

Research on Data Mining Algorithm Based on Micro-blog of Multi-view Clustering Model

Zhou Bei

Sichuan Agricultural University, Ya'an 625014, China

Keywords: micro blog; text data mining; eigenvectors; similarity matrix

Abstract: The traditional information extraction is mainly for the news, the extraction operation is relatively simple, and the characteristic of micro blog post - short refining needs to select a new and trial way to conduct the event extraction. In the course of these questions, this is a new algorithm that can think about the key word and time, set up the LDA model of the micro blog event, which is the first thing to set up the terms of the event, and calculate the semantics of the key words by the method of the DTW algorithm, the corresponding similarity matrix, and then use the method of co-training to do the same thing as the final eigenvector and finish the selection of the event. It is shown that that algorithm proposed in this paper has high accuracy and high efficiency compared with the previous algorithm.

1. Introduction

With the development of information technology and network technology, the Internet is now inseparable from our life. Its footprint has been involved in all walks of life, according to was the number of institutions, now the data rapid growth of the Internet, if in this huge database using traditional statistical analysis techniques to deal with, it will no doubt is a nightmare. Users are faced with "information explosion" and "lack of knowledge" at the same time. At this time, the emergence of search engine has greatly reduced the interference of useless information to query. The origins of search engines can be traced back to the year when students at the university of Montreal made a program to automatically index anonymous website files on the Internet. Its emergence has laid the foundation for the invention of the advanced search engine. Years, the first in the world a real modern search engines, Carnegie Mellon university's Michael Martin bo will "spider" program linked to his index, created the now familiar, it is mainly used for the university at the time of the digital library project. Then came the search engine, which began with a series of search services on behalf of yahoo!, which entered a period of rapid development. Nowadays, there are many Internet search engines, and the retrieval model is more and more extensive.

Data mining refers to the nontrivial process of revealing hidden, previously unknown and potentially valuable information from a large number of databases. It lies in the knowledge discovery in database (hereinafter referred to as knowledge discovery, its mostly in database, data warehouse and massive data set as a data source (as shown in the chart). Data mining are involved and many fields of study, including the database system, spatial database, machine learning, knowledge acquisition, statistics, based on the knowledge system, artificial intelligence and visualization, etc. It also has important practical utility in the field of medicine, business, recognition, teaching and so on, and it has further extended to other fields.

Swelled, however, as the amount of data on the Internet, the renewal speed up, the storage form of complex, traditional data mining techniques have apparently not suitable for the requirement of web mining, this will force the researchers must constantly innovate.

At the same time, the in-depth study of data mining has led to the development of data mining. In general, web pages simply list useful information to users, but because of the explosive growth of web content, it becomes increasingly difficult to find the information users need. It puts forward higher requirements on function and structure design, in fact, most of the current data in the form of structured link, so the page structure mining is one of the important research direction of data mining research.

