An Accurate Topic Mining Algorithm Based on Business Dictionary

Zhi Yang*, Feng Lin, Mu Hu, Qing-Qiang Meng, Hao-Quan Zheng

NARI Group Corporation /State Grid Electric Power Research Institute, Nanjing, China
* *yangzhi1@sgepri.sgcc.com.cn*

Abstract: The text mining is an important branch of data mining. Many scientific research institutions and teams are actively exploring and putting forward algorithms. Because of industry and scene difference, it is difficult to use the common analysis algorithm of log to mine the potential information accurately. For example, a topic is given in one scene, how to find the main related words is not easy. To deal with the problem, this paper provides the accurate topic mining algorithm based on business dictionary. In the algorithm, segmenting with business dictionary is achieved in the document set after screening the valid documents. In this step, the document set is split into professional terms and then the invalid words are removed. Finally, the qualitative analysis is transformed to quantitative analysis. With the relevance index, the relevance degree of every word is computed. The relevance matrix is returned to the user to analyze the relevance of the words and topic. The algorithm has been applied to PMS and the validation result shows the main related factors can be analyzed accurately.

1. Introduction

With the expansion of breadth and depth of data, the potential value of data, which can create wealth, is cognized by more and more enterprises, governments and other research communities [1]. As the key step of data knowledge discovery [2], data mining [3, 4, 5, and 6] is one of the research focuses. Text mining [7, 8] is an important branch of data mining. Nowadays, many algorithms have been proposed to solve data analysis in the text scenario. However, in electric power, there are too much related equipment, inconsistent statement, unfixed topics and many related fields in the descriptions of power operation and maintenance log. It is difficult for the existing text mining algorithms to find out its maximum value. How to achieve accurate data mining is becoming one of research hotspots.

In text mining, the text data is stored in the database or file in the form of semi-structured or unstructured data. It is difficult to dig out its value because of semantic information hidden in it. At present, there are many scientific research institutions and teams to provide algorithms. There are mainly two types. The first is text cluster [9-14]. The cluster analysis [9] is one of important method to realize text mining. The literatures [10, 11] provide two fuzzy cluster methods based on weighted characteristic. In them, the feature weight vector, which reflects the internal structure of data set, is obtained by using the supervised or unsupervised learning process. Then the distance function of weighted characteristic comes into being. Another representation method is automatic weighted characteristic technique [12, 13]. In the K-Means or FCM, the feature weight vector indicates importance of each feature on the whole data set. Besides, the literature[14]proposes the fuzzy clustering algorithm by integrating the feature weighting metric into the framework of soft subspace learning. In these algorithms, because that there are different key words in different topic and the descriptive ability of each key word for topic is different, it is difficult to find out the best feature vector and weight.

Another is topic mining [15-24]. In the PLSA (Probabilistic Latent Semantic Analysis) model [17, 18], the probability is related with special documents. So, there are some defects for dealing with new documents. In addition, over-fitting comes into being easily. The most famous topic model is LDA (Latent Dirichlet Allocation) [19, 20, and 21]. In the model, the document set is input. By setting the appropriate parameters, the final multiple topics and word-distribution in each topic could be obtained. On the basis of LDA, The extended LDA models (Twitter-LDA [22], Labeled-LDA [23], MB-LDA [24], etc.) are used for some scenarios. In these algorithms, the topic mining can be achieved and topic-related words are enumerated. However, business experts need to take time to analyze what is the topic and whether these words are related with the topic. Furthermore, the flexibility of algorithm is low and the topic-related words cannot be got for random topic.

The above algorithms can solve many text mining problems, but don't satisfy the text mining demand of electric power. In the scene of text mining, the log data is stored in the semi-structured database and in every

description of log, operation and maintenance information of power network equipment and systems (for example, equipment failure, maintenance procedure, etc.) are stored. The demand is that for given topic, the semantic-related words can be found out. At the same time, the semantic correlations of words are got. The potential value of data can be used for design and planning of power network. But, the above algorithms could not dig out the accurate result.

To address the aforementioned problems, we provide the topic mining algorithm with business dictionary to precisely find out the log value. In the algorithm, on the basis of the theory of topic mining, according to the process of document-word, by using the natural language semantic segmentation technology, the semantic impact factor of words are computed. Then, with the help of business dictionary, the accurate semantic-related words are returned to business expert. It is convenient for them to find out the potential knowledge.

The advantage of this algorithm lies in:

Search topic-related words with semantic technology. For the given topic, by use of semantic theory, conditional probability and business dictionary, the semantic-related words are found out accurately.

