления;

— измерение информативности (расчет количества информации, содержащейся в тексте и его составных частях);

— восстановление текстов (описание структуры на основании очень ограниченной исходной информации) и т.д.

Наибольший интерес вызывает применение статистического анализа для решения задачи ранжирования документов по темам, отраслям и т.п. Ранжирование (или категоризация) документов – это частный случай классификации. Ее задача состоит в том, чтобы отнести объект к некоторой категории. В свою очередь, классификация является одной из основных задач компьютерной лингвистики, поскольку к ней сводится ряд других.

Задачу автоматической классификации текстовых документов можно решить с помощью разных алгоритмов, большая часть которых построена на извлечении ключевой информации из документов [2-4]: TF-IDF, RAKE, YAKE, BERT, TopicRank, TextRank.

В работе для проведения статистического анализа документов (получения его ключевых характеристик) предлагается использовать алгоритм TextRank [5]. Основой данного алгоритма является представление текста в виде графа, где вершинами графа являются слова, с последующим вычислением весов вершин графа.

Связи между вершинами в общем случае формируются исходя из отношений между словами. Обычно используется связь совместного появления слов в пределах определенного окна размера N. Данное отношение является симметричным, и как следствие рёбра являются неориентированными. Веса вершин устанавливаются в некоторое начальное значение, равное для всех вершин. Конкретные значение не так важны, т.к. веса относительны. Далее веса итеративно обновляются по формуле:

$$S(V_i) = (1-d) + d * \sum_{V_j \in In(V_i)} \frac{w_{ij}}{\sum_{V_k \in Out(V_j)} w_{jk}} * S(V_j)\,, \quad (1)$$

где $V_i$ - вершина графа; $In(V_i)$ - множество вершин, имеющих связь, направленную в $V_i$; $Out(V_i)$ - множество вершин, в которые направлены связи из $Vi$ ;$d$ - коэффициент затухания от 0 до1 (обычно 0.85); $w_{ik}$ - вес ребра направленного из вершины $V_i$ к вершине $V_k$.

Обновление весов завершается после определенного кол-ва итераций или по условию, например, как только порядок изменения весов опустится до определенного значения. После вычисления всех весов они упорядочиваются по убыванию. Согласно данному алгоритму, чем выше вес, тем более ключевым является слово для текста.

При реализации алгоритмов анализа документов возникает проблема ресурсоемкости и временных затрат. Можно применить аппаратные решения, однако они будут сильно дорогостоящие и быстро достигнут предела прироста производительности. Перспективным подходом к решению этой проблемы является построение распределенной системы анализа массива документов.

В работе предлагается проектное решение информационно-аналитической системы (ИАС) с организацией распределенной обработки документов на основе брокеров задач (Рисунок 1).

Основными элементами ИАС выступают:

— Gateway – внешний веб-сервер, который доступен конечному пользователю. Основная задача - проксирование запросов на App Instance, а также балансировка нагрузки между ними.

— Database – единое место хранения данных.

— Message Broker – используется для отправки задач от App Instance к Worker Instance.

— App Instance – приложение, обрабатывающее не ресурсоемкие запросы (например, сделать запрос к базе данных и забрать результаты обработки текста, добавить текст в очередь на обработку).

— Worker Instance – приложение, выполняющее ресурсоемкие задачи (например, обработка и анализ текстов).

Оба модуля App Instance и Worker Instance взаимодействуют с брокером сообщений для эффективного распределения задач по обработке документов, а результаты сохраняют в базу данных.