Keyword clustering method can obtain the event cluster [1] by obtaining the critical period and corresponding word vector in the text, and then clustering the keywords and word vectors. Compared with ordinary text, micro-blog post has short and concise features, so the keywords contained in it can express the event content more effectively. Since the ninety s, the Internet developing at an amazing speed, the Web as an information manufacturing, distribution, processing and dealing with the main platform, a huge, heterogeneous, dynamic, semi-structured or unstructured information resources, and there are more than 80% of the information in these Web information in the form of Web text, the rapid growth of the capacity increase 1 million pages a day on average. With the expansion of the Internet and the emergence of a large number of online texts, it will mark this huge unstructured or semi-structured data ocean, with extremely rich useful information known as knowledge. How to find useful information and knowledge patterns in the vast amount of information provided by the Web has always been a problem for people to explore. Search engine classification browsing, retrieval effect is good, can help users find the information you need, but need artificial maintenance, the maintenance cost is high, information update slow, maintenance workload is big, at the same time, the search results accuracy is not high, the recall rate is limited, more can't satisfy the requirement of the user is given a special personalized service. Therefore, knowledge discovery based on Web text data comes into being. Web text categorization can effectively solve the above problems, and can be based on text semantic Web page contains a lot of Web page will be automatically classify, help people to better grasp the Web information, help the user to rapid and accurate positioning of target knowledge, to reduce the search space, to speed up the retrieval speed, improve the accuracy of the query. In recent years, the event extraction method based on keyword clustering has been increasingly favored by microblog event fetching tasks [2]. Paper [3] by extracting micro blog named entities in words, and these entities as keyword nodes, as well as on the dynamic relationship between the entity relationship diagram, the word clustering figure segmentation method is used to solve the problem, by segmentation to the same subgraph entities in the keyword are extracted as the description of an event. MABED [4] more consideration to the social relationship between the user, the article think certain social groups will be the one or a few incidents widely discussed, and through the @ operation to establish the linkage between twitter accounts or discussion. So in this paper, through the detection of weibo @ in operation, the preliminary screening alternative events, the algorithm and model will be the highest probability, influence the largest number of keyword as a representative of the extraction of events. But MABED exist some shortcomings, that is, there are many do not include the @ operation in weibo event cannot be extracted, which makes the algorithm of extracting accuracy affected by certain, practicality is not strong, in addition, MABED nor fully considering the timing relationships affect event extraction. DECoW [5] give full consideration to the role of temporal information, through the will of time sequence, the keyword appears to calculate fluctuation entropy so as to select the final event, but the algorithm only considers the temporal information, lead to the same time more than one event could not be extracted accurately[8-9].

Traditional information extraction work is primarily for news, news text and micro blog this compared with a detailed[10], complete, long length, etc, extraction work is relatively simple, and the characteristics of the micro blog this short refining, need to choose a new, trial event with the method of extraction. Event occurrence time two elements can use keywords and description, before a lot of work is aimed at the two aspects of algorithm and model design, but none fully use both information, therefore, in this paper, on the basis of work before, put forward a kind of can simultaneously consider keywords and occurrence time of the new algorithm, the algorithm is the core process is as follows: 1. After data preprocessing, weibo event collection of LDA model [6], generates keywords set as a marker of the description of the event, through the DTW algorithm [7] of event semantics between keywords, sequence similarity calculation, get the corresponding similarity matrix, finally USES the clustering method, the joint training where the iteration to produce the final feature vector and complete the event selection.

2. Multi-perspective Word Clustering Model Framework.

Can be based on the following three reasonable hypothesis to build the proposed event extraction method based on word clustering multiple points of view: a, an event can be expressed in a number of key words, such as happened on November 13, 2015, terrorist attacks can use "the French capital Paris," bomb "and other words to describe. 2, can be used in a certain algorithm themes and text extracted from weibo event related keywords, such as "terrorists" and "bomb" word is generally used to describe events such as terrorist attacks. Three, if a few key words often appear together, and the temporal development trend and so on, then these words are more likely to describe the same event. The latter two assumptions are the most important ones in the current popular time extraction methods and models. The core idea is that the topic text information and timing information in microblog have a significant impact on the extraction of events.

Based on the above assumptions, this paper proposes the method of event extraction of multi-angle word clustering model and USES the theme and temporal information. This model is mainly composed of three parts: first, the micro-blog post is processed, and the main processes include tagging, marking time, filtering low-frequency words and so on. Then, the model of subject extracting several key subject as a candidate of key events, then the keywords and related degree between each pair of key words to calculate, to build corresponding to the semantic similarity matrix and the timing of all key similarity matrix. Finally, a clustering result can be obtained by putting them into a multi-perspective pga algorithm, which contains the most accurate set of keywords that can describe an event.

2.1 Topic model

Subject the LDA model (Latent Dirichlet Allocation) is a classical probability model of the theme. LDA think each article does not belong to a certain theme, but belong to the collection of a set of have more than one topic, at the same time, the word is in the theme of each article must, on the basis of appear at a certain probability. Firstly, this paper constructs the micro-blog LDA model and USES each microblog as a mixed distribution of a group of events, and the occurrence probability of the word in microblog is a polynomial of the event. LDA can obtain the high probability word for each event, which is the characteristic word for this event.