Precisely dig out the words with the business dictionary. In this case, the interference of large number of unrelated words is avoided. The time and energy to search the business words from all the words can be economized.

Reduce the number of documents with topic. Using the topic, the irrelevant documents can be excluded. So, the efficiency is improved obviously.

This paper is organized as follows. Section 2 states the topic mining theory and evaluation criterion. The topic mining algorithm with business dictionary is proposed and analyzed in section 3. Section 4 verifies this system solution in the PMS. The last section draws to a conclusion.

2. Topic Mining Theory and Evaluation Criterion

2.1 Topic Mining Theory

In the scene of given topic, the ultimate objective of topic mining is that the optimal semantic-related word set can be found out. Because that actually the semantic-related words often exist in the sentence with topic word, the conditional probability is used for computing the semantic-related degree of word-topic.

In the document set, it is rare that there are different words in different topic. Usually, the words belong to two or more topics shown as figure 1. Because that it is simple that different words are in different topics, in this paper, it is discussed that the words belong to more topics. When In one sentence word A and word B appear together, they are described as co-occurrence.

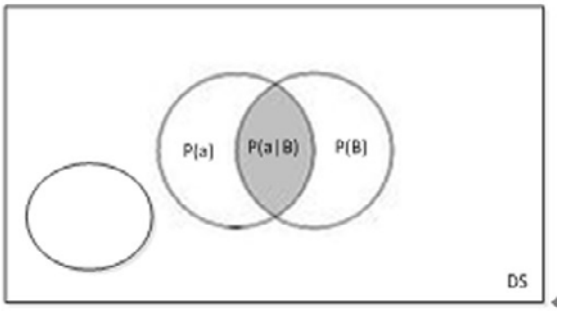

Figure 1. semantic-related probability

In the document set DS, there are many sentences S which describe one or more topics. In every sentence, there are many words W. From the above figure, we can see that there are some words a in the topic B. $P(a)$ is made up of many topic-related $\mathrm{P}(a_i)$. For the given topic B, the $P(a \mid B)$ is semantic-related degree.

The mathematical formula of semantic-related degree set is:

$$S = \{O_i\}, i = 1,2,3,... \quad (1)$$

In which, the O_i indicates the semantic-related degree of every word. Its mathematical formula is:

$$O_i = P(A_i \mid B) = {P(A_i B)}\Big/{P(B)}, i = 1,2,3,... \quad (2)$$

In which, $P(A_i B)$ indicates the co-occurrence probability of B and A_i in the document set. $P(B)$ Indicates the occurrence probability of topic B. $P(A_i B)$ indicates the occurrence probability of A_i under the premise of topic B.

The higher semantic-related degree is, the higher the O_i is. For the words with which the topic isn't together, the O_i is zero.

2.2 Evaluation Criterion

For the given topic, we can obtain many words from the above algorithm. There is different contribution of every word for the topic. The contribution of all words is 100%. For user's threshold, the fewer words, the better the effect. In this case, proportionality index is used to describe it.

$$proportionality = {numP}\Big/{numA}$$

In the formula, the $numP$ indicates word number of cumulative sum of contribution is greater the threshold. The $numA$ indicates the number of all the word number.

The proportionality is qualitative analysis index. The index cannot indicate the final result is good or not.

Combined with the actual business situation, user evaluates it.

3. Topic Mining Algorithm with Business Dictionary

The general principle of the algorithm is that the original data cannot be changed and disturbing by man-made factors should be avoided. Based on it, the topic mining algorithm is designed.

By using the topic set, the total amount of documents is reduced and topic-related documents are kept. Then, by means of business dictionary, the accurate words can be searched and saved. Finally, on the basis of relevance index, the semantic-related of word-topic can be obtained. The sorted result can be returned to user.

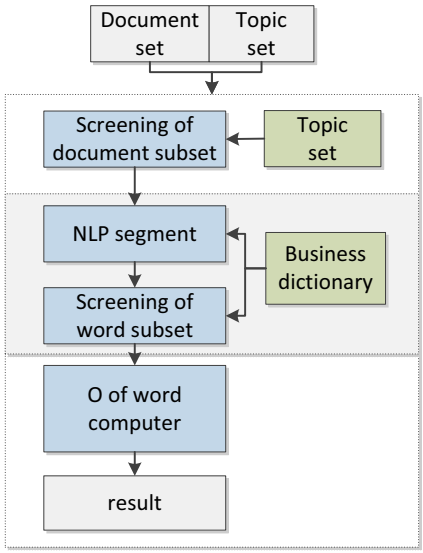

Figure 2. The process of algorithm

There are three steps in the algorithm.

