Search Terms Construction for Financial Advertisements Acquisition in Search Engines

Qun Dong [a], Changyong Guo [b], Zhaoxin Zhang [c] and Ning Li [d]

School of Computer Science and Technology, Harbin Institute of Technology, Weihai 264200, China;

[a]ouc_dongqun@126.com, [b]cyguo@hitwh.edu.cn, [c]heart@hit.edu.cn, [d]li.ning@hit.edu.cn

Keywords: nature language processing; graph-based algorithm; search terms construction.

Abstract: Internet financial advertisements have become the top priority in the digital economy and financial advertisements acquisition is the data basis for all relevant researches. We will study on the construction of search terms, a primary task of obtaining financial advertisements in search engines, which is a problem related to semantic extension. In this paper, an unsupervised, graph-based algorithm that simultaneously incorporates the position of words and their frequency in documents to choose search terms for financial advertisements acquisition is proposed. Experiments are carried out to prove that the algorithm compared with other algorithms is more efficient in terms of the time performance and can find more financial advertisements in search engines.

1. Introduction

With the development of Internet, the digital economy has become the focus of people's attention. And Internet financial advertisements are the top priority in the digital economy. Internet financial advertisements acquisition is the data basis for all relevant researches. Financial advertisements acquisition from search engines, the most widely used tool, is one of the most important ways. In light of our observation, when searching for a word with financial semanteme characteristics, search engines are more likely to return us a few financial advertisements at the top of the search results. Therefore, the paper will study on the construction of financial search terms that provides a basis for financial advertisements acquisition from search engines and analysis of financial advertisements. And it also provides an idea for other researches that need to get advertisements from search engines or extract information from text.

In this paper, the search terms with financial semanteme characteristics are obtained from a large number of financial texts, that is a research related to semantic extension. In this aspect of research, the word extensions based on dictionary date back to 1950.[1] By generating the synonyms dictionary, we can find the synonyms according to the dictionary and realize the word extension. The domain ontology is an extension of the dictionary-based approach.[2] And a keyword semantic expansion algorithm based on domain ontology is proposed.[3] But it still need to create a dataset of domain ontology which need the expertise of specific areas. The methods mentioned above generally require expertise in the field. In our case, these approaches require us to understand a lot of financial domain expertise. So it doesn't work for us who do not have the accumulation of financial knowledge. There are also ways to realize semantic extension by constructing word vectors based on deep learning.[4] Word2Vec[5, 6] that constructs word vectors from context and expresses semantic similarity by the distance between word vectors is a common word vector model. The ELMo[7] model that can handle the Word2Vec model's inability, dealing with polysemy, can do the same thing. These models based on deep learning are unfriendly to the computer performance and the time performance. So they are not suitable for training sets that change frequently. But When we get the advertisement texts, we will add it to the training set for iterative training. So these models are not suitable for us.

In natural language processing, there are many processing methods, one of which is graph-based. Especially in terms of keyword extraction, algorithms like TextRank, SingleRank and TopicRank are

all graph-based models that construct text into graphs and use some graph-based specific algorithms.[8, 9, 10] In terms of topology partitioning of graphs, Vincent D Blondel[11] proposed louvain algorithm that iterates according to the modularity. Modularity has been used to compare the quality of the partitions obtained by different methods, but also as an objective function to optimize. And experiments show that it has good performance in time complexity.

In view of the weaknesses of semantic extension and the inspiration of graph-based keyword extraction, the construction of financial search terms algorithm based on the graph structure is proposed for Chinese financial texts and the training set constantly updated. Firstly, the algorithm obtains the phrases with context information through word segmentation and part of speech selection. Then it constructs the graph through the phrases and use louvain algorithm to cluster. Finally, it chooses words in the same cluster with financial words like funds, loans and so on. And the words are the result of the search terms for financial advertisements. The algorithm, with a good efficiency in the time performance, is more suitable for the training set frequently updated. Because it is based on graph structure and considers the efficiency of topology partitioning of graphs.

2. Financial Search Terms Construction Model

In this section, we describe the construction of financial search terms, our fully unsupervised, graph-based algorithm, that simultaneously incorporates the position of words and their frequency in texts to choose the words with financial semanteme characteristics. Our algorithm involves three essential steps: (1) the graph construction at words level; (2) the clustering of the graph; and (3) the selection of clusters. The integrated architecture and relationship among steps are shown in figure 1. And these steps are detailed below.