The extraction method of the subject words refers to the total number of topics in K, and the specific state of the words under the KTH topic is recorded as the numerical value of the gibbs sampling method, which is acquired by the subject model LDA. For a topic k, the m word with the highest value is determined as the main inscription. At most, km will be a subject feature in the premise of allowing repetition. The subject words chosen by the above method are of great significance to the accurate description of the event and will also be used to describe the event.

Word similarity calculation using LDA model to extract on weibo event, usually put each topic is expressed as a theme phrases, these keywords can be used to describe an event, but a subject contains several event belong to the same subject of possibility is very great. For example, state visits without state leadership are a subject, but they belong to several events. Therefore, the results obtained above do not have practical value. In order to enhance the availability of the results, this paper introduced a new word clustering method, through the theme of the word distribution to clustering, clustering of key in the process of similarity is computed by Cosine algorithm.

After getting the distribution matrix of the word in the article about the topic, the key words can be used to calculate the similarity of the keywords by using the principal vector and the cosine similarity function of formula 1.

$$\text{Sim}_s(x,y) = \text{sim_cos}(\varphi_x, \varphi_y) \quad (1)$$

2.2 Timing analysis

EDCoW adopts micro-blog timing information for time extraction. It describes the change of the word in the data by constructing the trend curve of the occurrence frequency of each keyword in the micro-blog. The core assumption of EDCoW is that the keyword describing the same frequency

development trend is more likely to describe the same event. Based on the above theoretical method, the new calculation formula is defined to calculate the similarity.

Keywords timing signal similarity calculation This article calculate the tendency of signal of word method is as follows: first to extract weibo data collection C for segmenting the timestamp, assuming that the data C data collection for T day, every day as an event node, the word w signal can be classified as the following sequence:

$$S_w = [s_w(1), s_w(2), \dots s_w(T)] \quad (2)$$

Said t time keyword frequency, calculation method of reference EDCoW, as shown in formula (3), said at the time t contains words w the number of weibo, total number of weibo said corresponding period.

$$s_w(t) = \frac{N_w(t)}{N(t)} \times \log \frac{\sum_{i=1}^T N(i)}{\sum_{i=1}^T N_w(i)} \quad (3)$$

Dynamic Time Warping (DTW) is used to calculate the similarity of two sequential signals in order to improve the computational efficiency and accuracy of Time order similarity. DTW algorithm is a classical dynamic programming algorithm, which can be used to calculate matching similarity between sequences, DTW algorithm segmentation will be a big problem into several small problems, based on calculation of each small calculation result has been obtained, such as formula (4). Where, I,j is the node number of the sequence.

$$D(i, j) = d(i, j) + \min \begin{cases} D(i-1, j) \\ D(i, j-1) \\ D(i-1, j-1) \end{cases} \quad (4)$$

If m,n is the length of the sequence, the distance calculation formula of the final two sequences is shown as follows:

$$DTW(S_x, S_y) = D(m, n) \quad (5)$$

2.3 Multi-view clustering model.

Although the methods of the above two chapters can complete the extraction, there are problems: only partial data information is used. In order to make full use of the effective information of weibo, this paper presents a multi-angle extraction algorithm to solve the problem. The algorithm is suitable for the sample size to be decomposed directly, and the views of the samples are not affected by each other. The multi-angle clustering model algorithm flow is as follows:

Keywords similarity matrix constructed first, do not need to complete the clustering and multiple points of view for all word similarity matrix constructed, only need to select with rich meaning of N is the most important word structure and the cluster, so that on the one hand able to more accurately describe the events, also can relieve algorithm of time and a lot of overhead.

