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

УДК ____

Череватенко
Николай Павлович

Система обработки и визуализации лексем сервиса микроблогинга Twitter

АВТОРЕФЕРАТ

на соискание степени магистра технических наук
по специальности 1-40 80 03 Вычислительные машины и системы

_____

Научный руководитель
Самаль Дмитрий Иванович
к.т.н., доцент, доцент кафедры ЭВМ

_____

Минск 2015

ВВЕДЕНИЕ

На сегодняшний день одной из самых важных и заметных областей Web 3.0, ключевым принципом которой является участие пользователей в работе сайтов, являются сетевые дневники, или веб-логи, сокращённо называемые *блогами*. Концептуальным развитием блогов, обусловленным их широкой социализацией, являются микроблоги, которые имеют ряд характерных особенностей: ограниченная длина сообщений, большая частота публикаций, разнообразная тематика, различные пути доставки сообщений и т.д.

Первый и наиболее известный сервис микроблогов *Twitter* был запущен в октябре 2006 г. компанией *Obvious* из Сан-Франциско. К настоящему времени постоянно растущая аудитория сервиса составляет десятки миллионов человек. Очевидно, что автоматизированное выделение наиболее значимых терминов из потока сообщений, генерируемого сообществом Twitter, имеет практическое значение как для определения интересов различных групп пользователей, так и для построения индивидуального профиля каждого из них.

Современные приложения по обработке данных сервиса *Twitter* предоставляют лишь агрегированные данные о количестве упоминаний имени пользователя в сообщениях, анализ подписчиков по месторасположению, поиск по слову в глобальной истории сообщений, вывод последних сообщений с искомым словом. По сути отсутствует какой либо сервис по обработке содержания сообщений и нахождения в них ключевой информации с дальнейшей ее кластеризацией и обработкой. Исходя из количества публикуемых твитов в день - 55 миллионов, такой сервис по обработке поступающей информации был бы очень востребован людьми зарабатывающими на рекламе в *Twitter*, с целью привлечения новых подписчиков к себе на страницу, а так же людям интересующимися последними новостями и экономящими своя время.

Основной задачей данного проекта является обработка поступающих из сервиса сообщений, лексический анализ каждого сообщения с целью выделения ключевых слов в сообщении, дальнейший анализ схожести ключевых слов, последующее сокращение их, визуализация обработанных данных в удобочитаемом виде. Дополнительно, анализ пользовательского

текста сообщения перед последующей его отправкой и публикации в *Twitter* с целью анализа содержимого, нахождения ключевых слов и выделение их по популярности в данный момент, а так же помощь в расстановке специального микросинтаксиса - хештега. Хэштеги дают возможность группировать подобные сообщения, таким образом перейдя по хэштегу в сообщении можно получить набор сообщений, которые его содержат. Хэштеги также используются чтобы выразить контекст вокруг данного сообщения, без намерения фактически классифицировать сообщения для последующего поиска, обмена или по другим причинам. Это может помочь выражать юмор, волнение, печаль или другие эмоции.

Однако нужно отметить, что классические статистические методы экстракции ключевых терминов, основанные на анализе коллекций документов, малоэффективны в данном случае. Это обусловлено чрезвычайно малой длиной сообщений (до 140 символов), их разнообразной тематикой и отсутствием логической связи между собой, а также обилием редко используемых аббревиатур, сокращений и элементов специфического микросинтаксиса.

Для решения этой проблемы в представленной работе относительная значимость терминов в анализируемом контексте определяется с помощью данных о частоте их использования в качестве ключевых в интернет-энциклопедии Википедия. Работа алгоритма основана на расчёте "информативности" каждого термина, т.е. оценки вероятности того, что он может быть выбран ключевым в тексте. В дальнейшем к анализируемому набору терминов применяется ряд эвристик, результатом которых является список терминов, сочтённых ключевыми. Для получения информации из Twitter было использованы Twitter REST API и Twitter Streaming API. Twitter REST API предоставляет программный доступ к базе данных Twitter. Позволяет создавать новый твит, читать профиль пользователя и данные подписчиков. Twitter Streaming API предоставляет разработчику доступ к глобальному потоку Twitter данных. Поступающими данными являются сообщения пользователей, события ретвита и другие.

