Development and research of algorithms for clustering data of super-large volume

I.A. Rytsarev[1], A.V. Blagov[1], M.I. Khotilin[1]

[1]Samara National Research University, 34 Moskovskoe Shosse, 443086, Samara, Russia

Abstract

The work is devoted to the research of text data clustering algorithms. As the object of research, the social network Twitter was selected. At the same time, text data was collected, processed and analyzed. To solve the problem of obtaining the necessary information, studies in the field of optimizing the data collection of the social network Twitter were carried out. A software tool that provides the collection of necessary data from specified geolocation has been developed. The existing algorithms for clustering data, mainly of large volume were explored.

Keywords: data clustering algorithms; superhigh volume data; text analysis; k-means; tf-idf metric; lda; collective decision-making method

1. Introduction

The aim of the paper is to explore the algorithms of clustering text data of social networks collected on certain geolocations.

As the object of research data from the social network Twitter was used. To achieve the goal, the following tasks were set:

- collection of social network data,
- processing of the received data with extraction of the necessary information,
- research, approbation and modernization of data clustering algorithms.

During the research work the following algorithms were studied and tested:

- The k-means algorithm,
- LDA algorithm;
- algorithm of data classification by the judge method.

In addition to the algorithms, the following measures were tested:

- TF-IDF,
- Word2Vec.

A software product to collect data from the social network Twitter was developed. A software product for cluster analysis of collected data is also being developed.

2. Text data clustering

Clustering (or cluster analysis) is the task of dividing a set of objects into groups, called clusters [1]. Within each group there should be "similar" objects, and the objects of different groups should be as different as possible. At the same time, some measure must be defined. Unlike the classification for clustering, the list of groups is not clearly defined and is determined during the operation of the algorithm. The main goal of clustering is the search for existing structures [2-6].

The most popular approach to solving the classification problem is the classification of information through machine learning.

Machine learning is the process by which a machine (computer) is able to display behavior that has not been explicitly programmed into it. There are two types of training: inductive and deductive.

In the works of researchers engaged in cluster analysis of textual information in various types of search engines, there is often an inductive measure of Word2vec [7-8]. The most popular deductive approach can be considered Dirichlet's Latent Placement (LDA).

For a more detailed analysis, it is best to combine different approaches and methods depending on the amount of processed data.

3. Data collection from social network Twitter

To investigate the operation of the TF-IDF algorithm, a software tool that allows data to be collected directly from Twitter servers was developed. The implementation is built on the open interface Twitter API 2.0. The object of the study was a message from a twitter (tweet) of the Samara and Moscow regions. The main criterion for the selection of messages was the presence of a certain geolocation (including all settlements of the region).

To perform the collection, a request to the Twitter server containing the consumer key and the consumer secret key is sent. In response to the request, oauth.accessToken and oauth.accessTokenSecret were obtained, which allowed receiving data from the servers of the social network.

The second step in the implementation of data collection is the sending of a query, in response to which a set of tweets is returned.

4. Results and Discussion

Data for analysis and subsequent clustering were collected within 24 hours, according to two query-requests: Samara and Moscow regions. 1.5 GB of information was collected (> 40,000 messages). After that, the following algorithms were applied to this information: modified TF-IDF, LDA [9-10], data classification algorithm with the help of graphs.

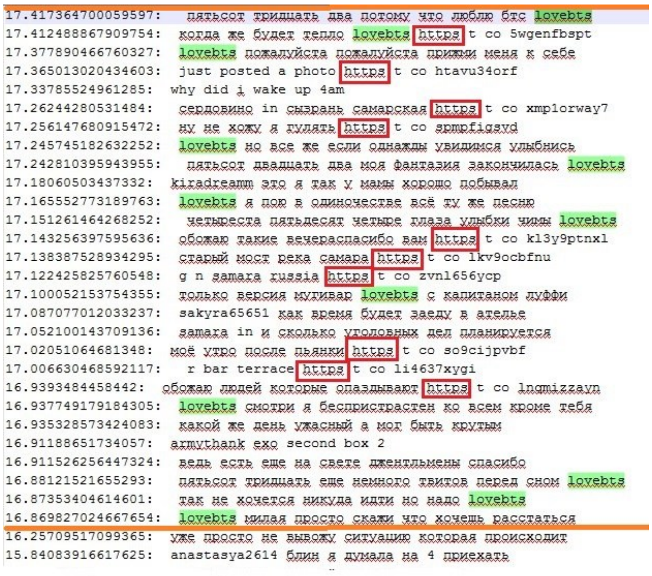
4.1. Processing with the modified TF-IDF algorithm

