КЛАССИФИКАЦИЯ ТЕКСТОВЫХ ДАННЫХ СОЦИАЛЬНОЙ СЕТИ TWITTER

И.А. Рыцарев, А.В. Благов

Самарский государственный аэрокосмический университет имени академика С.П. Королёва (национальный исследовательский университет) (СГАУ), Самара, Россия

В социальные сети играют большую роль в современном мире, важным при этом является определение значимых и популярных обсуждаемых тем. В данной статье рассматриваются вопросы сбора текстовых данных социальной сети twitter и дальнейшей кластеризации и классификации собранных данных.

Ключевые слова: big data, обработка данных, анализ данных, кластеризация, классификация, tf-idf, latent dirichlet allocation.

Введение

Данные сверхбольшого объёма (англ. big data) в информационных технологиях - наборы данных, размер которых превосходит возможности типичных баз данных (БД) по занесению, хранению, управлению и анализу информации [1]. Существует много серий подходов, инструментов и методов обработки структурированных и неструктурированных данных сверхбольшого объема.

Понятие больших данных подразумевает работу с информацией огромного объема и разнообразного состава, весьма часто обновляемой и находящейся в разных источниках в целях увеличения эффективности работы и создания новых.

На данный момент социальные сети находятся на пике популярности, уже сейчас миллионы пользователей используют Facebook и Twitter. Многим компаниям необходимо анализировать данные, полученные из социальных сетей, для оценки отношения пользователей к своим продуктам [4]. Кроме этого анализ данной области используется в решении вопросов безопасности [5]. Собрав и кластеризировав текстовые данные из социальной сети, можно определить основные темы и события, обсуждаемые пользователями социальных сетей в различных городах и странах..

Кластеризация текстовой информации на основе частотного анализа

Кластеризация (или кластерный анализ) — это задача разбиения множества объектов на группы, называемые кластерами. Внутри каждой группы должны оказаться «похожие» объекты, а объекты разных группы должны быть как можно более отличны, при этом должна быть определена некоторая мера. Главное отличие кластеризации от классификации состоит в том, что перечень групп четко не задана и определяется в процессе работы алгоритма. Основная цель кластеризации – поиск существующих структур [6].

Одним из основных методов частотного анализа является подсчет числа вхождений каждого слова в документе. На основе полученной информации можно составить так называемое «облако тегов» - визуальное представление веса слова в документе [7].

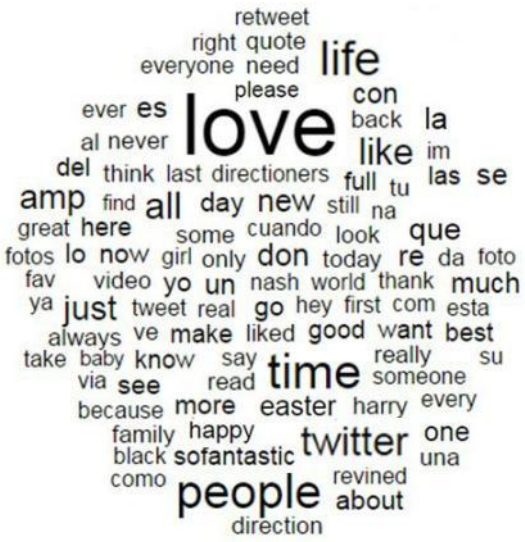

Рис. 1. «Облако тегов» англоязычного сегмента социальной сети Twitter

Однако в результате обработки данных на выходе мы получим, что у различных артиклей, предлогов и других служебных частей речи будет наибольшее число вхождений. Поэтому для наиболее честного оценивания значимости слова необходимо использовать меры, которые будут не только считать число вхождений слова в документе, но и учитывать количество вхождений слова в других документах. Примером такой меры является TF-IDF. При подсчете вес некоторого слова будет пропорционален количеству употребления этого слова в документе, и обратно пропорционален частоте употребления слова в других документах коллекции [8]. Таким образом, мера TF-IDF является произведением двух сомножителей:

$$tfidf(t,d,D) = \mathrm{tf}(t,d) \times idf(t,D), \qquad (1)$$

где $tf(t,d) = \frac{n_i}{\sum_k n_k}$, а $idf(t,D) = \log \frac{|D|}{|(d_i \ni t_i)|}$.

После определения с помощью меры TF-IDF кластеризацию можно проводить по различным алгоритмам, таким как, например, k-means [13].

Классификация текстовой информации на основе подхода с машинным обучением

Так же существует и другой подход к решению проблемы классификации – классификация информации через машинное обучение.

Машинное обучение — процесс, в результате которого машина (компьютер) способна показывать поведение, которое в нее не было явно заложено (запрограммировано). Различают два типа обучения: индуктивное и дедуктивное.

В работах исследователей, занимающихся кластерным анализом текстовой информации в различного рода поисковиках часто имеет место индуктивная мера Word2vec [9-10].