Разработанное приложение представляет из себя одностраничное приложение с веб-сервером и удаленной базой данных. Позволяющее пользователю просматривать информацию о популярных темах в реальном времени, с дополнительной информацией о количестве твитов содержащих популярное ключевое слово или хештег, а так же форму написания нового

твита с последующим анализом лексем введенного сообщения, выделением из них ключевых слов и подсказками по добавлению в сообщение микроразметки, ввиде хештегов, для лучшего выделения тематики сообщения. Сервер написан с использованием платформы Node.js, позволяющей писать серверные приложения на языке Javascript. Веб страница написана на фреймворке Backbone.js, который предоставляет программисту удобный способ структуризации приложения. В качестве удаленной базы данных выступает сервис FireBase, предоставляющую облачную NoSQL базу данных для приложений реального времени. Данный сервис предоставляет API для разработчиков, позволяющий синхронизировать данные приложения между клиентами и хранить их в облаке FireBase. Визуализация в свою очередь выполенна при помощи javascript библиотеки D3.js, позволяющей строить динамическую и интерактивную визуализацию данных в веб браузере при помощи технологий SVG, HTML5, CSS3. Верстка и графическое оформление интерфейса при помощи библиотеки Twitter Bootstrap.

ОБЩАЯ ХАРАКТЕРИСТИКА РАБОТЫ

Актуальность

Twitter представляет из себя социальную сеть для публичного обмена короткими сообщениями. Люди могут использовать различными образами: читать и публиковать новости, делиться своим мнением, вести обсуждения. Дневная аудитория составляет порядка ста миллионов человек, с колоссальным количеством ежедневных соощений в размере пятиста миллионов. Все это делает Twitter огромным источником последней информации о мире и людях.

В настоящее время активно развиваются алгоритмы лексического анализа текста. Так можно встретить ряд крупных проектов, например сервис по обработке текста от IBM под названием Alchemy. Данные системы позволяют выделять из теста ключевые слова и словосочетания, анализировать несет ли текст позитивный или негативный оттенок, выделение фактов из текста.

Данные системы будут развиваться, совершенствоваться и вскоре займут свою нишу в современном обществе. Они позволят упростить жизнь человека, ускорить восприятие информации, а в ряду случаев позволяют иключить участие человека, например автоматическое создание встречи в календаре исходя из полученного им сообщения.

Цель и задачи исследования

Целью диссертационной работы является накопление и анализ сообщений сервиса Twitter с целью выделения из них ключевых слова и тем разговоров. Провести анализ существующих алгоритмов выделения ключевых терминов из текста.

Достижение поставленной цели потребовало решения следующих основных задач:

1) Анализ существующих алгоритмов;
2) Проведение сбора и анализа сообщений;
3) Сопоставление полученных данных с реальными.

Объект исследования: пользовательские сообщения малой длины. Под сообщениями малой длины следует понимать текст длиной в 140 символов, из-за ограничений на максмальную длину сообщений в сервисе Twitter.

Методы исследования

Теоретические методы исследования основывались на методах лексической обработки теста, лемматизации, выделения ключевых слов и системного анализа. Экспериментальная часть исследования базировалась на обработке и анализе сообщений с помощью ЭВМ с последующей визуализацией результатов и сравнения их с текущими тенденциями сообщений. Для программной реализации разработанных алгоритмов использовались методы создания программных систем и программирование на языках высокого уровня, моделирование с помощью специализированных пакетов обработки лексем.

Личный вклад соискателя

Основные результаты и положения, выносимые на защиту, получены лично автором. Все алгоритмы, обсуждаемые в работе, были экспериментально исследованы автором самостоятельно. Научный руководитель принимал участие в постановке цели и задач работы, их предварительном анализе, планировании, а также в обсуждении полученных результатов.

Опубликованные результаты

По теме диссертационной работы опубликована 1 печатная работа. Из них 1 тезисы доклада на конференции [1].

Структура и объём диссертации

Диссертация изложена на 57 страницах. Она состоит из введения (3 стр.), общей характеристики работы (3 стр.), трёх глав (49 стр.), заключения (2 стр.). Работа содержит 5 иллюстраций (2 стр.) и 1 таблицу (1 стр.), список использованных источников, состоящий из 39 наименований (3 стр).

КРАТКОЕ СОДЕРЖАНИЕ РАБОТЫ

В первой главе выполнен обзор программных средств необходимых для построения разрабатываемого приложения, обзор этапов обработки пользовательских сообщений, а также был рассмотрено API предоставляемое сервисом Twitter для доступа к своей базе данных. Основные выводы этой главы:

1) Для осуществления доступа к сервису Twitter разработчику необходимо зарегистрировать свое приложения в самом сервисе и получить специальный код разработчика. Дальнейший доступ к базе данных осуществляется посредством REST API и Streaming API. REST API предоставляет возможность поиска по базе данных с широкими возможностями фильтрации результатов сообщений, так же позволяет выгружать публичные сообщения конкретных пользователей или групп. Streaming API позволяет получать все сообщения пользователей в режиме реального времени без ограничений и какой-либо возможности фильтрации входного потока сообщений.

