Разработка и исследование алгоритмов кластеризации данных сверхбольшого объема

И.А. Рыцарев[a], А.В. Благов[a]

[a] *Самарский национальный исследовательский университет имени академика С.П. Королева, 443086, Московское шоссе, 34, Самара, Россия*

Аннотация

Работа посвящена исследованию алгоритмов кластеризации текстовых данных. В качестве объекта исследования были выбраны данные социальной сети Twitter. Собирались, обрабатывались и анализировались при этом текстовые данные. Для решения задачи получения необходимой информации были проведены исследования в области оптимизации сбора данных социальной сети Twitter. Разработано программное средство, обеспечивающее сбор необходимых данных из заданных геолокаций. Исследованы и апробированы существующие алгоритмы кластеризации данных, преимущественно большого объема.

Ключевые слова: алгоритмы кластеризации данных; данные сверхбольшого объема; текстовый анализ; k-means; tf-idf метрика; lda; метод судьи

1. Введение

Целью работы является исследование алгоритмов кластеризации текстовых данных социальных сетей, собранных по определенным геолокациям.

В качестве объекта исследования использовались данные социальной сети Twitter. Для достижения цели были поставлены задачи:

- сбор данных социальной сети,
- обработка полученных данных с извлечение необходимой информации,
- исследование, апробация и модернизация алгоритмов кластеризации данных.

В ходе научно-исследовательской работы изучены и апробированы следующие алгоритмы:

• алгоритм k-means,
• алгоритм LDA;
• алгоритм классификации данных методом судьи.

Помимо алгоритмов были исследованы и апробированы следующие меры:

• TF-IDF,
• Word2Vec

Разработан программный продукт по сбору данных социальной сети Twitter, а так же ведется разработка программного продукта по кластерному анализу собранных данных

2. Кластеризация текстовых данных

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

Наиболее популярный 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]. Наиболее популярным дедуктивным подходом можно считать Латентное размещение Дирихле (LDA).

Для более детального анализа лучше всего сочетать различные подходы и методы в зависимости от количества обрабатываемых данных.

3. Сбор данных социальной сети Twitter

Для исследования работы алгоритма TF-IDF было разработано программное средство, позволяющее собирать данные напрямую с серверов Twitter. Реализация построена на открытом интерфейсе Twitter API 2.0. Объектом исследования были взяты сообщения из твиттера (твитты) Самарской и Московской областей. Основным критерием отбора сообщений было наличие определенной геолокации (включая все населенные пункты области).

Для осуществления сбора к серверу сети Twitter отправляется запрос, содержащий в себе consumer key и consumer secret key. В ответ на запрос были получены oauth.accessToken и oauth.accessTokenSecret, которые позволили получать данные с серверов социальной сети.

Вторым шагом к осуществлению сбора данных является отправка query-запроса, в ответ на который возвращается набор твиттов.

4. Использование алгоритмов кластеризации и полученные результаты

Данные для анализа и последующей кластеризации собирались в течение шестидесяти дней по двум query-запросам: Самарская и Московская области. Было собрано 1,5 Гб информации (>600000 сообщений). Затем на этой информации были применены следующие алгоритмы: модифицированным TF-IDF, LDA, алгоритм классификации данных методом судьи.

4.1. Работа модифицированного алгоритма TF-IDF

Путем применения модифицированной метрики TF-IDF:

$$tfidf(t,d,D) = k \cdot \mathrm{tf}(t,d) \cdot idf(t,D), \quad (1)$$

где $tf(t,d) = n\_i/(\sum k * n\_k)$, $idf(t,D) = \log |D|/|(d\_i \ni t\_i)|$, a k – поправочный коэффициент, на слова, являющиеся хештегами;

и алгоритма k-means было получено 22 кластера. На примере одного из полученных кластеров (рисунок 1) видно, что сообщения близки по значению, но среди них есть сообщения с «чужой» тематикой.

Подобный не совсем точный результат вероятнее всего был получен в связи с тем, что исследуемые сообщения в твиттере имеют ограничение на 140 символов.

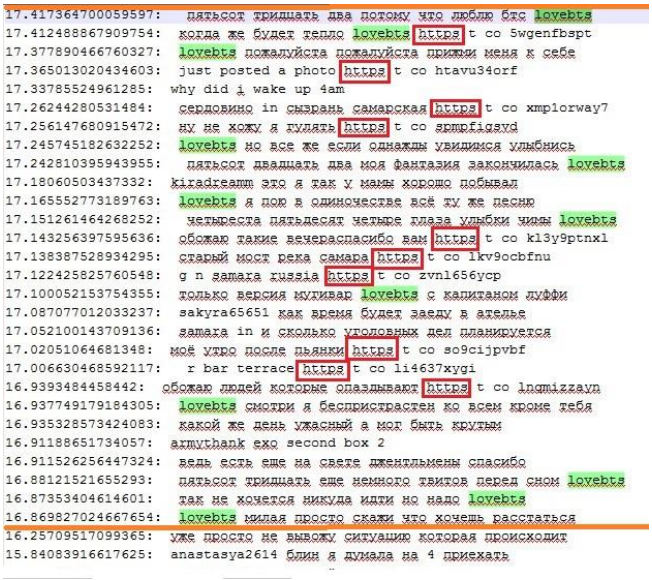

Рис. 1. Пример одного из полученных кластеров.