By applying the modified TF-IDF metric:

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

where $tf(t,d) = n\_i/(\sum k * n\_k\ )$, $idf(t,D) = \log |D|/|(d\_i \ni t\_i\ )|$, k – correction factor, for words that are hashtags; and the k-means algorithm, 22 clusters were obtained. On the example of one of the obtained clusters (figure 1) it is clear that the messages are close in meaning, but among them there are messages with "foreign" subjects.

Such an inaccurate result was most likely obtained due to the fact that the researched messages on Twitter have a 140 character limit.



```
17.417364700059597:   пятьсот тридцать два потому что люблю бтс lovebts
17.412488867909754:  когда же будет тепло lovebts https t co 5wgenfbspt
17.377890466760327:  lovebts пожалуйста пожалуйста прими меня к себе
17.365013020434603:  just posted a photo https t co htavu34orf
17.33785524961285:  why did i wake up 4am
17.26244280531484:   сердовино in сызрань самарская https t co xmp1orway7
17.256147680915472:  ну не хожу я гулять https t co spmpfigsvd
17.245745182632252:  lovebts но все же если однажды увидимся улыбнись
17.242810395943955:   пятьсот двадцать два моя фантазия закончилась lovebts
17.18060503437332:  kiradreamm это я так у мамы хорошо побывал
17.165552773189763:  lovebts я пою в одиночестве всё ту же песню
17.151261464268252:   четыреста пятьдесят четыре глаза улыбки чимы lovebts
17.143256397595636:  обожаю такие вечераспасибо вам https t co kl3y9ptnxl
17.138387528934295:  старый мост река самара https t co 1kv9ocbfnu
17.122425825760548:  g n samara russia https t co zvnl656ycp
17.100052153754355:  только версия мугивар lovebts с капитаном луффи
17.087077012033237:  sakyra65651 как время будет заеду в ателье
17.052100143709136:  samara in и сколько уголовных дел планируется
17.02051064681348:  моё утро после пьянки https t co so9cijpvbf
17.006630468592117:   r bar terrace https t co li4637xygi
16.9393484458442:  обожаю людей которые опаздывают https t co lnqmizzayn
16.937749179184305:  lovebts смотри я беспристрастен ко всем кроме тебя
16.935328573424083:  какой же день ужасный а мог быть крутым
16.91188651734057:  armythank exo second box 2
16.911526256447324:  ведь есть еще на свете джентльмены спасибо
16.88121521655293:   пятьсот тридцать еще немного твитов перед сном lovebts
16.87353404614601:   так не хочется никуда идти но надо lovebts
16.869827024667654:  lovebts милая просто скажи что хочешь расстаться
16.25709517099365:  уже просто не вывожу ситуацию которая происходит
15.84083916617625:  anastasya2614 блин я думала на 4 приехать
```

Fig. 1. Example of one of the obtained clusters.

In addition, the high density of clusters (figure 2) indicates a low accuracy of the metric.

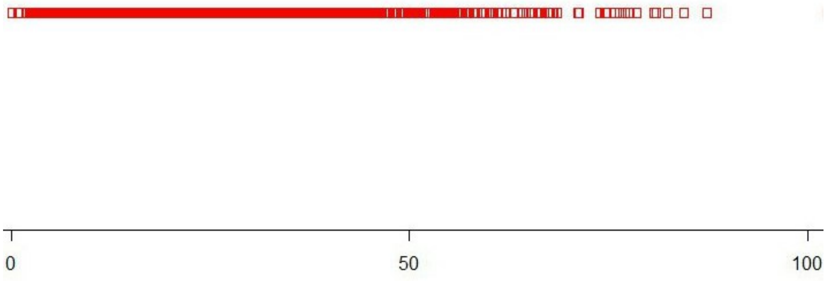


Fig. 2. Distribution of the values of the TF-IDF metric of the processed data on the number line.

