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

УДК 004.415.2

Белов
Александр Владимирович

Информационная система анализа социальных сетей

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

_____

Научный руководитель
Лукашевич Марина Михайловна
кандидат технических наук, доцент

_____

Минск 2018

ВВЕДЕНИЕ

В настоящей диссертации освещается разработка информационной системы анализа социальных сетей. Основными задачами разрабатываемой системы являются:

– сбор данных из избранных социальных сетей;

– возможность импорта данных из сторонних источников с целью увеличения объема доступной для анализа информации;

– дополнение собранных данных о местоположении пользователей географическими координатами;

– анализ полученных данных доступными способами;

– экспорт полученных результатов универсальных форматах данных;

– визуализация результатов анализа в виде таблиц и интерактивных графиков;

– формирование отчетов по результатам анализа.

Наиболее пригодными для анализа социальными сетями признаны Facebook, LinkedIn и Github. Проект Github является по своему основному назначению сервисом для организации совместной работы, однако имеет в своем составе все необходимые для социальной сети элементы, и часто используется, как социальная сеть для разработчиков программного обеспечения.

В качестве одного из инструментов анализа данных, представляющих интерес своей реализацией и результатами, используется построение рекомендательных систем. В диссертации подробно рассматривается выбор конкретных методов и деталей реализации рекомендательной системы, использующей в качестве исходных данных данные о предпочтения пользователей в системе Github. Результаты представляют собой направления, признанные пригодными для изучения конкретным пользователем и проекты, схожие с интересующими пользователя на данный момент.

Несмотря на то, что задачей проектируемой информационной системы является общий анализ социальных сетей, с целью получения более интересных для анализа результатов принято решение использовать для анализа профили студентов специальности ВМСиС за несколько последних лет. Таким образом, становится возможность использовать результаты диссертации в повышении качества обучения на отдельно взятой специальности с перспективой использования получившейся информационной системы с другими социальными группами с целью достижения иных целей и задач.

Проектированию, разработке и реализации информационной системы анализа социальных сетей и посвящена настоящая диссертация.

ОБЩАЯ ХАРАКТЕРИСТИКА РАБОТЫ

Цель данной работы – реализовать информационную систему анализа социальных сетей, пригодную для изучения публично доступной информации о студентах специальности ВМСиС.

Объектом исследования являются социальные сети. Предметом исследования – алгоритмы сбора, анализа и визуализации данных социальных сетей.

Для достижения данной цели были признаны необходимыми следующие задачи:

– выбор социальных сетей, представляющих наибольший интерес для анализа и его обоснование;

– выбор наиболее подходящих технологий из существующих на сегодняшний день для построения интерфейса информационной системы и анализа полученных данных;

– реализация необходимых алгоритмов получения, анализа и визуализации данных.

Для реализации алгоритмов анализа данных наиболее подходящим признан язык программирования Python, а также библиотеки алгоритмов машинного обучения scikit-learn, быстрых численных вычислений numpy, научных расчетов scipy, а также библиотека для построения рекомендательных систем implicit.

Реализация сбора данных реализована с применением библиотек Octokit для работы с API сервиса Github, Koala для использования небольшого числа доступных для использования методов работы с данными социальной сети Facebook, Linkedin-scraper, позволяющий произвести импорт данных пользователя из публично доступного профиля в профессиональной социальной сети LinkedIn, а также библиотеки Mechanize, служащей для получения данных, недоступных с помощью иных способов через эмуляцию действий пользователя в процессе работы с веб-интерфейсами социальных сетей.

Часть данных, необходимых, в частности, для построения рекомендательной системы, получена с использованием облачного сервиса Google BigQuery и SQL запросов особого вида.

Для реализации пользовательского интерфейса используется веб-фреймворк Ruby on Rails для языка программирования Ruby, а также Javascript библиотека построения интерактивных графиков d3.js.

СОДЕРЖАНИЕ РАБОТЫ

Диссертация состоит из пяти глав.

В первой главе рассматриваются существующие социальные сети, их особенности, доступные инструменты разработчика для работы с данными каждой, а также основные проблемы в использовании. Описываются основные подходы к анализу данных социальных сетей, а также производится обзор существующих аналогов и выбор на основе их технологических стеков языков программирования, основных библиотек и алгоритмов для использования в настоящей диссертации.

Вторая глава посвящена теоретическим основам рекомендательных систем. Рассмотрены вопросы пригодности рекомендательных систем для анализа данных социальных сетей, имеющиеся методы их построения и основные алгоритмы, пригодные для реализации. Описан процесс выбора конкретного типа рекомендательных систем и алгоритмов.

Третья глава рассматривает вопросы проектирования и программной реализации информационной системы анализа данных социальных сетей. В данной главе более подробно описаны алгоритмы получения и анализа данных, построения рекомендательных систем, а также алгоритмы визуализации данных и типы используемых интерактивных графиков. Описываются основные сущности модели данных, а также классы, используемые в программной реализации и их взаимосвязи.

В четвертой главе описывается процесс тестирования реализованной информационной системы, описываются характеристики ее работы и выявляются точки, пригодные для дальнейшей оптимизации. Оценивается также точность работы рекомендательной системы с данными пользователей Github и наглядность используемых средств для визуализации данных.

ЗАКЛЮЧЕНИЕ

В процессе работы над данной диссертацией была спроектирована и разработана информационная система анализа данных социальных сетей. В качестве социальных сетей, пригодных для сбора релевантных данных, были использованы социальные сети Facebook и LinkedIn, а также сервис совместной работы Github. Для анализа репозиториев программных проектов системы Github была построена рекомендательная система коллаборативной фильтрации по методу соседства с использованием алгоритма чередующихся наименьших квадратов.

В качестве объектов для тестирования получившейся системы были выбраны студенты специальности ВМСиС и их профили в названных социальных сетях и сервисах. Среди информации, представляющей большой интерес, следует отметить данные о распределении студентов и мест работы по странам мира, декларируемые в профессиональной социальной сети LinkedIn навыки, а также стек используемых в работе над проектами с отрытым исходным кодом технологий, полученный с помощью Github. Для пользователей, имеющих аккаунты в системе Github и достаточное количество отмеченных программных проектов, были сформированы рекомендации по наиболее интересным для дальнейшего изучения программным проектам и направлениям.

К преимуществам полученной информационной системы можно отнести:

– возможность выполнять сбор и анализ данных, распределенных между несколькими социальными сервисами;

– наличие возможности исследовать практические результаты работы каждого пользователя в системе Github и соотносить их с декларируемыми навыками и учебными программами;

– наличие пользовательского интерфейса и интерактивных графиков, позволяющих представить анализируемую информацию в виде, наиболее благоприятном для восприятия человеком.

К недостаткам, в свою очередь, могут быть отнесены:

– фокусирование на более узкой социальной группе и отсутствие подтверждения корректной работы на других выборка;

– недостаточная гибкость пользовательского интерфейса;

– относительно небольшое количество интерактивных графиков;

– недостаточная воспроизводимость результатов работы рекомендательной системы при ее переобучении.

СПИСОК ПУБЛИКАЦИЙ СОИСКАТЕЛЯ

[1–А] Белов А.В. Ключевые особенности анализа данных профессиональных социальных сетей / А. В. Белов // Компьютерные системы и сети: материалы 53-й научной конференции аспирантов, магистрантов и студентов – Минск, 2017 – С. 17 – 18.