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

УДК 002.06

Краснов
Андрей Юрьевич

Алгоритмы объединения и визуализации ресурсов пользователей на примере музыкальной социальной сети

АВТОРЕФЕРАТ
на соискание степени магистра технических наук

по специальности 1-40 81 01 - Информатика и технологии разработки программного обеспечения

_____

Научный руководитель
Воронов А.А
кандидат технических наук

_____

Минск 2017

ВВЕДЕНИЕ

На сегодняшний день в связи с повсеместным распространением интернет технологий, а также активным развитием проектных форм организации работ (проектного менеджмента), которые являются ответом на такие вызовы времени как ускорение темпов научно-технического прогресса, и высокая скорость социально- экономических изменений, возникает необходимость оперативного группового обсуждения ситуаций и проблем, в условиях отсутствия экспертов извне, а также недостатка времени. Это обусловливает использование знаний, опыта и ресурсов ограниченного числа людей в рамках решения совместных задач.

Развитие социальных сетей, различных типов программного обеспечения совместной работы (groupware) и такого направления как краудсорсинг обуславливает актуальность проблемы объединения ресурсов.

В рамках магистерской диссертации была поставлена цель разработать метод синтеза знаний, опыта и ресурсов, использующий технологии, присущие программным средствам совместной работы, а также технологии краудсорсинга и различные машинные методы обработки контента. Так как разработанный метод нуждается в интеграции с какой-то реальной платформой, где его можно было бы опробовать, он был интегрирован с собственной разработкой – веб-приложением «социальная музыкальная сеть».

Помимо технологий, присущих groupware и краудсорсинговым платформам, были использованы технологии интеллектуального анализа данных, такие как: text mining, ocr, а также распознавание аудио (audio-fingerprinting). Text Mining – это нетривиальный процесс обнаружения действительно новых, потенциально полезных и понятных шаблонов в неструктурированных текстовых данных. Text mining включает в себя большое число методов интеллектуального анализа текстов и позволяет группе экспертов быстро проанализировать имеющуюся у них информацию. Ключевым механизмом text mining, для решения задачи интеграции текстовых ресурсов, являются алгоритмы кластеризации текстов. При кластеризации использовались плоские и иерархические алгоритмы.

Также, учитывая специфику платформы, в которую будет внедряться метод, был произведен анализ методов и сервисов по обработке аудио контента с целью его последующего идентификации и распознавания. Помимо аудио контента, одним из типов ресурсов являются изображения, которые часто содержат текстовую информацию. Для упрощения идентификации текста, были добавлены методы OCR позволяющие распознать текст с изображения для его последующего анализа.

ОБЩАЯ ХАРАКТЕРИСТИКА РАБОТЫ

Целью работы является разработка метода объединения ресурсов пользователей, а также его реализация и последующая интеграция в веб-приложение «социальная музыкальная сеть».

Для достижения цели в работе были поставлены и решены следующие задачи:

- разделить этапы структурирования области сбора информации и этапы заполнения этой области ресурсами;
- упростить процесс заполнения ресурсами используя техники анализа данных (text mining, ocr, audio-fingerprinting);
- иерархически структурировать область сбора информационных ресурсов;
- осуществить отбор информации, опираясь на мнение большинства участников группы;
- предоставить результат слияния в удобной визуальной форме.

Научная значимость разработки состоит в применении сначала плоских методов кластеризации, а затем последовательного применения иерархических методов кластеризации в рамках кластеров, определённых на предыдущем этапе. Такой подход позволяет более точно определить наиболее схожие тексты в рамках рассматриваемого корпуса документов.

Практическая значимость работы заключается в том, что разработанный метод и приложение позволит группе пользователей произвести объединение своих идей и ресурсов. Данные разработки могут быть применены не только в рамках музыкальных социальных сетей, но и в других сферах социальной или проектной деятельности.

Следующие результаты проведенной работы были представлены на конференциях и опубликованы в сборниках:

1. Доклад, посвященный методам синтеза знаний, опыта и ресурсов пользователей в контексте «программного обеспечения совместной работы» был представлен на 52 студенческой конференции БГУИР, которая проходила в Минске в 2016 году.

2. Доклад, посвященный методам text mining в рамках разработки алгоритма по объединению ресурсов пользователей, был опубликован в сборнике XV международной конференции «Развитие информатизации и государственной системы научно-технической информации – 2016», которая проходила 17 ноября 2016 года в ГНУ «Объединённый институт проблем информатики Национальной академии наук Беларуси», город Минск.

СОДЕРЖАНИЕ РАБОТЫ

Магистерская работа состоит из четырех глав, в которых приводятся: обзор аналогов разрабатываемого программного обеспечения, описание разрабатываемого метода объединения ресурсов пользователей, описание используемых алгоритмов, архитектура разработанного программного средства.

Во введении рассмотрено современное состояние проблемы исследования алгоритмов объединения ресурсов пользователей, определены основные направления исследований, а также дается обоснование актуальности темы магистерской работы.

В первой главе приводится анализ текущего состояния сферы программного обеспечения совместной работы, а также приводится классификация данного программного обеспечения и обосновывается место разрабатываемого продукта в этой классификации. В тексте дается определение термину краудсорсинг, а также приводится обоснование его связи с программным обеспечением, предоставляющим методы и алгоритмы для объединения ресурсов. В главе приводится классификация программного обеспечения, использующего краудсорсинговые методы, а также описываются плюсы и минусы данного подхода.

