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,
ст. 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 организации и исследователей. Преобразование контента социальных сетей в текстовую информацию и выделение ключевых концепций стало важным условием для порождения знаний и формулирования стратегий. Анализ полученной информации 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, разработавшие проприетарные системы сбора информации для визуализации данных, и исследователи, разработавшие экспертные системы для анализа настроений [1].

Пользователи социальных сетей активно публикуют данные о своей ежедневной активности, чувствах и мыслях, выражая свое мнение и позицию. Благодаря этому в социальных сетях образуются группы пользователей (сообщества), имеющих общие интересы. Для выявления ключевых концепций и тематик присущих группе пользователей используется контекстная обработка данных, генерируемых участниками группы. В данной работе контекстная обработка данных основана на идеях и принципах тематического моделирования. Результаты такой обработки данных 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

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] Pang-Ning Tan, Michael Steinbach and Vipin Kumar. Introduction to Data Mining. Addison-Wesley, 2006. 769 P.

[5] Toby Segaran. Programming Collective Intelligence. O'Reilly Media, 2007. 362 P.

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

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

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

[9] Воронцов

[10] ссылка на machine learning