Identifying Obstacles in Data Sharing by Automatic Extraction of Problematic Points in Documents

Yan Wan
School of Economics and Management
Beijing University of Posts and Telecommunications
Beijing, China
e-mail: wanyan@bupt.edu.cn

Guanhao Chen
School of Economics and Management
Beijing University of Posts and Telecommunications
Beijing, China
e-mail: wangyalu@bupt.edu.cn

Yalu Wang
School of Economics and Management
Beijing University of Posts and Telecommunications
Beijing, China
e-mail: 17888843379@163.com

Jinping Gao
School of Economics and Management
Beijing University of Posts and Telecommunications
Beijing, China
e-mail: doc.gao@126.com

***Abstract*— This paper aims to propose an automatic viewpoint extraction method using open data obstacle extraction as an example. Open data (data sharing) is very important because it reduces job repeatability and increase productivity and openness of work. However, open data in China is not as well developed as we wish. It is hindered by various problems, such as the willingness to share, the incompatible of data formats, etc. In order to identify different problems, then allocate to relevant parties to tackle these problems, we adopt an automatic extraction algorithm of natural language processing techniques, to automatically identify problematic points (obstacles) of data sharing from relevant literature. In this paper, we first construct a vocabulary for "obstacles", so that machines can find "obstacles" in literature more accurately. Then, an extraction algorithm combined with word2vec and Pointwise Mutual Information (PMI) is proposed, to automatically find the sentences that talk about "obstacles" of open data in documents. An experiment of this method is carried out and analyzed. It shows that the proposed method can be a very good tool for similar tasks that need to find viewpoint from a large amount of documents but cannot be done by simple keyword searches.**

Keywords-open data; feature extraction; data sharing obstacles.

I. Introduction

In recent years, with the rapid growth of data production and the extensive application of information technology, data volume is accumulated at an unprecedented speed. While data is constantly being produced at an unimaginable speed, it is also distributed in different institutions that own data, including governments, enterprises, scientific research institutions and other departments. How to make full use of this massive data is highly valued by government and researchers.

The main theme of big data era lies in data sharing. By analyzing and mining the value of data, we can save resources, improve the quality of product and make profits in ways that are more effective. Nowadays data openness is indispensable for achieving these purposes. The concept of data openness has been raised by Auer et al. [1], which is defined as "Open data is the idea that some data should be freely available to everyone to use and republish as they wish, without restrictions from copyright, patents or other mechanisms of control." The shared data include government data, science data and enterprise data. In open data process, there exists various problems, which hinder data dissemination, data explore and value realization. In order to solve these problems, we need first to identify these problems then assign them to responsible parties to seek solutions. This paper focus on problem identification, by extracting problematic points from open data literature using natural language processing technology.

In most open data studies, scholars carried out their research by reading a large amount of documents. This is time-consuming and may miss some important points. This paper suggests an automatic sentence extraction method which finds out issues encountered in open data process automatically in a more comprehensive and efficient way. The method we propose is Problematic Points Extraction Model (PPEM). It combines word2vec model and Pointwise Mutual Information (PMI) based method. Its performance is validated by comparing it with word2vec and PMI.

In brief, this paper includes three parts:

a) Propose a sentence extracting model to identify the problematic points from documents discussing open data issues.

b) Evaluate the above model using text mining evaluation method.

c) Classify the found open data problematic points.

In Section II, a survey of the related work is described. In Section III, we introduced the method and the procedure we carried our research. In Section IV, the result of our proposed method is described and compared with other possible methods. Finally, in Section V, conclusion and future work are discussed.

II. Related Works

There exist many studies about the problems in open data process, but seldom do researchers focus on the automatic problems extraction. In this section, we examine three kinds of literature that contains open data, defect detection and vocabulary construction. The open data literature mainly demonstrates the research status of open data problems. Then we review the study of defect detection, which is a reasonable reference to our research. As we employ the vocabulary construction method to extract problematic points, we also make a brief introduction of this technology.

A. Open Data

In today’s society, customers using mobile phone to accomplish many activities in their everyday life. Meanwhile, their personal information and behavior data has been stored by the service provider and related institute such as government departments, banks, enterprises and so on. The massive data stored in the above places has significant value for the development of economics and enhancement of people’s living standard. However, the reality is that most of the data remains unused and the value of data is being wasted. It is time that we undertake some strategies for data liberalization.

In the process of data liberalization, the government as the master of a large number of high-value data resources and the standardization of the use of information resources are the main force to promote the open sharing of data. To promote the e-government programs, Dawes designed an electronic government information access programs in his research [2]. In Conradie and Choenni’s research [3], they regarded the way of data acquisition, storage and use, and the suitability of data openness, as crucial indicators for open data releasing. They conducted interviews and workshops and finally classified the barriers to data release as fear of false conclusions, financial effects, opaque ownership and unknown data locations, low priority [3].