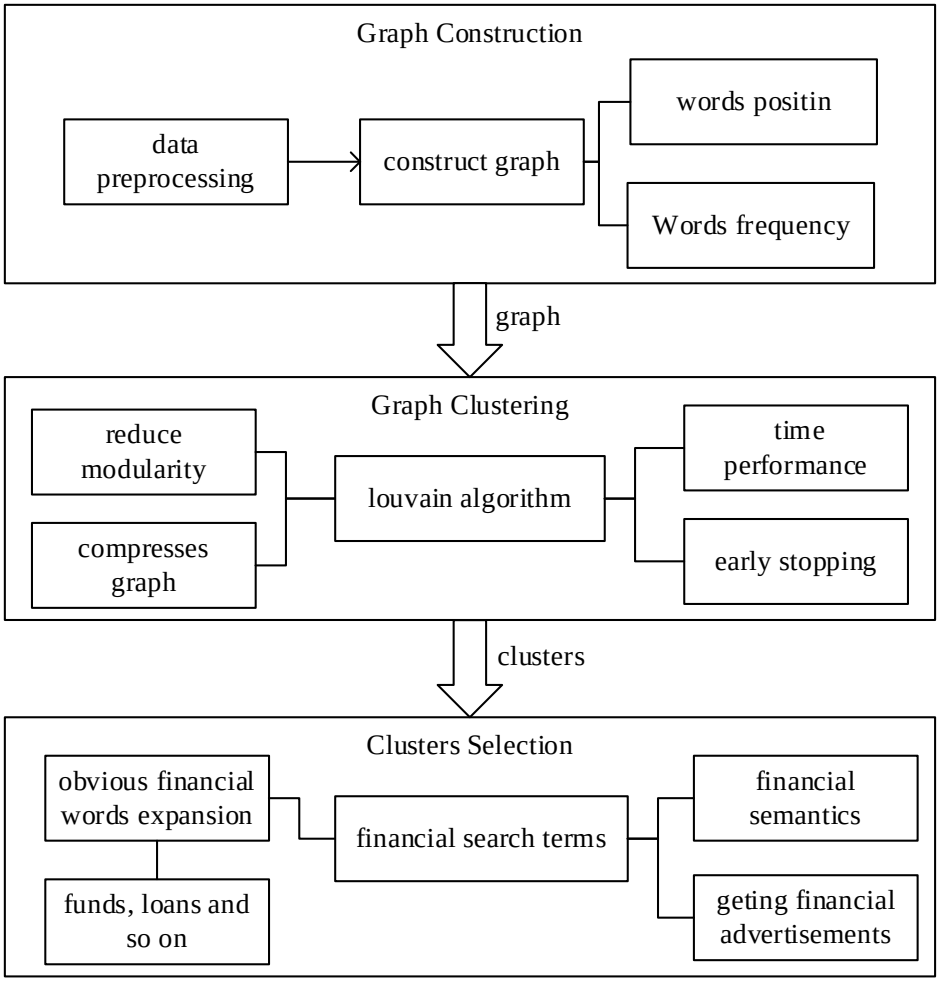

Figure 1: Model Architecture

2.1 Graph Construction.

In our model, the graphs are built from Chinese natural language texts that are related to finance. For languages like English one can assume that word boundaries are given by whitespace or punctuation. However words are never segmented by whitespace in Chinese. So the texts need to be processed by word segmentation before the graph construction. And our purpose is to build a batch of search terms, so we only choose nouns and verbs for the graph construction. The graph built from the processed texts may include multiple or partial links between the vertices. And it may be useful to include the 'strength' of the connection between two vertices. So we use an undirected weighted graph to represent the relationship among different words in the texts.

Formally let t be the texts processed by word segmentation. We build a word graph G = (V, E, w) for t. And we define a window whose size is n. Two words are in the same window if they co-occur within a window of n contiguous tokens. In a word graph, each unique word corresponds to a node. Two nodes v_i and v_j are connected by an edge $(v_i, v_j) \in E$ if the words corresponding to these nodes appear in the same window according to t. The weight of an edge $(v_i, v_j) \in E$ is computed based on the count of the two words appearing in the same window.

When constructing the graph, we take the context semantic information of the word into full consideration by the position of the word and the relative position among different words. Therefore, the relationship among words will be shown by the way of edge and weight in the graph.

2.2 Graph Clustering.

From the above description, we can turn a text set into a graph. Then in terms of topology partitioning of graphs, the nodes, corresponding words, on the graph are clustered, and the words located in the same cluster with financial words, such as funds, loans, can be found. Since the context semantic information of words is fully considered in the construction of the graph, the words in the same cluster with financial words is related to financial semantics.