Принцип работы меры заключается в нахождении связей между контекстами слов согласно предположению, что слова, находящиеся в похожих контекстах, имеют тенденцию значить похожие вещи, т.е. быть семантически близкими. Более формально задача стоит так: максимизация косинусной близости между векторами слов (скалярное произведение векторов), которые появляются рядом друг с другом, и минимизация косинусной

близости между векторами слов, которые не появляются друг рядом с другом. Рядом друг с другом в данном случае значит в близких контекстах.
Word2vec анализирует контексты употребления слов и делает вывод, что являются или не являются близкими по смыслу. Так как подобные выводы word2vec делает на основании большого количества текста, выводы оказываются вполне адекватными. Алгоритмы, на которых базируется word2vec подробно изложены в работах [11-12].
Пример векторных расстояний, полученных по word2vec, приведен в таблице 1.

Табл. 1. Векторные расстояния между словом «France» и другими словами по мере word2vec

Слово	**Векторное расстояние**
Paris	0.978443
Spain	0.665923
Belgium	0.665923
Netherlands	0.652428
Italy	0.633130
Portugal	0.577154
Russia	0.571507
Germany	0.563291

Одним из видов дедуктивного подхода можно считать Латентное размещение Дирихле (LDA) - это порождающая модель, позволяющая объяснять результаты наблюдений с помощью неявных групп, что позволяет получить объяснение, почему некоторые части данных схожи. Как правило, при использовании данного подхода определяется ограниченное количество тем – «топиков» и далее утверждается, что каждый документ представляет собой смесь этого небольшого количества тем [14].
Для более детального анализа лучше всего сочетать различные подходы и методы в зависимости от количества обрабатываемых данных.

Заключение

Вопросы, связанные с кластеризацией и дальнейшей классификацией текстовых данных являются актуальными в связи с колоссальным распространением социальных сетей и интернет сервисов во всем мире.
Подходы и методы, представленные в статье планируются к апробации над текстовыми данными, собираемыми из социальной сети Twitter в российском сегменте. Сбор необходимых данных ведется при помощи разработанного программного комплекса с учетом времени и геолокационных зон. Планируется развить данную тему в направлении вывода и оптимизации параллельных алгоритмов кластеризации.

Литература

1. Dean, J. MapReduce: simplified data processing on large clusters / J. Dean, S. Ghemawat // Communications of the ACM. – 2008. – Т. 51. №. 1. – P. 107-113.

2. Vossen, G. Big data as the new enabler in business and other intelligence / G/ Vossen //Vietnam Journal of Computer Science. – 2014. Т. 1. №. 1. – P. 3-14.
3. Tamhane, D.S. Big Data Analysis Using Hace Theorem / D. S. Tamhane, S. N. Sayyad //International Journal of Advanced Research in Computer Engineering & Technology (IJARCET) Volume. – 2015. Т. 4.
4. Tan, W. Social-network-sourced big data analytics / W. Tan, M. B. Blake, I. Saleh, S. Dustdar. // IEEE Internet Computing. – 2013. №. 5. – P. 62-69.
5. Васильков, А. Как «большие данные» помогают улучшить безопасность [Электронный ресурс]. – // Компьютерра: сетевой журн. 2014 – URL: http://www.computerra.ru/108760/security-n-big-data/ (дата обращения: 24.09.2015).
6. Чубукова, И. Задачи Data Mining. Классификация и кластеризация [Электронный ресурс]. – URL: http://www.intuit.ru/studies/courses/6/6/info / (дата обращения: 14.01.2016).
7. Big Data Instruments for Social Media Analysis / A. Blagov, I. Rytcarev, K. Strelkov, M. Khotilin // Proceedings of the 5th International Workshop on Computer Science and Engineering. – 2015. – P. 179-184.
8. Using tf-idf to determine word relevance in document queries / Ramos J. //Proceedings of the first instructional conference on machine learning. – 2003.
9. Wang H. Introduction to Word2vec and its application to find predominant word senses [Электронный ресурс]. – URL: http://compling.hss.ntu.edu.sg/courses/hg7017/pdf/word2vec%20and%20its%20application%20to%20wsd.pdf (дата обращения: 02.02.2016).
10. Yu, M. Improving lexical embeddings with semantic knowledge / M. Yu, M. Dredze //Association for Computational Linguistics (ACL). – 2014. – P. 545-550.
11. Efficient Estimation of Word Representations in Vector Space / Tomas Mikolov, Kai Chen, Greg Corrado, Jeffrey Dean. // In Proceedings of Workshop at ICLR. – 2013.
12. Distributed Representations of Words and Phrases and their Compositionality / Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, Jeffrey Dean // In Proceedings of NIPS. – 2013.
13. MacQueen, J.. Some Methods for Classification and Analysis of Multivariate Observations / J. MacQueen // In Proc. of the 5th Berkeley Symposium on Mathmatical Statistics and Probability. – 1967. – P. 281–297.
14. Blei, D.M. Latent dirichlet allocation / D.M. Blei, A.Y. Ng, M. I. Jordan //the Journal of machine Learning research. – 2003. – Т. 3. – P. 993-1022.