The first step: Screening of document set

There are many topics in the document set. When one or more topics are analyzed, the related documents need to be filtered out. In this way, the analysis scope can be reduced to improve efficiency. The filtering method is that on the base of topic (or related topic) set, the fuzzy search can be used to traversal of all documents. When the topic is contained in the document, the document is related, otherwise it is independent. According to the topic set, the irrelevant documents are taken away.

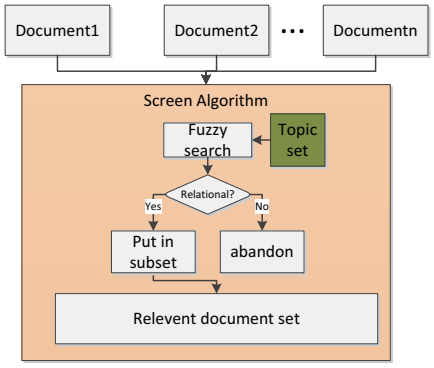

Figure 3. The process of screen algorithm

The documents of subset are related to topic. When there are none in the subset, it indicates there is no document related to topic.

The second step: Segment with business dictionary

There are different technical terms in different industries. In the same industry, there are some specific terms in different scenarios. If these terms, which are not included in the general dictionary, cannot be identified, the valid information hidden in the documents will be lost and some data value cannot be found. So, in order to avoid the loss of useful information, the defect can be remedied in the stage. The business dictionary depends on specialized person. There are industry term, special scene term and some common term. The segment process with business dictionary is that segmenting word to split document set, then, filtering out the form word, figure, and so on. The reserved word set is effective which the base for precise theme mining is.

The detailed procedure is as follows:

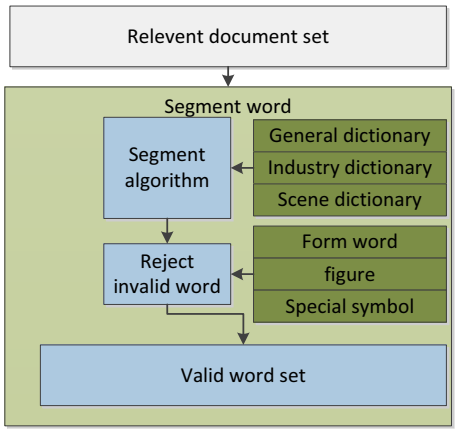

Figure 4. The process of segmenting word

The third step: relevance index computing

Because the words and topic appear in the same document, the words describe the topic from different aspect or angle. The correlation degree of different word and topic is different. The higher the correlation, more intimate the word and topic. In this case, the accurate characterization can be got from the word set. The related words reflect the customer's focus to a certain degree. In the paper, the relevance index is used to describe the relationship. The relevance index is illustrated with the ratio of word frequency and topic frequency.

The formula to compute the relevance is as follow:

$$R = WF/TF = \sum w_i / \sum t, i = 1,2,3,... \quad (3)$$

In the formula, the R indicates the relevance. The WF indicates the word frequency. The TF indicates the topic frequency.

After computing R of all words, a data matrix is formed which describes the relevance of each word and topic.

The data matrix as result is returned to the user. On the base of the data matrix, the data can be displayed with network diagram etc.

The algorithm analysis:

By appoint the topic and analyze the topic in the document set, the user can obtain the relevant words. In this case, the relationship between the topic and related words is shown accurately. At the same time, it is avoided that the minor topic is ignored.

In the algorithm, there are no all the topics but appointed topic. So, the workload can be saved to improve efficiency.

By use of business dictionary, the specific business words can be retained. It is very good to grasp the topic accurately. At the same time, the high frequency and not related words are removed. In this case, the related words are focused.

Using the relevance index formula, the relevance of topic and word as qualitative analysis index is used for the user to support making the policy decision.

4. Experiment and Analysis

The above solution has been applied to the power production management system (PMS) of State Grid of China.

The PMS is the most advanced power production management system in the world. It is one of SG186 engineering applications, which is one of the most massive and complex applications. The investment of PMS1.0 is over billion. The investment of PMS2.0 is much larger than PMS1.0. Including in the low pressure data, the total amount of data of PMS2.0 is over 150 billion. In these data, the log is very large and contains much information such as grid operation and fault information. In the log, there is a lot of implicit information which are not found. In this paper, the fault data is selected to verify the above algorithm.

In the PMS of some province, there is a lot of fault and defect data of main transformer, line, tower and other equipment in the log. The log describes the related equipment, phenomenon, process, consequence and failure analysis. From the data, we can analyze the main factor and secondary factor. The direct or indirect causes may be obtained.

