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

I.A. Rytsarev¹, A.V. Blagov¹, M.I. Khotilin¹

¹Samara 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.accessTokenType 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:
\[
tf\text{idf}(t, d, D) = k \times tf(t, d) \times idf(t, D),
\]
where \( tf(t, d) = \frac{n_i}{(\sum k \times n_{k})} \), \( idf(t, D) = \log \frac{|D|}{|d_i \ni t_i|} \), \( k \) – 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.

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; $\text{Mult} (\theta)$
- Select a word $w_{n \alpha} \sim p (w_{n \alpha} \mid z_n, \beta)$ with probabilities given in $\beta$.

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

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

(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 $\theta$,

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.

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

4.3. Algorithm of classification 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 Twitter 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