Министерство образования Республики Беларусь
Учреждение образования
Белорусский государственный университет
информатики и радиоэлектроники

УДК 004.422.8

Костенич
Антон Михайлович

Система анализа новостей с учетом актуальных тенденций в социальных сетях

АВТОРЕФЕРАТ
на соискание степени магистра информатики и вычислительной техники
по специальности
1-40 81 02 «Технологии виртуализации и облачных вычислений»

Научный руководитель
Насуро Е. В.
кандидат технических наук

МИНСК 2018

КРАТКОЕ ВВЕДЕНИЕ

Развитие информационных технологий и увеличение доступности информации привело рынок новостных ресурсов к состоянию постепенного перехода от классического формата публикаций в виде газет и журналов с определенным интервалом к электронному формату. Наиболее ярким и смелым примером является американский еженедельный новостной журнал Newsweek, который на момент перехода на цифровой формат в январе 2013 года являлся вторым по величине еженедельником в США, незначительно уступая Time по тиражу.

На сегодняшний день на рынке существует большое количество новостных источников, начиная изданиями, требующими платную подписку, заканчивая новостными аккаунтами в социальных сетях. Такое количество источников порождает проблему информационного шума, при которой пользователи могут получать большое количество «ненужной» им информации. Для решения этой проблемы создаются так называемые новостные агрегаторы.

Новостной агрегатор – ресурс, занимающийся сбором и структурированием новостей.кр, как правило – в автоматическом режиме. В большинстве случаев, агрегаторы создаются специально для крупных поисковых систем или информационных сайтов (например, Яндекс.Новости, Новости Mail.ru), однако с ними конкурируют продукты, специализирующиеся на сборе новостей с различных информационных ресурсов, такие как Flipboard или Feedly.

На данный момент существующие агрегаторы не обладают функционалом по подбору новостей на основе социальных сетей конкретного пользователя, что приводит к своеобразной ограниченности подбора ресурсов: новости будут предложены исключительно на основании выбранных пользователем категорий.

В данной диссертационной работе подробно исследуются существующие подходы для подбора новостей и обработки текстовых корпусов с последующим выделением ключевых слов. Также в рамках работы происходит реализация системы анализа новостей с учетом актуальных тенденций в социальных сетях.

ОБЩАЯ ХАРАКТЕРИСТИКА РАБОТЫ

Цели и задачи исследования

Целью диссертационной работы является анализ возможности реализации и непосредственная разработка высоконагруженной системы, агрегирующей новостные ресурсы с учетом данных, получаемых из социальных сетей конкретного пользователя.

Для достижения поставленных целей были выделены следующие *задачи*:

1. Разработать архитектуру высоконагруженной и устойчивой системы для улучшения опыта взаимодействия пользователей, оптимизации затрат на поддержку системы и усовершенствования ее работоспособности.

2. Проанализировать существующие методы обработки лингвистических корпусов, определить наиболее быстрый и подходящий метод и разработать программное обеспечение, реализующее его.

3. Разработать унифицированный подход работы с интерфейсом социальных сетей для создания абстракции над непосредственно используемыми инструментами.

Объектом исследования являются высоконагруженные системы и системы обработки лингвистических корпусов.

Предметом исследования являются алгоритмы выделения ключевых слов и архитектура высоконагруженных и устойчивых веб-систем.

Практическая значимость работы заключается в создании системы, использующей алгоритмы и подходы, не имеющие на данный момент аналогов на рынке, с возможным последующим ее продвижением.

Связь работы с приоритетными направлениями научных исследований и запросами реального сектора экономики

Результаты работы магистерской диссертации будут использованы при решении производственных задач в направлении исследования новых технологических возможностей и рынков для дальнейшего развития.

Личный вклад соискателя

Результаты, приведённые в диссертации, получены соискателем лично. Вклад научного руководителя Е. В. Насуро заключается в формулировке целей и задач исследований, а также руководстве ходом работы.

Апробация результатов диссертации

Технические аспекты доложены на 53-ой научной конференции аспирантов, магистрантов и студентов (Минск, БГУИР, май 2017).

Опубликованность результатов диссертации

По теме диссертации суммарно опубликовано одна печатная работа в сборнике 53-ой научной конференции аспирантов, магистрантов и студентов.

Структура и объём диссертации

Диссертация состоит из общей характеристики работы, введения, трех глав, заключения, списка использованных источников, списка публикаций автора.

Общий объем работы составляет 55 страниц, из которых основного текста – 43 страницы, 9 рисунков на 7 страницах, список использованных источников из 17 наименования на 2 страницах и 1 приложения на 6 страницах.

КРАТКОЕ СОДЕРЖАНИЕ РАБОТЫ

В первой главе освещаются типы агрегаторов и динамика их развития, а также раскрывается понятие индексации и подходы к ее проведению.

Обзор различных особенностей реализации высоконагруженных систем представлен во второй главе. Исследуются различные подходы и архитектуры необходимые для реализации систем подобного типа. Рассматриваются средства реализации с подробным обоснованием.

Помимо этого, вторая глава освещает существующие подходы к обработки лингвистических корпусов, способы выделения ключевых слов и фраз. Проведено обоснование выбора конкретного алгоритма.

Третья глава диссертации подробно рассматривает процесс реализации высоконагруженной системы выделения ключевых слов. Подробно рассматриваются особенности построения высоконагруженных систем: выбор архитектуры каждого модуля и связи между ними. Также подробно освещается алгоритм выделения ключевых слов.

ЗАКЛЮЧЕНИЕ

Основные научные результаты диссертации

1. Проведено исследование подходов к построению архитектуры высоконагруженных систем.

2. Выполнен обзор различных алгоритмов выделения ключевых слов.

3. Разработано программное обеспечение, реализующее алгоритм выделения ключевых слов, удовлетворяющее поставленным требованиям и стандартам.

4. Разработан унифицированный подход работы с внешним интерфейсом, предоставляемым социальными сетями.

Рекомендации по практическому использованию результатов

1. Результаты теоретического исследования могут быть использованы для обучения подходам к построению высоконагруженных систем.

2. Полученные результаты формируют теоретическую и практическую базу для разработки программного обеспечения, обрабатывающего лингвистические корпусы. Они могут быть использованы при разработке новых и совершенствовании существующих проектов в данной области.

3. Разработанная система может быть доведена до стадии возможности использования конечным пользователем и выведена на рынок для коммерческого использования.

СПИСОК ОПУБЛИКОВАННЫХ РАБОТ

1. Костенич, А. М. Система анализа новостей с учетом актуальных тенденций в социальных сетях / А. М. Костенич // Компьютерные системы и сети: материалы 53-й научной конференции аспирантов, магистрантов и студентов. (Минск, 2 – 6 мая 2017 года). – Минск : БГУИР, 2017. – С. 21 - 23.