Санкт-Петербургский государственный университет
Кафедра технологий программирования

Башарин Егор Валерьевич

Выпускная квалификационная работа бакалавра

Контекстный анализ данных социальных сетей

010400
Прикладная математика и информатика

Научный руководитель,
старший преподаватель
Попова С.В.

Санкт-Петербург
2016

Содержание

Введение

В настоящее время явление социальных сетей достаточно распространено. Социальные сети уверенно вошли в жизнь современного человека и теперь занимают в ней значимую часть. Главным образом они оказывают влияние на поведение, предубеждения, ценности и намерения человека, что отражается во всех сферах его деятельности. Оказываемое влияние, быстрый рост популярности и открытый доступ к контенту привлекли к социальным сетям внимание правительства, финансовых организаций и исследователей. Преобразование контента социальных сетей в текстовую информацию и выделение ключевых концепций стало важным условием для порождения знаний и формулирования стратегий. Анализ полученных данных помогает исследователям улучшить понимание об информационных потоках, о формировании и распространении мнений, о связи ценностей и предубеждений пользователя и генерируемого им контента. Тем не менее число качественных и количественных исследований в данной области слишком мало. Самый значительный барьер при использовании социальных сетей — это отсутствие методологии для выбора, сбора, обработки и анализа информации, полученной с сайтов социальных сетей. Однако, существуют компании по производству программного обеспечения, разрабатывающие проприетарные системы сбора информации для визуализации данных, и исследователи, занимающиеся разработкой экспертных систем для анализа настроений [1].

Пользователи социальных сетей ежедневно публикуют данные о своей активности, чувствах и мыслях, выражая свое мнение и позицию. Благодаря этому в социальных сетях образуются группы пользователей (сообщества), имеющие общие интересы. Для выявления ключевых концепций и тематик присущих группе пользователей используется контекстная обработка генерируемого ими контента. В данной работе контекстная обработка данных основана на идеях и принципах тематического моделирования. Результаты такой обработки могут использоваться для мониторинга заболеваний, например гриппа [2], или для предсказания поведения рынка.

Постановка задачи

Целью данной работы является изучение методов контекстной обработки данных социальных сетей, в основе которых лежат при принципы и идеи тематического моделирования. Для того чтобы достичь поставленной цели предлагается выполнить следующий ряд задач:

1. Анализ предметной области
   - 1.1 Обзор социальных сетей
   - 1.2 Выбор социальной сети в качестве источника данных
2. Подготовка данных
   - 2.1 Загрузка данных с веб-страниц социальной сети
   - 2.2 Предобработка данных
   - 2.3 Разбиение данных на обучающую и тестовую части
3. Выбор тематической модели
   - 3.1 Анализ тематических моделей
   - 3.2 Выбор подходящей тематической модели
4. Построение тематической модели
   - 4.1 Анализ методов построения тематической модели
   - 4.2 Реализация программного модуля, реализующего выбранную тематическую модель
   - 4.3 Обучение тематической модели
5. Оценка качества модели
   - 5.1 Обзор и анализ оценок качества тематических моделей
   - 5.2 Оценка качества построенной модели
6. Анализ полученных результатов

Обзор литературы

Тема данной работы тесно пересекается с информационным поиском, основы которого подробно рассмотрены в книге Кристофера Майнинга "Introduction to Information Retrieval" [3]. Особое внимание стоит уделить главам 2 и 18. В главе 2 описываются методы подготовки и предобработки текстовой информации. Глава 18 сосредотачивает внимание на подходах латентно-семантического анализа, который является ценным инструментом в тематическом моделировании. В конце каждой главы приведены ссылки на литературу для более подробного изучения темы.

Вероятностное латентно-семантическое моделирование стало логичным продолжением идей латентно-семантического моделирования и нашло свое применение в тематическом моделировании. Это стало причиной появления вероятностных тематических моделей. Основные принципы вероятностного латентно-семантического анализа (probabilistic latent semantic analysis - pLSA) были описаны Томасом Хоффманом в 1999 году в статье [4]. Затем они были развиты Дэвидом Блеем в его статье 2003 года [5], в которой была введена и рассмотрена тематическая модель латентного размещения Дирихле (latent dirichlet allocation - LDA). Статья Д.Блея описывает основные преимущества LDA перед pLSA, а также методы построения и оценки качества тематической модели LDA. В статье Д.Блея 2012 года [7] рассматриваются связь LDA с другими вероятностными тематическими моделями, а также применение LDA в тематическом моделировании.

В техническом отчете Грегора Хейнрича "Parameter estimation for text analysis" [6] рассматриваются методы оценки параметров моделей для тематического анализа текстов. В отчете подробно разобраны темы, связанные с основными подходами оценки параметров, сопряженными распределениями и Байесовскими сетями, а также применение данных тем для построения тематической модели LDA.