Кроме того, высокая плотность кластеров (рисунок 2) говорит низкой точности метрики.

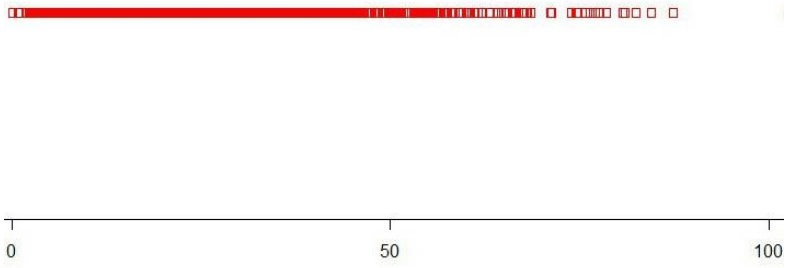

Рис. 2. Распределение значений метрики TF-IDF обработанных данных на числовой прямой.

4.2. Использование LDA алгоритма

LDA алгоритм основан на определении наиболее употребляемых топиках (темах), которые могут образовывать кластеры.

Модель LDA решает классическую задачу анализа текстов: создать вероятностную модель большой коллекции текстов (например, для information retrieval или классификации).

Очевидно, что у одного документа может быть несколько тем. Подходы, которые кластеризуют документы по темам, никак этого не учитывают. LDA — это иерархическая байесовская модель, состоящая из двух уровней:

- на первом уровне – смесь, компоненты которой соответствуют «темам»;
- на втором уровне – мультиномиальная переменная с априорным распределением Дирихле, которое задаёт «распределение тем» в документе.

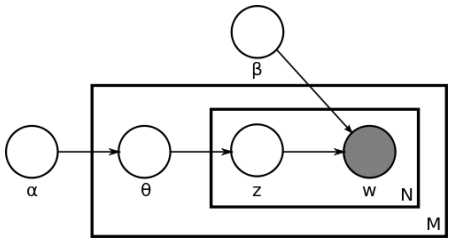

Рис. 3. Граф модели.

Рассмотрим работу модели, для этого необходимо:

- выбрать длину документа N;
- выбрать вектор $\theta \sim (\alpha)$ — вектор «степени выраженности» каждой темы в этом документе;
- для каждого из N слов w:
  - выбрать тему z_n по распределению $Mult(\theta)$;
  - выбрать слово $w_n \sim p(w_n \mid z_n, \beta)$ с вероятностями, заданными в β.

Для простоты мы фиксируем число тем k и будем считать, что β – это просто набор параметров $\beta_{ij} = p(w^j = 1 \mid z^i = 1)$, которые нужно оценить, и не будем беспокоиться о распределении на N. Совместное распределение тогда выглядит так:

$$p(\theta, , , N \mid \alpha, \beta) = p(N \mid \xi) p(\theta \mid \alpha) \prod_{n=1}^{N} p(z_n \mid \theta) p(w_n \mid z_n, \beta). \quad (2)$$

В отличие от обычной кластеризации с априорным распределением Дирихле или обычным распределением Байеса, мы не выбираем кластер один раз, а затем «накидываем» слова из этого кластера, а для каждого слова сначала выбираем по распределению θ тему, а уже потом «набрасываем» это слово по данной теме. [18]

В ходе работы экспертным путем было выявлено, что наиболее оптимальное число исходных кластеров – шесть.

Пример результатов работы алгоритма приведен в таблице 1. В ней представлены вероятности принадлежности текста к каждому из 6 кластеров.

Таблица 1. Пример результатов работы алгоритма

Пример текста	Кластер №1	Кластер №2	Кластер №3	Кластер №4	Кластер №5	Кластер №6
Сила спорта. https://t.co/dqQPfanNl5	35.461	19.335	23.081	25.019	17.128	27.564
Лучший отдых только с ними, с моими самыми дорогими https://t.co/XBrFrQgzJv	13.050	15.807	7.784	3.168	12.623	6.424
Что за игры ты затеял???	33.112	33.637	10.533	4.718	13.466	20.149
I'm at ТРК «Космопорт» - @cosmoport_s in Самара, Самарская обл. https://t.co/9XTbwSdfSq	27.597	14.618	35.064	11.965	24.892	14.070
Ну совсем, как сыр в микроволновке	12.495	11.094	2.253	5.891	13.178	28.925
Почему надо мной смеётся вся трибуна ????	34.784	32.829	0.224	23.805	28.408	14.602
@ElenaVa49244597 так,я в гости собралась	19.198	7.329	35.520	2.090	19.682	2.462
ИИИИИГОООООООРЬЬЬ СКОРО 43	1.760	5.355	5.138	7.474	2.507	9.669
У самой лучшей и любимой мамочки день рождения https://t.co/0hcQC3fqwM	20.529	18.061	8.178	2.262	22.329	27.982
Сидеть на лекции с температурой 38 кайф	14.629	10.081	7.070	21.961	34.937	8.447
Холодно настолько,что аж зубы сводит	23.311	35.638	9.650	17.149	9.982	5.856
В универе даже в куртке холодно	9.131	39.143	3.376	6.635	31.993	9.763