Through formula (1), we can calculate the semantic similarity between keywords, calculate the complete set of keywords separately, and construct the semantic similarity matrix G between different keywords. The G matrix is the semi-positive definite matrix of NxN, the matrix item $G_{i,j}$ represents the semantic similarity of keyword w_{ij} . In combination with the key words of the previous chapter, the time sequence distance calculation method is adopted, and the distance formula of the sample is converted into the similarity formula, and the time series similarity calculation formula of w_{ij} is obtained:

$$Sim_{s(x,y)} = \frac{1}{1+DTW(S_i, S_j)} \quad (6)$$

According to the semantic similarity matrix G, we can define the time order similarity matrix T, the matrix item T_i , and j for the time order similarity of the keyword w_{ij} .

The multi-angle clustering algorithm adopts the principle of cooperative training, and the semantic similarity matrix G and the time order similarity matrix T are used as input to calculate the

clustering results. Is the core of the method through an Angle of view of the similarity matrix constantly affect another perspective of clustering results, namely the semantic similarity matrix G and sequential similarity matrix T correct each other, after several rounds of iteration, is a combination of the two perspectives of the optimization of clustering results.

The process of the above algorithm is: in the $i-1$ training process, the matrix of these two different perspectives can be calculated by the p-clustering algorithm, and the eigenvector $ui-11$ and $ui-12$ can be obtained. In the next iteration, $ui-11$ and $ui-12$ are brought into the algorithm, and the clustering results can be obtained respectively from this perspective. By combining the initial matrix G and T , the new semantic similarity matrix $S1$ is recomputed, and $S1$ is corrected by $ui-12$, and the new time order similarity matrix $S2$ is also obtained. The constant iterative process, which makes interaction between two matrices and correction, the iterative process is completed, matrix has been fixed eigenvector $U_i - 11$ G , T eigenvector matrix has been fixed $U_i - 12$, this paper argues the importance of semantic information is much greater than the temporal information, so the $U_i - 11$ chosen as the final computation results, the results at the same time under the influence of semantics and sequential two perspectives, this article named the method for multiple Angle of view MVCM extraction method.

3. Experiment and Result Analysis.

This section introduces the experimental data sets and experimental methods, and compares and analyzes the experimental results.

3.1 Experiment setting

In this paper, two public data sets are selected for the experiment: FSD2011[8] and Event2012[9]. Each data in the data set are marking the event, FSD2011 as small sample data set, selecting 20 of these weibo number of normal data, experiments with Event2012 as large sample data set, in order to further analysis of differences in the number of different ways in different samples, the data sets were randomly divided into 50, 100, 150 and 200 events packets.

3.2 Experimental results

Table 1 comparison of accuracy of different algorithms.

Dataset Method	FSD2011		
	Precision	Recall	F-score
MABED	0.9100	0.5600	0.6928
TSC	0.7100	0.6100	0.6561
LDA	0.9100	0.7100	0.7975
MVCM	0.9100	0.8100	0.8571

Dataset Method	Event2012_1			Event2012_2		
	Precision	Recall	F-score	Precision	Recall	F-score
MABED	0.8433	0.1900	0.5600	0.8059	0.3600	0.4962
TSC	0.6100	0.5700	0.6100	0.4600	0.4200	0.4391
LDA	0.7700	0.7100	0.7100	0.6700	0.6400	0.6547
MVCM	0.8700	0.7700	0.8100	0.7700	0.7200	0.7441

Dataset Method	Event2012_3			Event2012_4		
	Precision	Recall	F-score	Precision	Recall	F-score
MABED	0.7005	0.1967	0.3039	0.6529	0.1700	0.2662
TSC	0.3767	0.3100	0.3400	0.2450	0.2250	0.2346
LDA	0.5967	0.5767	0.5865	0.4700	0.4600	0.4649
MVCM	0.6433	0.6100	0.6262	0.5150	0.4950	0.5048

The core of MVCMM is based on the extraction of keyword clustering. In this paper, MVCMM is compared with three other methods based on keyword clustering.

Table 1 shows the experimental results of four algorithms for five data sets. Both TSC and LDA are based on the single-perspective pga algorithm. Their experimental results are basically the same as that of MVCMM, and LDA algorithm is not as good as the performance of MVCMM in all data sets. With events, an increase in the number of data sets, all the methods of extraction result are falling, the fall in TSC is most obvious, the main reason is the core of the TSC is to distinguish between temporal trend of different keywords, so it can't identify the similar events happened in time. In the process of increasing the data volume, the MVCMM method always maintains high accuracy.