4.2. LDA algorithm processing

LDA algorithm is based on the definition of the most used topics (themes) that can form clusters.

The LDA model solves the classical problem of text analysis: create a probabilistic model of a large collection of texts (for example, for information retrieval or classification).

- Obviously, one document can have several topics; Approaches that cluster documents on topics do not take this into account. LDA is a hierarchical Bayesian model consisting of two levels:
- • on the first level - a mixture, the components of which correspond to "themes";
- • at the second level, a multinomial variable with a priori Dirichlet distribution, which specifies the "distribution of topics" in the document.

Complex models are often the easiest to understand so - let's see how the model will generate a new document:

- choose the length of the document N (this is not drawn on the graph - it's not that part of the model);
- select a vector $\theta \sim (\alpha)$ — the vector of the "degree of expression" of each topic in this document;

- for each of the N words w:
  - select a topic z_n by distribution ; $\mathrm{Mult}(\theta)$
  - Select a word $w_n \sim p(w_n \mid z_n, \beta)$ with probabilities given in β.

For simplicity, we fix the number of topics k and assume that β is simply a set of parameters $\beta_{ij} = p(w^j = 1 \mid z^i = 1)$, Which need to be evaluated, and we will not worry about the distribution on N. The joint distribution then looks like this:

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

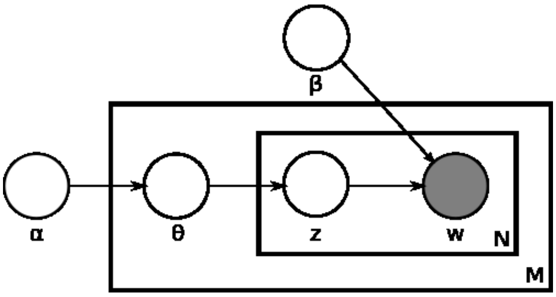
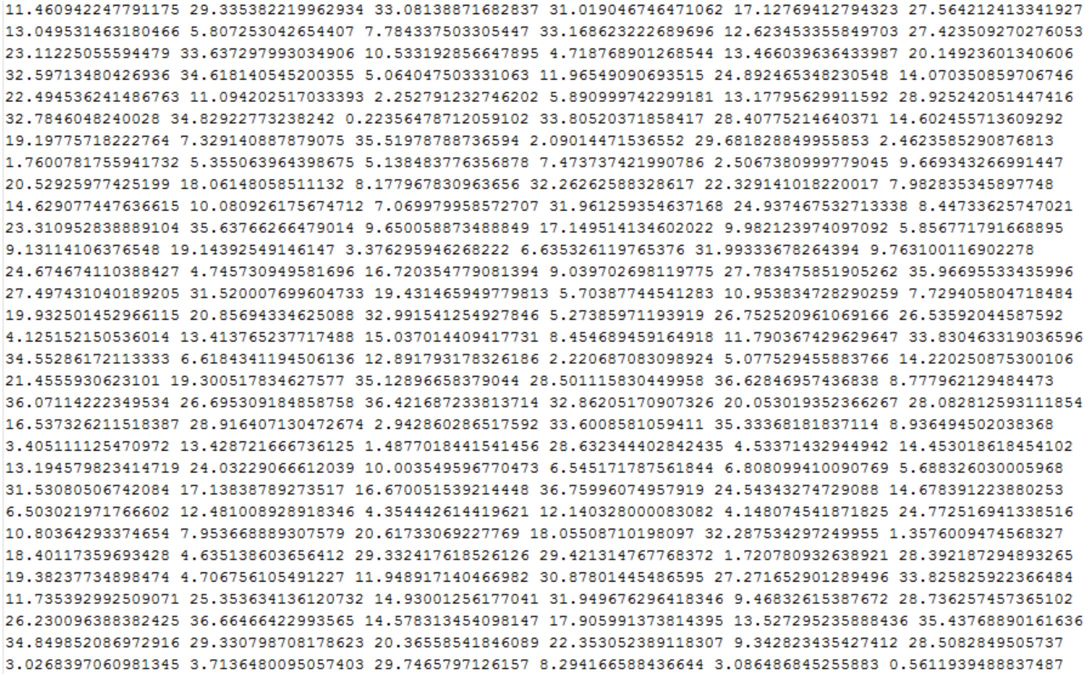