In order to obtain more financial advertisements, we will continue to add the financial advertisement texts to the dataset of the graph construction. That is a process of dynamic expansion. Therefore, we have a high demand for the time performance of graph partitioning algorithm. And our goal is to obtain the search term. So the semantic information of the final words is not as high as that of synonyms.

Therefore, we finally chose to use louvain algorithm that finds high modularity partitions of large networks in a short time. It is a heuristic algorithm that divides nodes in a graph by reducing the modularity values, and then compresses the graph. It repeats the above two steps until the modularity is no longer reduced. But when louvain algorithm is used to construct financial search terms, it should be completed before the modularity not reducing. Because the number of clusters divided by the unmodified louvain algorithm is small and the number of words in each cluster is large, the semantic information among words may not have such great relevance. Therefore, clusters are selected after the first iteration in this paper.

2.3 Clusters Selection.

We finally select the desired clusters. Find some words with obvious financial semantic characteristics, such as funds, loans and so on. Then find the clusters of these words and select all the words in those clusters as financial search terms. Since the context semantic information of words is fully considered in the construction of the graph, the words in the same cluster with financial words should be related to financial semantics. So we can get financial advertisements through the selected search terms.

3. Experiments

To demonstrate the effectiveness of the proposed method, we used sohu financial news and sina financial news to perform the experiments. Table 1 provides detailed information about each dataset.

Before the experiments, we need to perform the word segmentation on the Chinese text. And our purpose is to build a batch of search terms, so we only choose nouns and verbs for the graph construction. We use the jieba library for word segmentation and part of speech selection. And the gensim tool was used to construct word vectors as a comparative experiment. And the experiment is conducted on a computer configured with core 2.2GHz CPU and 8GB memory, and the operating system is Ubuntu 18.04.4 LTS.

Table 1 Datasets Summary

Dataset	Piece	Len	Lang
sohu FN	5000	612	CN
sina FN	10000	618	CN

Table 1: A summary of the datasets, including the piece of news, the average text length after word segmentation and the language of the dataset.

sohu FN: souhu financial news from the news datasets of sogou labs. The news datasets includes news data of 18 channels. And we adopt the financial news as the experimental data.

sina FN: sina financial news from the training set of THU Chinese Text Classification. The datasets include 16 classes. And we also adopt the financial news as our experimental data.

In terms of the graph construction, we set windows size as 5 by reference to TextRank. At the same time, Word2Vec model, which does also not require additional professional knowledge, is used to train word vectors and synonyms based on vector distance are selected for word extension. That is used as a comparative experiment. The details of the experiment and the time efficiency between different methods are listed in table 2.

Table2 Methods Detail

Dataset	Nodes number	edges number	Iteration number	graph clustering time(s)	word vector cons- truction time(s)
sohu FN	35914	756376	3	0.24062	53.39749
sina FN	40844	1041778	3	0.238706	79.65998

Table 2: The details of our algorithm and the time efficiency between the two methods, including the number of nodes of the graph, the number of edges of the graph, the iteration number of the graph clustering, the time of graph clustering and the time of word vector construction based on Word2Vec model.

We can find that our algorithm took much less time than Word2Vec model. Our algorithm takes about a second whereas the Word2Vec model takes about a minute. And we can find that the number of clusters after the second iteration is very small from figure 2. So the number of words in each cluster is large and the semantic information among words in each cluster may not have such great relevance. Therefore, clusters are selected after the first iteration in this paper.