In the R language [25] environment, based on the real data, the verification can be achieved. The main analysis process is as follows:

4.1 Preprocess Stage

The fault log is stored in the structured database. So, three fields are selected to analyze the algorithm. The three fields are as follow:

Table1: the description of three fields

No	Field name	Description
1	DSMCE	city name
2	JSYC	reason of failure
3	GZQFK	The description information of failure

In the selected log data, because that the null value is no influence to analysis result. So, all the null values are deleted.

In order to determine the topic set, the categorical data (JSYC) is analyzed. From the data, the "40499" value is selected as analysis object. Then, the data about it is analyzed with word cloud algorithm. The biggest words are selected as topics. The topic set is:

{“switch trip”,“phase fault”,“Lightning”}

4.2 Verification Process

First step: screening of the document set. Using the topic set, all the documents are filtered with the fuzzy algorithm. In the fuzzy algorithm, the filter condition is “like”. If the document contains the topic, the document is remaindered. The remainder is the valid data which number is 4528.

Second step: segment with business dictionary. According to the state grid and the fault scenario, the professionals can formulate the business dictionary.

In the result of segmenting, there are many invalid words. There are many form words, single letters, and figures, and so on in the invalid words which are shown as following.

{ A,C,B,Q,30,23,64,5,7,17,... }

Without the business dictionary, it is difficult to find out the main factor.

Because there are a lot of words after segmenting, some words, which frequency is higher, are selected to show word frequency chart. The top 150 words are related in every chart. From the chart, we can see that with the business dictionary, it is easy to see the main factor set which is {“switch trip”,“phase fault”,“Lightning”}.

Third step: with the above formula (3), the topic one by one is used to analyze the related words. The relevance index of every word can be computed. The user's threshold is 90%. With the help of the evaluation criterion formula, we find out the proportionality of top 10 is more than 90%. So, the top 10 is selected to analyze. For example, the result of the topic “switch trip” is shown as following.

Fail to coincide	0.37
Line protection action	0.26
Overcurrent I share protection action	0.24
Red two, 203 lines	0.24
Advanced 609 wire	0.22
Bump	0.21
crack of needle bottle	0.21
Zhoucun 634 line	0.2
Unsuccessful Reclosing	0.18
phase fault	0.18

Figure 5. The related words of topic “switch trip”

From the figure, we can see that the word frequency of "Fail to coincide" is closest to topic "switch trip". The result is returned to the user. Based on the result, the algorithm plot in the igraph package is used to show the result. The result is shown as follow.

Figure 6. Relevance chart of related words and topic

In the figure, the wider circles are topics, the smaller circles are related words. The connection line between the big circle and small circle indicates relation of related words and topic. The thickness of the connection indicates relevance degree.

The business professional can analyze real influence factors. For example, there are several lines relating the topic "switch trip". It indicates these line faults often cause the topic" switch trip".

Analysis

Based on the algorithm, the most relevant words with topic can be shown with the thick line. In this case, it is convenient to the business workers to analyze the main factors and secondary factors accurately.

Screening the document set with the topic set can reduce the amount of documents. So, the efficiency of topic analysis in small document set is improved evidently.

Segment with business dictionary can retain more potential information about documents. By the help of it, the professional may analyze the relevant words accurately.

Using the relevance index, the qualitative analysis is transformed into the quantitative analysis. On the base of this, the relevant degree can be described accurately.

5. Conclusion

Because that it is difficult to accurately analyze the related words to document set from the log data, this paper provides the accurate topic mining algorithm based on business dictionary. In the algorithm, firstly, with the topic set, the valid document set can be filtered with fuzzy algorithm. After creating the industry and scene business dictionary, the documents are segmented into a lot of words. In this case, the invalid words are removed. At last, by using the relevance index, the relevance of every word can be computed. The qualitative analysis is transformed into quantitative analysis. The algorithm is applied to PMS. The validation result shows the main related factors of topic can be analyzed accurately.

Although the algorithm can solve the problem of analyzing the log data accurately, when drawing up the business dictionary, the business expert must be need. The quality of the business dictionary is constrained by business expert ability. Furthermore, it is the research direction in the future how to improve the efficiency of the algorithm in the distributed environment.