Next MVCMM compared with extraction method based on text clustering, LSH [8] is the typical representative of such a method, its core is to only use text semantic information to calculate the text similarity, the use of local sensitive hashing (LSH) can improve the calculation efficiency. In order to compare the MVCMM, this paper adopts the method of assigning keywords clustering to the key word cluster to deal with MVCMM. In addition to the data indicators mentioned in the previous section, this section USES the Information (NMI) indicators to evaluate, and the experimental results are shown in table 2.

Table 2 NMI comparison of different algorithms.

Dataset Method	LSH				MVCMM			
	Precision	Recall	F-score	NMI	Precision	Recall	F-score	NMI
FSD2011	0.2079	0.9440	0.3365	0.7280	0.8393	0.8441	0.8417	0.9072
Event2012_1	0.1640	0.8821	0.2717	0.7227	0.7099	0.9054	0.7698	0.9009
Event2012_2	0.0579	0.8187	0.1004	0.5817	0.6477	0.8581	0.7380	0.8762
Event2012_3	0.0374	0.8735	0.0632	0.6072	0.5733	0.8685	0.6903	0.8763
Event2012_4	0.0427	0.8278	0.0728	0.5661	0.4632	0.8208	0.5914	0.8486

4. Conclusion

In this paper, a multi-perspective clustering model (MVCMM) is proposed to perform the micro-blog event extraction task, and the multi-view framework is constructed to utilize the topic information and timing information of the microblog simultaneously. Compared with other similar algorithms, MVCMM proves that MVCMM has remarkable efficiency and accuracy.

References

- [1] Yeh A S, Hirschman L, Morgan A A. Evaluation of text data mining for database curation: lessons learned from the KDD Challenge Cup[J]. Bioinformatics, 2003, 19 Suppl 1(suppl_1):i331.
- [2] Tiffin N, Kelso J F, Powell A R, et al. Integration of text- and data-mining using ontologies successfully selects disease gene candidates.[J]. Nucleic Acids Research, 2005, 33(5):1544-1552.
- [3] Yeh E S, Hirschman L, Morgan E A. Evaluation of text data mining for database curation: lessons learned from[J]. Bioinformatics, 2003, 19 Suppl 1(suppl_1):i331.
- [4] Goldman J, Chu W D, Goldman R. Term domain distribution analysis: a data mining tool for text databases.[J]. Methods of Information in Medicine, 1999, 38(2):96-101.
- [5] Karimi S, Wang C, Metke-Jimenez A, et al. Text and Data Mining Techniques in Adverse Drug Reaction Detection[J]. Acm Computing Surveys, 2015, 47(4):1-39.
- [6] Jeske D R, Liu R Y. Mining and Tracking Massive Text Data: Classification, Construction of Tracking Statistics, and Inference under Misclassification[J]. Technometrics, 2007, 49(2):116-128.
- [7] Gijóncorreas J A, Andradenavarro M A, Fontaine J F. Alkemio: association of chemicals with biomedical topics by text and data mining[J]. Nucleic Acids Research, 2014, 42(W1):422-9.

- [8] Koike S, Yamaguchi S, Ojio Y, et al. Effect of Name Change of Schizophrenia on Mass Media Between 1985 and 2013 in Japan: A Text Data Mining Analysis[J]. Schizophrenia Bulletin, 2015, 42(3):sbv159.
- [9] Kim K H, Lee S, Shim J B, et al. A text-based data mining and toxicity prediction modeling system for a clinical decision support in radiation oncology: A preliminary study[J]. Journal of the Korean Physical Society, 2017, 71(4):231-237.
- [10] Botros, Lu, Z. X, et al. CLINICAL NOTES AS INDICATORS FOR VITAMIN B12 LEVELS VIA TEXT DATA MINING.[J]. Pathology, 2014, 46:S84.