B. Difficulty Discovery

In this paper, we define the concept of problematic words, which are problem-oriented words in sentences to indicate problematic sentences. The research conducted by Abrahams et al. [4] in Virginia Tech provides a good example for us. In their research, they first applied automatic defect discovery approach on discussion threads of vehicle forums and proposed Vehicle Defect Discovery System [4]. They also defined ‘smoke words’ concept to better recognize vehicle defects. Subsequently, they undertook another study about automotive defect and consumer electronics defect, and proposed Social Media Analytic framework using Text(SMART) for Quality Management (QM) tasks [5]. Later in 2017, they extended their methods into the defect discovery of toys and dishwasher appliances [6][7].

C. Vocabulary Construction

Vocabulary construction mostly appears in sentiment analysis tasks, by doing so, researchers can distinguish the linguistic unit between positive and negative polarities. Vocabulary, also called lexicon, can be divided into two categories, domain-oriented and domain-independent. Kanayama and Nasukawa [8] proposed an unsupervised lexicon building method, in which they used the “polar atoms” as the linguistic unit and calculate context coherency. Our research aims at the problems exist in open data domain, and we suppose to establish a domain-oriented lexicon.

III. METHODS

Using keywords “open data”, and then eliminating duplicate and too short articles, 486 documents are obtained from CNKI database (the most popular Chinese academic database). 436 documents are used to train the Word2vec model and the remaining 50 documents are for experiment. In this section, the PPEM’ s procedures, including data pre-processing, lexicon construction and problematic points extraction, will be explained in detail.

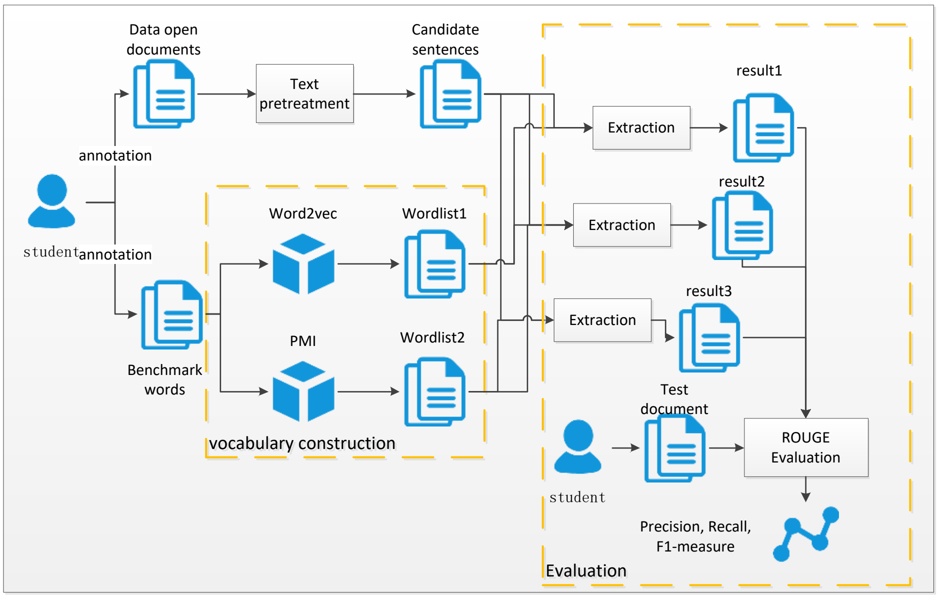

Figure 1. Procedure of problematic points extraction model (PPEM)

TABLE I. PROBLEMATIC WORDLIST CONSTRUCTION

Benchmark words	Problematic words
问题(problem)	挑战(challenge),难题(difficulty), 障碍(obstacle),后顾之忧(future trouble)
缺乏(shortage)	缺少(shortage), 不够(not enough), 薄弱(weak), 缺失(lack)
不足(shortcoming)	诸多(numerous)，障碍(obstacle)，明显(obvious)，差异(difference), 突出(prominent)
困难(difficulty)	现象(phenomenon), 严重(serious), 安全隐患(potential safety hazard), 难(difficult)
差距(gap)	差异(difference), 差别(difference), 明显(obvious),成就(achievement), 落后（fall behind）
无法(cannot)	不能(unable), 难以(difficult to), 能够(able to), 很难(hard)
尚未(not yet)	并未(wasn’t),尚(yet), 尽管(despite), 较晚(later), 虽然(although)
阻碍(hinder)	制约(restrict), 起到(serve as), 忽视(ignore), 严重(serious), 改变(alter)
挑战(challenge)	机遇(opportunity), 难题(puzzle), 问题(problem), 面临(confront)

A. *Text Pre-processing*

After the collection of open data documents, the first step is text pre-processing to clean and formalize the texts. First, we eliminate the graphs, tables and annotations of the text. Next, we write a computer program to remove the literature serial numbers, non-Chinese characters, and other redundant content that influence the quality of analysis. For stop word elimination and word segmentation, we employ the python modules, *jieba* and *snownlp*, which are useful natural language processing tools.