4.3. Классификация данных методом судьи

Метод классификации данных, известный как метод судьи, основан на том, что каждое слово можно отнести к той или иной категории (классу). Тогда в результате обработки текст будет представлять собой набор «голосов» принадлежности каждого слова в тексте к тому или иному классу. Проанализировав полученный вектор мы можем принять решение к какому классу относится текст.

В настоящее время авторами осуществляется разработка алгоритма с использованием предложенного метода судьи. Результаты будут представлены позже.

5. Заключение

Кластеризация текстовых данных, представленных короткими сообщениями (140 символов) социальной сети Twitter является актуальной нетривиальной задачей в связи с колоссальным распространением социальных сетей и интернет сервисов во всем мире.

В результате её решения был создан программный комплекс, позволяющий осуществлять сбор данных социальной сети Twiiter по определенным геолокациям. С помощью программного комплекса был осуществлён сбор данных по Самарской и Московской областям.

В ходе исследования было установлено, что используя алгоритмы, основанные на применении метрики TF-IDF, сложно получить качественную кластеризацию текстовой информации, содержащейся в коротких сообщениях социальной сети Twitter. Сделан вывод о малой пригодности метрики TF-IDF для кластеризации коротких текстовых сообщений, либо о необходима существенная модификация данной метрики.

Алгоритмы, основанные на «машинном обучении», в свою очередь, показали хорошие результаты - было выявлено шесть кластеров сообщений: «учеба», «эмоции», «обмен фотографиями», «городская среда», «новости города», «политика». Это говорит об «омоложении» аудитории социальной сети.

Алгоритм классификации данных с использованием метода судьи (в настоящий момент) находится в разработке.

В ходе дальнейшей работы планируется сравнение реализованного алгоритма классификации текстовых данных и алгоритма LDA, а также оптимизация параллельных алгоритмов кластеризации.

Литература

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

[2] Vossen, G. Big data as the new enabler in business and other intelligence //Vietnam Journal of Computer Science. 2014. Т. 1. №. 1. Pp. 3-14.

[3] Tamhane, D.S., Sayyad, S.N. Big Data Analysis Using Hace Theorem //International Journal of Advanced Research in Computer Engineering & Technology (IJARCET) Volume. 2015. Т. 4.

[4] Tan, W., Blake, M. B., Saleh, I., Dustdar, S. Social-network-sourced big data analytics //IEEE Internet Computing. 2013. №. 5. Pp. 62-69.

[5] Васильков, А. Как «большие данные» помогают улучшить безопасность [Электронный ресурс] – Режим доступа: http://www.computerra.ru/108760/security-n-big-data/ (24.09.2015).

[6] Чубукова, И. Задачи Data Mining. Классификация и кластеризация. -ИНТУИТ.ру.

[7] Blagov, A., Rytcarev, I., Strelko,v K., Khotilin, M. Big Data Instruments for Social Media Analysis //Proceedings of the 5th International Workshop on Computer Science and Engineering, 2015. Pp. 179-184.

[8] TF-IDF [Электронный русурс] – Режим доступа: https://ru.wikipedia.org/wiki/TF-IDF (20.09.2015).

[9] Wang, H. Introduction to Word2vec and its application to find predominant word senses. – 2014.

[10] Yu, M., Dredze, M. Improving lexical embeddings with semantic knowledge //Association for Computational Linguistics (ACL). – 2014. – С. 545-550.

[11] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient Estimation of Word Representations in Vector Space. In Proceedings of Workshop at ICLR, 2013.

[12] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. Distributed Representations of Words and Phrases and their Compositionality. In Proceedings of NIPS, 2013.

[13] MacQueen, J.. Some Methods for Classification and Analysis of Multivariate Observations. In Proc. of the 5th Berkeley Symposium on Mathmatical Statistics and Probability, pages 281–297, 1967.

[14] Blei, D.M., Ng, A.Y., Jordan, M. I. Latent dirichlet allocation //the Journal of machine Learning research. – 2003. – Т. 3. – С. 993-1022.

[15] Меньшиков, И.Л. Анализ тональности текста на русском языке при помощи графовых моделей / И.Л. Меньшиков // Доклады Всероссийской научной конференции АИСТ’2013: сб. ст. – Екатеринбург, 2013. – С. 151-155.

[16] Обзор алгоритмов кластеризации данных [Электронный русурс] – Режим доступа: https://habrahabr.ru/post/101338/ (10.11.2016).

[17] Рекомендательные системы: LDA [Электронный русурс] – Режим доступа: https://habrahabr.ru/company/surfingbird/blog/150607/ (23.11.2016).