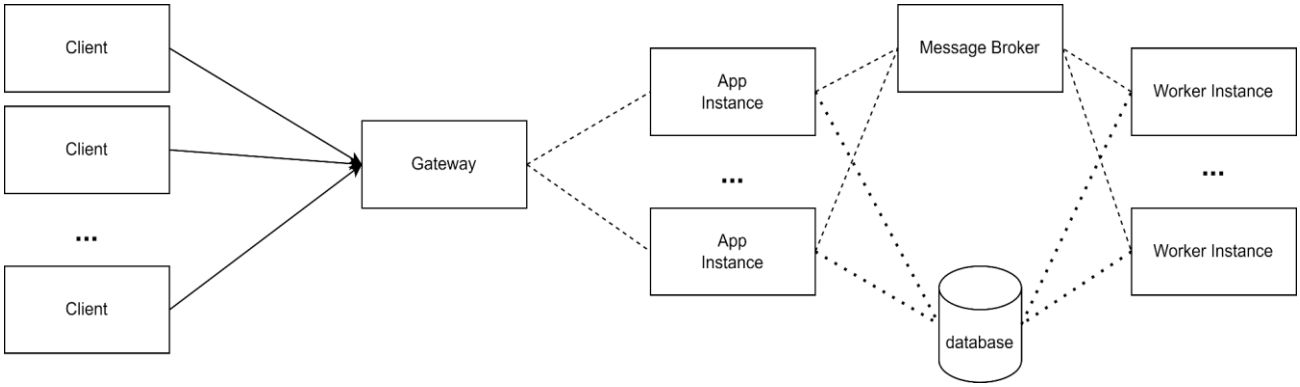


Рисунок 1 – Общая схема распределенной системы.

Для эффективной работы ИАС необходимо учитывать проблему эффективного распределения нагрузки между разными сетевыми ресурсами [6]. В данном случае вычислительными ресурсами выступают приложения по обработке запросов (Instances).

При программной реализации предложенной архитектуры использовались такие технологии:

В качестве Gateway элемента – высокопроизводительный веб- и прокси-сервер Nginx, который может обрабатывать большое количество запросов с минимальной задержкой. Кроме того, он также имеет встроенную балансировку нагрузки, кеширование, защиту от DDoS-атак и многое другое. Модули App Instance реализованы на CPython версии 3.10 с использованием веб-фреймворка FastAPI. Данный фреймворк показывает хорошие показатели на тестах производительности, побеждая по некоторым позициям своих основных конкурентов Flask и Falcon [7].

Для реализации механизма постановления задач в очередь и их выполнения Worker Instance применена технология Celery, общая архитектура обработки запросов в Celery представлена на рисунке 2. В этой архитектуре в качестве брокеров могут выступать Redis и RabbitMQ, а в качестве бэкендов Redis, RabbitMQ, SQLAlchemy (базы данных). В работе использовались и показали эффективность в качестве брокера RabbitMQ и SQLAlchemy как бэкенд для хранения результатов. Так же, как и для App Instance, Worker Instance реализованы на CPython версии 3.10.

В данном решении за работу с данными отвечает мощная, открытая объектно-реляционная система управления базами данных Postgres, которая использует и расширяет язык SQL и имеет множество функций, которые безопасно хранят и масштабируют самые сложные нагрузки по данным.

В ходе масштабирования приложения часто возникают проблемы с ростом соединений к базе данных. Для решения этой проблемы можно использовать легковесный пулер соединений для PostgreSQL – PgBouncer. Он может создавать и повторно использовать соединения к одной или нескольким базам данных (на одном или разных серверах) и обслуживать клиентов по TCP и Unix-

сокетам, поддерживает несколько режимов пула соединений, таких как session, transaction и statement. PgBouncer позволяет уменьшить накладные расходы на открытие и закрытие соединений к PostgreSQL, которые могут быть дорогостоящими в терминах производительности и ресурсов; ограничить количество одновременных соединений; распределять нагрузку между несколькими серверами СУБД; обеспечить высокую доступность СУБД, переключаясь на резервный сервер в случае сбоя основного.

