1 Введение

В современном мире информация быстро устаревает, поэтому способы вовремя находить нужные данные — постоянный объект для исследований. Одним из направлений в этой области является извлечение данных из микроблоггинговых платформ.

Платформы микроблогов стали очень популярным способом размещения данных в Сети. В них можно найти сообщения пользователей практически на любую тему, начиная стихийными бедствиями и заканчивая рейтингами музыкальных исполнителей. Правильная обработка доступной информации — нетривиальная задача, которая имеет множество областей применения. Последние несколько лет эта тема активно исследуется во многих университетах мира.

Отслеживание сообщений о стихийных бедствиях в реальном времени поможет вовремя организовать спасательные операции и сохранить жизни людей[1]. Руководствуясь сообщениями пользователей микроблогов можно судить о популярности товаров и вовремя принимать экономически целесообразные решения. Можно делать предположения о рейтингах политических деятелей и эффективности рекламы на основании информации в микроблогах. Помимо перечисленных способов применения доступной информации в микроблоггинговых платформах можно привести множество других.

В данной работе в дальнейшем будет рассматриваться сервис микроблогов Твиттер[1]. В нем помимо текстовой информации можно публиковать фото, видео и геотеги, что так же может быть использовано при анализе, но в этой работе не рассматривается. В доступном наборе данных можно проводить анализ разных сущностей, эта работа посвящена выявлению событий среди потоков информации. Поскольку трактовка событий в сообщениях микроблогов может быть субъективной, выделим несколько свойств, которыми характеризуется событие.

Событие в первую очередь должно быть чем-то аномальным на фоне остальных данных. Оно определяется резким изменением частотных характеристик некоторых слов в сообщениях. События в микроблогах носят взрывной характер, в течении нескольких часов частоты релевантных слов возрастают в десятки раз и так же быстро опускаются до нормального уровня. Примером события могут быть: стихийное бедствие, выход спорного законопроекта на резонансную тему, получение фильмом награды на кинофестивале.

Исторически подходы к описанной задаче менялись, в следующем разделе будет рассмотрена формальная постановка задачи и эволюция способов ее решения.

2 Описание задачи

Цель данной работы состоит из нескольких частей:

- исследовать существующие подходы по извлечению событий из сети Твиттер,
- исследовать возможность применения иерархического процесса Дирихле для решения описанной задачи,
- разработать метод для извлечения событий на основе иерархоческого процесса Дирихле,

[1] http://www.twitter.com

- продемонстрировать работу алгоритма на реальных данных.

Объект изучения этой работы — алгоритм, который по входным данным строит множество событий. В качестве данных для задач подобного рода служит мультимножество документов (сообщений) $\{D_i \,|\, i \in \overline{1,n}\}$. Будем считать что каждый документ имеет временную метку. Документ D_i определяется как упорядоченный набор слов $\{w_j \,|\, w_j \in D_i\}$. При этом слова в документах принадлежат некоторому словарю V.

Событие — некоторая сущность, которая характеризуется временем возникновения и ключевыми словами. Оно вызывает резкий подъем частотных характеристик некоторых слов. Событием может быть футбольный матч и музыкальный концерт. В социальной сети Твиттер есть популярные темы, которые всегда содержат много сообщений. Например это сообщения с ключевыми словами iphone и ipad. Но такие сообщения нельзя считать событиями. Также событиями нельзя считать еженедельные пятничные сообщения о конце рабочей недели[2].

Особенности социальной сети Твиттер состоят в следующем:

- короткие сообщения (до 140 символов),
- наличие шума и ошибок,
- большая плотность сообщений.
- взрывной характер событий,

3 Обзор существующих подходов

3.1 NED

Первым подходом к извлечению событий принято считать NED (New Event Detection). NED предназначен для того, чтобы находить первое сообщение на новую тему. Следующие сообщения на эту же тему уже не будут новыми и алгоритм их не пометит.

Pic from NED article

4 Полученные результаты

В качестве данных были использованы сообщения пользователей Twitter с 4 июня 2013 года по 31 июня 2013 года, содержащие в себе хэштег #texas. Всего корпус включает порядка 240 тысяч сообщений и 1.5 миллиона слов.

5 Заключение

Список литературы

[1] Imran, Elbassuoni, Castillo, Diaz and Meier; Extracting Information Nuggets from Disaster-Related Messages in Social Media; 2013.

[2] Xun Wang, Feida Zhu, Jing Jiang, Sujian Li; Real Time Event Detection in Twitter; 2011.

[3] Thorsten Brants, Francine Chen, Ayman Farahat; A System for New Event Detection; 2003.