2) Процесс лексического анализа входящих сообщений представлен следующим набором операций: получение информации из сервиса, удаление стоп-слов, лемматизация, выделение ключевых слов основываясь на лексической базе IBM Alchemy.

3) Необходимым программным средством является наличие сервера через который будет осуществляться получение данных реального времени и исторических данных из сервиса Twitter. Наличие удаленной базы данных для хранения пользовательской информации и результатов обработки входящих сообщений. Веб-приложения предоставляющего интерфейс пользователю для работы с удаленным сервером и сервером базы данных.

Во второй главе был произведен анализ синтаксиса сообщений сервиса Twitter. Сообщения ограничены 140 символами. Помимо обычных слов в сообщении может присутствовать сылки на других пользователей сервиса, специальная микросинтаксис представленный слештегами, кештегами, хештегами. Данный микросинтаксис позволяет группировать сообщения по категориям. Описывается способ извлечения ключевых терминов из сообщений микроблогов с использованием информации, полученной при помощи интернет-сервиса Alchemy. Работа алгоритма основана на расчёте

для каждого термина его "информативности", т.е. оценки вероятности того, что он может быть выбран ключевым в тексте.

Во третьей главе были описаны требования, которым должно соответствовать разрабатываемое экспериментальное приложение. Также была описана разработка архитектуры и приведена структура разрабатываемых классов данного приложения. . В качестве демонстрации была произведена выборка сообщений за период международного конкурса Евровидение. Был произведен анализ полученных сообщений с целью выявления ключевых терминов и дальнейшая визуализация результатов.

В ходе тестирования разработанного алгоритма полученные результаты совпали с преобладающими тенденциями разговоров за рассматриваемый период.

ЗАКЛЮЧЕНИЕ

В процессе разработки магистерской диссертации был реализован сервис по анализу и обработке лексем сервиса микроблогинга Twitter, написан сервер, получающий данные в режиме реального времени через Twitter Streaming API а так же производящий поиск в базе данных Twitter посредством REST API. Создано веб-приложение по отображению данных поступающих с сервера и позволяющее пользователю создавать и отправлять новые сообщения исходя из популярны в это время тем и обсуждений. Серверная часть была написана посредством технологии Node.js, веб-фреймворка Express, базы словарных данных IBM Alchemy, что позволило получить новое в своем роде программное обеспечение, позволяющее обрабатывать поступающие сообщения с разбивкой его на лексемы, все серверное приложение было написано на языке Javascript. Несмотря на ограничения накладываемые сервисом Twitter по запрсу данных с их серверов, для пользователя имеется большой запас ресурсов для выполнения его поисковых запросов и обработке введенных им сообщений. Разработанное программное средство открывает новые возможности для лексического анализа сервиса Twitter.

В ходе выполнения работы был использован алгоритм извлечения ключевых терминов из минимально структурированных текстов сообщений микроблогов. В качестве примера возможного практического использования алгоритма в рамках разработанной системы был проведен анализ сообщений в период проведения музыкального конкурса Евровидение. Проведённое экспериментальное исследование показало, что алгоритм работает корректно и эффективно.

Новизна данного проекта заключается в отстутствии каких либо сервисов по анализу лексем сервиса Twitter. Существует множество сервисов по показу статистической информации, данных о количестве упоминаний имени пользователя в сообщениях других пользователей, анализ подписчиков по месторасположению но схожему с разработанным сервисом нет.

Безусловно, как у любой первой версии, у данного модуля есть недостатки связанные с хранением и актуализацией данных. Есть необходимость хранения всей уже полученной ранее информации в неагрегированном виде. Однако количество посупающих сообщений

заставляет агрегировать инфомацию что исключает возможность дальнейшего полнотекстового поиска по ней.

В продолжение начатой работы планируется улучшение сервиса, с целью анализа сообщений пользователя и его связей для повышения релевантности выводимой информации. Пользователю при этом необходимо предоставить полный доступ к данным своего аккаунта, в том числе и к текстам личных сообщений.

Одной из возможных перспектив использования разработанного алгоритма является возможность предустановки пользователем нескольких категорий с целью просмотра только тех сообщений, которые им соответствуют. Возможна фильтрация не только по извлечённым с помощью IBM Alchemy терминам, но и путём простого поиска по текстам сообщений после препроцессинга. Такой сервис был бы востребованным [19] и позволял бы улучшить релевантность показа контекстной рекламы на канале пользователя в зависимости от популярности среди людей того или иного вида информации.

СПИСОК ПУБЛИКАЦИЙ СОИСКАТЕЛЯ

[1] Череватенко Н.П. Архитектура одностраничного приложения по обработке данных сервиса микроблогинга Twitter/ Н.П. Череватенко // Материалы 51-ой научной конференции аспирантов, магистрантов и студентов БГУИР. – 2013. – с. 41.