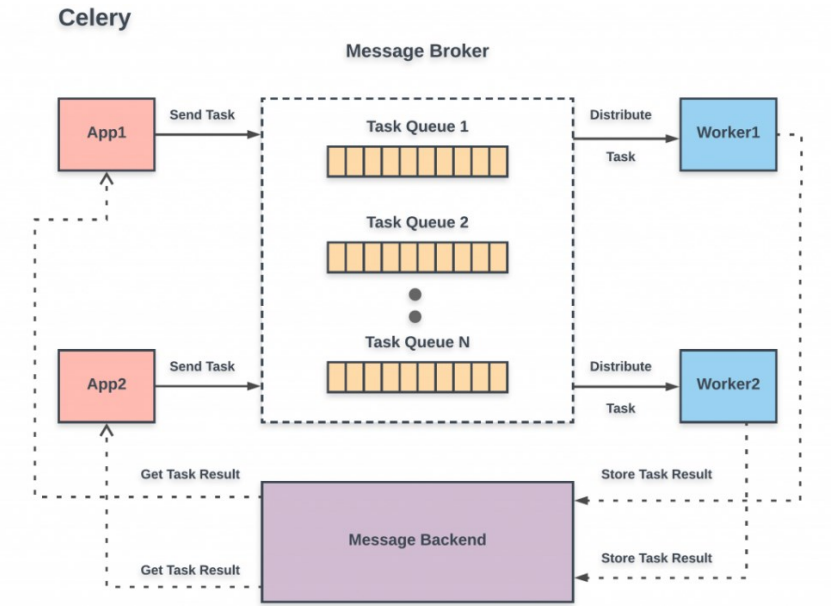


Рисунок 2 – Общая архитектура Celery

***Заключение.*** Таким образом получилась хорошо масштабируемая ИАС, способная достаточно быстро и точно обрабатывать большие массивы документов и классифицировать их по заранее подготовленным категориям. Ожидается, что система позволит повысить эффективность работы с документами за счет сокращения времени поиска.

***Список использованных источников:***

*1. Митина, О.В. Методы анализа текста: методологические основания и программная реализация / О В. Митина, А.С. Евдокименко // Челябинск: Вестник ЮУрГУ. — 2010. — № 40 (216). — С. 29-38.*

*2. Методы автоматической классификации текстов [Электронный ресурс]. – Режим доступа: http://swsys.ru/index.php?page=article&id=4252/. – Дата доступа: 01.04.2023.*

*3. Keyword Extraction: from TF-IDF to BERT [Электронный ресурс]. – Режим доступа: https://towardsdatascience.com/keyword-extraction-python-tf-idf-textrank-topicrank-yake-bert-7405d51cd839/. – Дата доступа: 04.04.2023.*

*4. TopicRank: Graph-Based Topic Ranking for Keyphrase Extraction [Электронный ресурс]. – Режим доступа: https://www.researchgate.net/ publication/258908054_TopicRank_Graph-Based_Topic_Ranking_for_Keyphrase_Extraction/. – Дата доступа: 28.03.2023.*

*5. Милкова, М. А. Современные методы извлечения ключевой информации из нормативных документов / М. А. Милкова, И. В. Неволин, Д. П. Пигорев. // Экономическая наука современной России. — 2021. — № (2). — С. 101-114.*

*6. Северин, К. М. Программный менеджер распределенных вычислений в мультиагентной среде / К. М. Северин, А. И. Парамонов // Информационные технологии и системы 2022 (ИТС 2022) = Information Technologies and Systems 2022 (ITS 2022) : материалы Международной научной конференции, Минск, 23 ноября 2022 / Белорусский государственный университет информатики и радиоэлектроники ; редкол.: Л. Ю. Шилин [и др.]. – Минск : БГУИР, 2022. – С. 165–166.*

*7. Web Framework Benchmarks [Электронный ресурс]. – Режим доступа: https://www.techempower.com/benchmarks/#section=data-r21&hw=ph&test=composite&a=2&l=zijzen-6bj&c=d&b=4&d=e3. – Дата доступа: 04.04.2023.*

UDC 004.75+004.91

# DISTRIBUTED SYSTEM OF STATISTICAL ANALYSIS OF DOCUMENTS

*Severin K. M.*

*Belarusian State University of Informatics and Radioelectronics,*
*Minsk, Republic of Belarus*

*Paramonov A. I. – Candidate of Engineering Sciences, Associate Professor*

**Annotation.** The methods of text analysis are considered. The analysis of the tasks solved by the statistical method is carried out. The problem of automatic classification of text documents is studied, the TextRank algorithm for document classification is considered. The design solution of the information and analytical system of classification of documents is proposed.

**Keywords.** Document analysis, statistical analysis, TextRank, algorithm, text classification, information and analytical system, graph, distributed system, server, client..