Среди русскоязычной литературы следует обратить внимание на работы К. В. Воронцова. В работе [8] подробно описаны основные идеи вероятностного тематического моделирования. В первой части данной работы ставится задача тематического моделирования. Далее рассматриваются основные вероятностные тематические модели pLSA, LDA и их модифицированные версии, а также методы их построения. Работа завершается рассмотрением способов оценки качества тематических моделей.

Глава 1. Подготовка данных

1.1 Обзор социальных сетей

В данной работе под социальной сетью понимается веб-сайт, который предназначен для поддержания социальных взаимоотношений с помощью Интернета. Не смотря на то, что социальные сети появлись около 20 лет назад их популярность растет с каждым годом. На рисунке 1 показан график, отображающий рост числа пользоватлей социальных сетей во всем мире. По итогам 2015 года число пользователей социальных сетей превысило отметку в 2 миллиарда человек и по прогнозам их количество будет только расти [9].

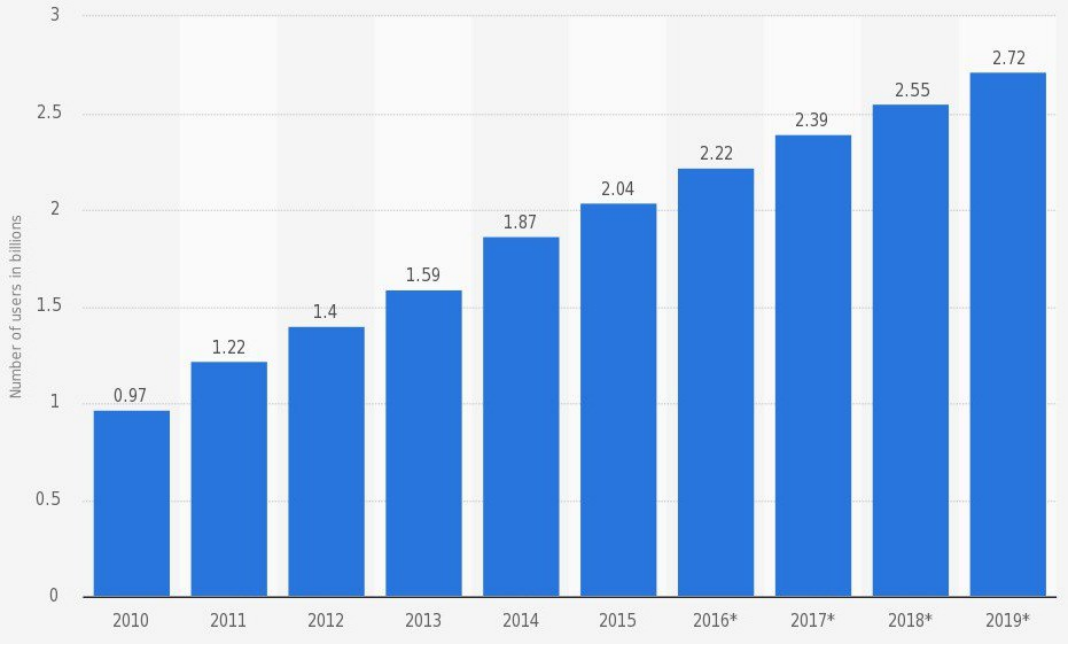

Рис. 1: Число пользователей социальных сетей по годам

Так как число социальных сетей Число социальных сетей огромно, на рисунке 2 приведена сравнительная статистика самых популярных социальных сетей по числу активных пользоватей на апрель 2016 года [10].