Во второй главе приводится детальное описание разрабатываемого метода объединения ресурсов; формулируются цели и задачи метода; вводятся основные составляющие метода: краудсорсинговые технологии, методы data mining (text mining), алгоритмы audio-fingerprinting и optical character recognition; приводится пошаговое описание метода в виде блок схем.

В третьей главе приведено подробное описание алгоритмов по обработке контента, применяемых в разработанном во второй главе методе. В рамках описания методов, ключевую роль получили методы text mining. В тексте приводятся основные этапы работы text mining алгоритмов, такие как: поиск информации, предварительная обработка документов, извлечение информации, применение методов text mining, интерпретация результатов. В главе приводится классификация методов кластеризации, а также дается описание применяемых методов плоской и иерархической кластеризации. В рамках обработки аудио контента приводится описание технологии audio-fingerprinting и ее ключевых этапов. Рассмотрен процесс интеграции с audio-fingerprinting системой, а также другими публичными api для агрегации информации об аудио контенте. В рамках обработки изображений приводится обзор технологии OCR, а также дается описание ее основных этапов.

В четвертой главе приводится описание разработанного программного средства, использующего описанные в главах 2-3 методы и алгоритмы. В ней

приводится описание выбранного технологического стека, а также обоснование выбора той или иной технологии, фреймворка, языка. В разделе «Системное проектирование» приводятся архитектура системы, а также краткое описание всех модулей, входящих в нее. В разделе «Функциональное проектирование» приводится детальный анализ каждого разработанного модуля, описывается его принцип работы, внутренняя архитектура, а также процесс взаимодействия с другими модулями. В разделе «Результаты работы и оценка эффективности» приводятся результаты работы разработанного метода и приложения, а также результаты тестирования и оценка эффективности алгоритмов, описанных в главе 3.

ЗАКЛЮЧЕНИЕ

В данной работе были рассмотрены проблемы объединения ресурсов пользователей в социальных сетях. Была проведена классификация программного обеспечения совместной работы, а также указано место разрабатываемой системы в этой классификации. Также были рассмотрены методы краудсорсинга, и приведена классификация программного обеспечения, которое использует данные методы. Был разработан и реализован алгоритм объединения ресурсов, базирующийся на краудсорсинговых технологиях, а также методах машинной обработки различного типа контента.

Данный метод позволит решить такие задачи как: разделение этапов структурирования области сбора информации и этапов наполнения этой области ресурсами; упрощение процесса заполнения ресурсами, используя техники анализа данных (data mining, ocr, audio-fingerprinting); иерархическое структурирование области сбора информационных ресурсов, а также осуществление отбора информации с опорой на мнение большинства участников группы, и представление результатов слияния в удобной визуальной форме.

Был произведен анализ методов и алгоритмов, используемых в технологии интеллектуального анализа текстов text mining. При решении задачи объединения текстовых данных были реализованы алгоритмы предобработки корпуса текстов, построены частотные диаграммы популярности слов в корпусе текстов, был произведен анализ оптимального числа кластеров, используя силуэтные индексы, произведена плоская кластеризация текстов, а также иерархическая агломеративная кластеризация текстов. На основе результатов кластеризации были построены диаграммы и дендограммы, отражающие распределение текстовых по кластерам. Было произведено тестирование

применяемых методов на различных текстовых выборках.

Были проанализированы принципы работы и реализованы OCR методы, а также были реализованы методы предобработки естественных изображений, содержащих текстовые области. В рамках данных методов был осуществлен поиск связных регионов, а также их последующая фильтрация по геометрическим параметрам, произведена фильтрация, используя метод вариации штриха, произведено объединение полученных регионов используя метод перекрытия.

Были реализованы методы audio-fingerprinting с целью распознавания и идентификации аудио ресурсов, а также произведена интеграция с такими внешними сервисами как Acrcloud, предоставляющий базу отпечатков аудио треков, а также last.fm и Вконтакте для получения и агрегации информации о музыкальных треках, загружаемых в систему.

Все описанные выше методы были реализованы и интегрированы в веб-приложение, написанное с использованием современного технологического стека.

По теме изучаемой проблемы были сделаны две публикации, в рамках которых была проведена классификация программного обеспечения совместной работы, а также произведен обзор алгоритмов text mining для решения задач объединения ресурсов.

Разработанный метод и приложение позволит группе пользователей, произвести объединение своих идей и ресурсов и могут быть применены не только в рамках музыкальных социальных сетей, но и в других сферах социальной или проектной деятельности.

Список публикаций соискателя

1-А. Краснов, А.Ю. Методы синтеза знаний, опыта и ресурсов пользователей в контексте «Программного обеспечения совместной работы» / А.Ю. Краснов // Компьютерные системы и сети: материалы 52-й научной конференции аспирантов, магистрантов и студентов. Минск, 25-30 апреля 2016 года. – Минск: БГУИР, 2016. – С. 28-30.

2-А. Воронов, А.А. Методы text mining в рамках разработки алгоритма по объединению ресурсов пользователей / А.А. Воронов, А.Ю. Краснов // Развитие информатизации государственной системы научно-технической информации: материалы XV Междунар. конф., Минск 17 ноября 2016 г. / ГНУ «Объединённый институт проблем информатики Национальной академии наук Беларуси». – Минск, 2016. – С. 158-164