References

[1] Jiawei Han,Micheline Kamber,Jian Pei.Data Mining Concepts and Techniques Third Edition. Ming Fan, Xiaofeng Meng.China Machine Press,2015.7

[2] Luo, Qi.Advancing knowledge discovery and data mining.Proceedings-1st International Workshop on Knowledge Discovery and Data Mining,WKDD,pp:3-5,2008

[3] Mehmed Kantardzic.Data Mining:Concepts,Models,Methods,and Algorithms, Second Edition. Xiaohai Wang,Zhigang Wu.Tsinghua University Press,2014.12

[4] Yu Wu.Accurate Data Mining For Big Data.Chemical Industry Press.2014.3

[5] Mhamdi Faouzi1,Elloumi Mourad1.A new survey on knowledge discovery and data mining.Proceedings of the 2nd International Conference on Research Challenges in Information Science,RCIS 2008, pp:427-432

[6] Tang Zhihang,Li Zhenhui,Yang Baoan.Knowledge discovery and data mining to assist natural language understanding.Journal of Computational Information Systems, v5, n1, pp:317-322, February,2009

[7] Al Fawareh, Hejab Ma'azer,Jusoh, Shaidah,Osman, Wan Rozaini Sheikh.Ambiguity in text mining.Proceedings of the International Conference on Computer and Communication Engineering 2008, ICCCE08: Global Links for Human Development, pp:1172-1176

[8] Roncero, V.G., Costa, M.C.A., Ebecken, N.F.F..Text mining on a grid environment.WIT Transactions on Information and Communication Technologies, v 42, p 13-21, 2009

[9] Bezdek J C . Pattern Recognition with Fuzzy Objective Function Algorithms . New York : Plenum Press, 1981

[10] Wang Li Juan, Guan Shou Yi, Wang XiaoLong, Wang XiZhao. Fuzzy C Mean algorithm based on feature weights Chinese Journal of Computers , 2006, 29(1D) : 1797—1803(in Chinese)

[11] Li Jie, Gao XinBo, Jiao LiCheng. A new feature weighted fuzzy clustering algorithm. Acta Electron Sinica, 2006, 34(1),pp:89—92(in Chinese)

[12] Huang J Z,Ng M K，Rong H，Li Z．Automated variable weighting in k-means type clustering．IEEE Transactions on Pattern Analysis and Machine Intelligence，2005，27(5),pp:1—12

[13] Wang Qiang，Ye Yunming，Huang J Z．Fuzzy k-means with variable weighting in high dimensional data analysis/Proceedings of the 9th International Conference on Web—Age Information Management．Zhangjiajie，China，2008,pp:365—372

[14] WANG Jun, WANG Shi—Tong, DENG Zhao—Hong.A Novel Text Clustering Algorithm Based on Feature Weighting Distance and Soft Subspace Learning.chinese journal of computers,Vol.35 No.8,2012,pp:1655-1665

[15] Lu Ming-Yu,Yao Xiao-Na,Wei Shan-Ling.BBS hot topic mining algorithm based on fuzzy clustering.Dalian Haishi Daxue Xuebao/Journal of Dalian Maritime University, v34,n4, pp:52-54+58,2008

[16] Zhang Chenyi,Sun Jianling,Ding Yiqun.Topic mining for microblog based on MB-LDA model.Jisuanji Yanjiu yu Fazhan/Computer Research and Development, v48, n10, pp:1795-1802,October,2011

[17] Steyvers M,Griffiths T.Probabilistic topic models.Handbook of latent semantic analysis,2007,427(7),pp:424-440

[18] Sethi A,Upadrasta B,Bangalore K.Introduction to Probabilistic Topic Modeling.2012

[19] Wainwright M J,Jordan M I.Graphical models,exponential families,andvariational inference.Foundations and Trends in Machine Learning,2008,1(1-2)pp:1-305

[20] Heinrich G.Parameter estimation for text analysis.Technical report,Technical report,2005

[21] Blei D M,Jordan M I,et al.Variational inference for Dirichlet process mixtures.Bayesian analysis,2006,1(1),pp:121-143

[22] Zhao W X,Jiang J,Weng J,et al.comparing twitter and traditional media using topic model[M].Advances in information Retrieval.Springer Berlin Heidelberg,2011,pp:338-349

[23] Ramage D,Hall D,Nallapati R,et al.Labeled LDA:A supervised topic model for credit attribution in multi-labeled corpora[C].proceedings of the 2009 comference on Empirical Methods in Natural Language Processing:Volume 1-Volume 1.Association for Computational Linguistics,2009,pp:248-256

[24] Zhang Chenyi,Sun Jianling,Ding Yiqun.Topic Mining for Microblog Based on MB-LDA Model.Joural of Computer Research and Development,48(10),pp:1795-1802,2011

[25] R https://www.r-project.org/