1.2 Выбор социальной сети и получение данных

1.3 Предобработка данных

1.4 Результаты

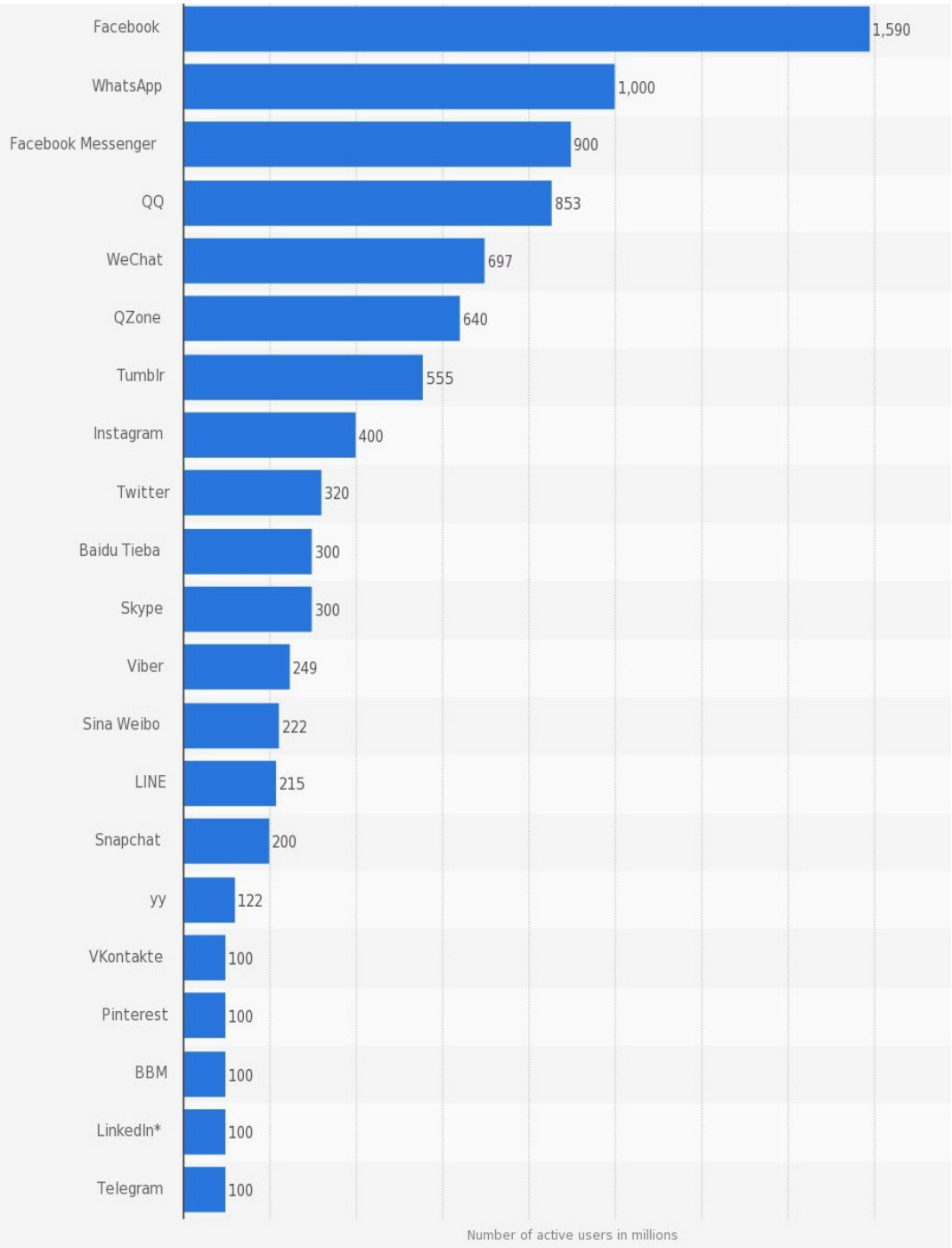

Рис. 2: Рейтинг самых популярных социальных сетей на апрель 2016 года

Глава 2. Выбор и построение тематической модели

3.1 Тематическое моделирование

3.1 Выбор тематической модели

3.2 Описание тематической модели LDA

3.3 Реализация тематической модели

3.4 Эксперименты

Глава 3. Оценка качества модели

4.1 Обзор оценок качества тематических моделей

4.2 Оценка качества построенной тематической модели

Анализ результатов

Заключение

Список литературы

[1] Arturas Kaklauskas. Biometric and Intelligent Decision Making Support // Springer, 2015, P. 220.

[2] Paul, MJ. and M. Dredze. You Are What You Tweet: Analyzing Twitter for Public Health. // In Proc. of the 5th International AAAI Conference on Weblogs and Social Media (ICWSM), 2011.

[3] Christopher D. Manning, Prabhakar Raghavan and Hinrich Schütze. Introduction to Information Retrieval. Cambridge University Press, 2008. 506 P.

[4] Thomas Hofmann. Probabilistic latent semantic indexing. Proceedings of the Twenty-Second Annual International SIGIR Conference, 1999.

[5] David M. Blei, Andrew Y. Ng, Michael I. Jordan. Latent Dirichlet Allocation // Journal of Machine Learning Research 3, 2003. P. 993 – 1022.

[6] Gregor Heinrich. Parameter estimation for text analysis. Technical report, Fraunhofer IGD, Darmstadt, Germany, 2005.

[7] David Blei. Introduction to Probabilistic Topic Models // Communications of the ACM, 2012. P. 77–84.

[8] Воронцов К.В. Вероятностное тематическое моделирование // www.machinelearning.ru : web. — 2013.

[9] http://www.statista.com/statistics/278414/number-of-worldwide-social-network-users/

[10] http://www.statista.com/statistics/272014/global-social-networks-ranked-by-number-of-users/

[11] ссылка на machine learning

[12] ссылка на апи вк

[13] можно всякие документации добавить