Fig 3. Graph of the model.

Unlike the usual clustering with the a priori Dirichlet distribution or the usual naive Bayesian, we do not select the cluster once, and then we insert words from this cluster, and for each word we first select the topic by the distribution of θ, and then we sketch this word on this topic [11].

In the course of the work, it was revealed by expert means that the optimal number of initial clusters is six.

The result of the algorithm can be seen in figure 4.



```
11.460942247791175 29.335382219962934 33.08138871682837 31.019046746471062 17.12769412794323 27.564212413341927
13.049531463180466 5.807253042654407 7.784337503305447 33.168623222689696 12.623453355849703 27.423509270276053
23.11225055594479 33.637297993034906 10.533192856647895 4.718768901268544 13.466039636433987 20.14923601340606
32.59713480426936 34.618140545200355 5.064047503331063 11.96549090693515 24.892465348230548 14.070350859706746
22.494536241486763 11.094202517033393 2.252791232746202 5.890999742299181 13.17795629911592 28.925242051447416
32.7846048240028 34.82922773238242 0.22356478712059102 33.80520371858417 28.40775214640371 14.602455713609292
19.19775718222764 7.329140887879075 35.51978788736594 2.09014471536552 29.681828849955853 2.4623585290876813
1.7600781755941732 5.355063964398675 5.138483776356878 7.473737421990786 2.5067380999779045 9.669343266991447
20.52925977425199 18.06148058511132 8.177967830963656 32.26262588328617 22.329141018220017 7.982835345897748
14.629077447636615 10.080926175674712 7.069979958572707 31.961259354637168 24.937467532713338 8.44733625747021
23.310952838889104 35.63766266479014 9.650058873488849 17.149514134602022 9.982123974097092 5.856771791668895
9.13114106376548 19.14392549146147 3.376295946268222 6.635326119765376 31.99333678264394 9.76310011690227B
24.674674110388427 4.745730949581696 16.720354779081394 9.039702698119775 27.783475851905262 35.96695533435996
27.497431040189205 31.520007699604733 19.431465949779813 5.70387744541283 10.953834728290259 7.729405804718484
19.932501452966115 20.85694334625088 32.991541254927846 5.27385971193919 26.752520961069166 26.53592044587592
4.125152150536014 13.413765237717488 15.037014409417731 8.454689459164918 11.790367429629647 33.830463319036596
34.5528617211333 6.618434119450613б 12.891793178326186 2.220687083098924 5.077529455883766 14.220250875300106
21.4555930623101 19.300517834627577 35.12896658379044 28.501115830449958 36.62846957436838 8.777962129484473
36.07114222349534 26.695309184858758 36.421687233813714 32.86205170907326 20.053019352366267 28.082812593111854
16.537326211518387 28.916407130472674 2.942860286517592 33.6008581059411 35.33368181837114 8.936494502038368
3.405111125470972 13.428721666736125 1.4877018441541456 28.632344402842435 4.53371432944942 14.453018618454102
13.194579823414719 24.03229066612039 10.003549596770473 6.545171787561844 6.808099410090769 5.688326030005968
31.53080506742084 17.13838789273517 16.670051539214448 36.75996074957919 24.5434327472908B 14.678391223880253
6.503021971766602 12.481008928918346 4.354442614419621 12.140328000083082 4.148074541871825 24.772516941338516
10.80364293374654 7.953668889307579 20.61733069227769 18.05508710198097 32.287534297249955 1.3576009474568327
18.40117359693428 4.635138603656412 29.332417618526126 29.421314767768372 1.720780932638921 28.392187294893265
19.38237734898474 4.706756105491227 11.948917140466982 30.87801445486595 27.271652901289496 33.825825922366484
11.735392992509071 25.353634136120732 14.93001256177041 31.949676296418346 9.46832615387672 28.736257457365102
26.230096388382425 36.66466422993565 14.578313454098147 17.905991373814395 13.527295235888436 35.43768890161636
34.849852086972916 29.330798708178623 20.36558541846089 22.353052389118307 9.342823435427412 28.5082849505737
3.0268397060981345 3.7136480095057403 29.7465797126157 8.294166588436644 3.086486845255883 0.5611939488837487
```