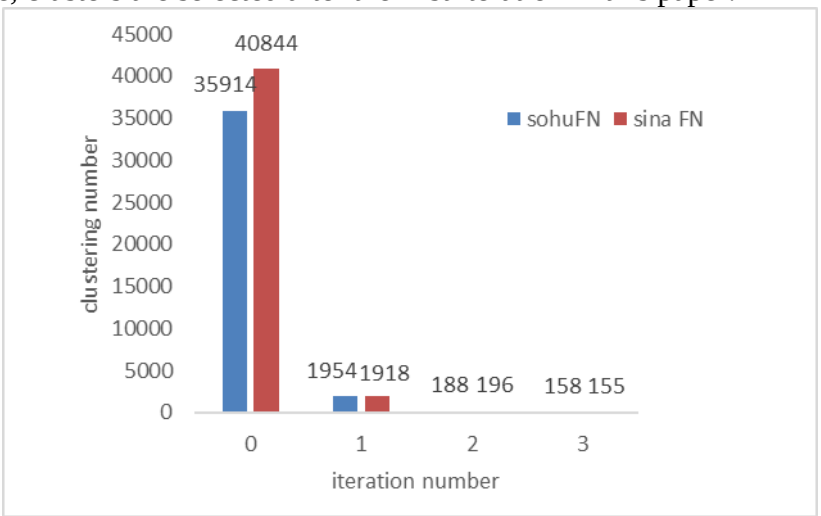

Figure 2: Relation between iterations number and clusters number

In terms of the selection of clusters of the experiment, we choose loan as an example in sohu financial news. The words in the same cluster with the loan after the first iteration are selected as financial search terms. And we ended up using the bing search engine as an example for advertising. The results are listed in table 3. We can find that only 3 words appear in Word2Vec model and not in ours. And only " credit", as search terms, of the 3words find advertisements in the search engine. But the words, listed in table3, only occurring in our model can all find financial advertisements.

Of course, some words without obvious financial semanteme characteristics also appear in our model. But our purpose is to get search terms, not synonyms. And search terms with financial semanteme characteristics may not get advertisements. So words with financial semanteme characteristics or not, which can not find advertisements for consecutive days in search engine, will be removed from the search terms in later data acquisition project.

Table 3 results of models

Seed	Words in both models	words only in Word2Vec model	words only in our model
loan	housing loan, bank loan, house purchase, loan limit, loan granting, mortgage	lending, credit, deposit	foreign exchange reserves, original stocks, stock speculation, auto loans, etc

Table 3: the result of the construction of financial search terms, including the words occurring in both models, only in the Word2Vec model and only in our algorithm.

Through the above attack experiments, it can be stated that our algorithm is more efficient in terms of time, which is more suitable for the continuous updating of data sets. And we can find more financial advertisements according to our algorithm.

4. Conclusion and future work

The algorithm of financial search terms construction based on graph is proposed in this paper. The method can also be used in the problem of semantic similarity, which is more efficient than the current popular neural network model. And the search terms constructing by our algorithm can get more financial advertisements in search engines.

The problem that the number of clusters is small and each cluster has more words is solved by stopping the iteration early in this paper. But there's no theory behind it. In the future work, we will choose an appropriate value, related to the modularity, as a condition to stop the iteration.

Acknowledgments

This work is supported by National Key R&D Program of China No.2018YFB0804703, the National Information Security 242 Project of China under Grant No.2019A035

References

[1] Qi M, Yong-Feng H. Query expansion based on Web knowledge base and search engine[J]. Journal of Computer Applications, 2012.

[2] Navigli R, Velardi P. An analysis of ontology-based query expansion strategies[C]// Proceedings of the ECML /PKDD-2003 Workshop on Adaptive Text Extraction and Mining. CavtatDubrovnik, Croatia:[s. n.], 2003: 42 – 49.

[3] Zou G, Xiang Y. Information search model based on domain ontology[J]. Journal of Tongji University(Natural Science), 2009, 37(4):545-549.

[4] Fu R, Guo J, Qin B, et al. Learning Semantic Hierarchies: A Continuous Vector Space Approach[J]. Audio, Speech, and Language Processing, IEEE/ACM Transactions on, 2015, 23(3):461-471.

[5] Mikolov T, Chen K, Corrado G, et al. Efficient Estimation of Word Representations in Vector Space[J]. Computer Science, 2013.

[6] Mikolov T, Sutskever I, Chen K, et al. Distributed Representations of Words and Phrases and their Compositionality[J]. Advances in Neural Information Processing Systems, 2013, 26:3111-3119.

[7] Peters M E, Neumann M, Iyyer M, et al. Deep contextualized word representations[J]. 2018.

[8] Mihalcea R, Tarau P. TextRank: Bringing Order into Texts[J]. Emnlp, 2004:404-411.

[9] Wan X, Xiao J. CollabRank: towards a collaborative approach to single-document keyphrase extraction[C]// International Conference on Coling. DBLP, 2008.

[10] Bougouin A, Boudin F, Daille B. Topicrank: Graph-based topic ranking for keyphrase extraction[C]. //Proc of the 6th International Joint Conference on Natural Language Processing, pages 543–551, 2013

[11] Blondel V D, Guillaume J L, Lambiotte R, et al. Fast unfolding of communities in large networks[J]. Journal of Statistical Mechanics: Theory and Experiment, 2008, 2008(10):0-0.