Fig. 4. The result of the algorithm.

Figure 4 shows the probabilities of the text belonging to each of the 6 clusters.

4.3. *Algorithm of classification of data by the collective decision-making method*

The algorithm for classifying data by the collective decision-making method is based on the idea that each word relates to one or another category (class). Then, as a result of processing, the text will be a set of "voices" of the affiliation of each word in the text to one or another class. Analyzing the resulting vector, we can decide which class the text belongs to.

Currently, the algorithm is being developed. The results will be presented for comparison later.

5. Conclusion

As a result of research work, a software package that allows to collect data from the social network Twiiter for certain geolocations was written. With the help of this complex, data collection was carried out in the Samara and Moscow regions.

It was found that using algorithms based on the use of the TF-IDF metric, it is difficult to obtain a qualitative clustering of the textual information contained in short messages of the social network Twitter. From this we can conclude that the TF-IDF metric is not suitable for short text messages, or about the necessary modernization of this metric.

Algorithms based on "machine learning", in turn, demonstrated good results - six clusters of messages were identified: "study", "emotions", "photo sharing", "urban environment", "city news", "politics". This suggests "rejuvenating" the audience of the social network..

The data classification algorithm by the judge's method (currently) is under development.

Questions on clustering and further classification of text data are relevant in connection with the enormous spread of social networks and Internet services around the world.

In the course of further work, it is planned to compare the implemented algorithm for classifying text data and the LDA algorithm, as well as studying the issue in the direction of output and optimization of parallel clustering algorithms.

Acknowledgements

The work has been performed with partial financial support from the Ministry of Education and Sciences of the Russian Federation within the framework of implementation of the Program for Improving the Samara university Competitiveness among the World's Leading Research and Educational Centers for the Period of 2013-2020s.

References

[1] Dean J, Ghemawat S. MapReduce: simplified data processing on large clusters. Communications of the ACM 2008; 51(1): 107–113.

[2] Tan W, Blake MB, Saleh I, Dustdar S. Social-network-sourced big data analytics. IEEE Internet Computing 2013; 5: 62–69.

[3] Chubukova I. Tasks of Data Mining. Classification and ckusterization. URL: http://www.intuit.ru.

[4] Belim SV, Kutlunin PE. Boundary extraction in images using a clustering algorithm. Computer Optics 2015; 39(1): 119–124. DOI: 10.18287/0134-2452-2015-39-1-119-124.

[5] Protsenko VI, Kazanskiy NL, Serafimovich PG. Real-time analysis of parameters of multiple object detection systems. Computer Optics 2015; 39(4): 582–591. DOI: 10.18287/0134- 2452-2015-39-4-582-591.

[6] Protsenko VI, Serafimovich PG, Popov SB, Kazanskiy NL. Software and hardware infrastructure for data stream processing. CEUR Workshop Proceedings 2016; 1638: 782–787. DOI: 10.18287/1613-0073-2016-1638-782-787.

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

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

[9] Blei DM, Ng AY, Jordan MI. Latent dirichlet allocation. The Journal of machine Learning research 2003; 3: 993–1022.

[10] Gong S. et al. Linear Discriminant Analysis (LDA).

[11] Reference systems: LDA. Surfingbird Blog. Habrahabr. URL: https://habrahabr.ru/company/surfingbird/blog/150607/ (23.11.2016).