B. *Lexicon Construction and Problematic Point Extraction*

As we can see in Figure 1, next to the text pre-processing is vocabulary construction step. In this step, we introduce Word2vec and PMI to calculate the similarity of words. Traditionally, for information extraction, many researchers use some kind of algorithm to score the sentences and select the sentences that are highly ranked. In our study, we suppose to collect the similarity scores and select the top n words to form the problematic vocabulary. The wordlist constructed by word2vec and PMI is named as wordlist1 and wordlist2, illustrated in table 3. In the next, we can use these scores to calculate the sentence score, which contain words in above vocabulary. The sentence scores generate by different models are certainly different. In Figure 1, we define the outcome of PPEM, Word2vec and PMI as result1, result2 and result3.

The PPEM model is organized by a simple combining algorithm and the wordlists generated by word2vec and PMI. The metric of PPEM is calculated as in (1).

$$Weight(w) = \rho\ Weight_{w2v}(w) + (1-\rho) Weight_{PMI}(w) \qquad (1)$$

In (1), ρ weighs the significance of word2vec, and 1-ρ weighs the significance of PMI. We can adapt optimal parameters by repeated experiment to our extraction model.

C. *ROUGE Evaluation*

The ROUGE evaluation method is extensively used in text mining field, which represents Recall-Oriented Understudy for Gisting Evaluation. This evaluation model operates through comparing candidate result with reference result [8][9]. The candidate result is generated by computer, while reference result is generated by experts. To validate the capability of PPEM model in a relatively convincing way, we choose three kind of metrics in ROUGE evaluation system, namely ROUGE-1, ROUGE-2, ROUGE-L. ROUGE-1 and ROUGE-2 represents the overlap of unigrams and bigrams between candidate result and reference result. ROUGE-L metric represents the overlapping longest common subsequence.

IV. RESULT AND EVALUATION

An experiment is designed to extract the problematic points by sentence extracting model PPEM, which employs the word2vec model and PMI indices. We use python program to realize the methods of word2vec and PMI, and ask three students who major in management science and engineering to accomplish the manual annotation task.

TABLE II. ROUGE METRICS OF PPEM, WORD2VEC, PMI

Model		**Word2vec**	**PMI**	**PPEM**
ROUGE-1	**P**	58.6	60.3	61.6
	R	82.8	80	87
	F1	68.6	68.8	72.1
ROUGE-2	**P**	39.3	42.0	44.0
	R	55.5	55.6	62.1
	F1	46.2	47.9	51.5
ROUGE-3	**P**	51.4	53.5	55.0
	R	72.6	71.0	77.7
	F1	50.2	51.0	54.4

As is shown in Table 2, we calculate the precision, recall, F1-measure of three metrics, which represented by P, R, F1. Precision is the percentage of true positive samples comparing with the true samples. In contrast, recall is the percentage that true positive samples divided by all positive samples. F1-measure represents the harmonic mean of precision and recall.

The output data shows that the sentence extracted by word2vec has higher recall but is lower in precision than PMI. Comparing the F1-measure, it can be summarized that the PPEM model have better performance in all these metrics, which denote that

PPEM is a more suitable approach for problematic points extraction.

V. Conclusion and Future Work

In this paper, we have studied the problems in open data field by employing word2vec and PMI techniques. The main contributions of our research are as follows:

(1) applying natural language processing approach to automatically extract the problems in open data field from a large number of documents.

(2) proposing a new model PPEM which combines word2vec and PMI. Evaluation of the performance of the model is conducted.

A vocabulary for problem discovery is constructed to improve the performance of PPEM.

By observing the experiment results, we can draw conclusions as follows:

The PPEM model performs better than extracting by word2vec or PMI alone. In other words, it is a reasonable choice for researchers to apply PPEM model to coping with problems in specific domains. The output sentences extracted by PPEM represent the main problems in open data field in China. To summarize, the main problematic points of open data focus on four aspects, as shown in Table 3.

TABLE III. Problems in different fields

Field	Problems
Data source	• government neglect of data liberalization • weak data storage • lack of personnel in charge of building open data departments • scattered data of enterprises • lack of a unified open standard
Data dissemination	• privacy protection • inconsistent standard • imperfect data legislation
Data analysis	• short of data analysis professionals • the task of data analysis is not yet clear
Data application	• big data technology has not been popularized among the general public • less data open enterprises • enterprises lack of funds

In general, in order to promote further open data, issues should be considered from national strategic level and a special leading group can be set up to coordinate data openness. Local governments need to implement the opening-up policy, strengthen the construction of an open platform for data and eliminate the isolation of information. Legislative and judiciaries need to promote data-related legislation and regulate the way data is used. Large-scale enterprises need to regulate data formats and desensitize sensitive data. They should abide by the data usage rules, not use the data for illegal activities, and on the other hand, enhance their understanding of the value of data and their ability to obtain and analyze data. Therefore, the process of opening up the data requires the joint promotion of the three parties, including the government, enterprises and individuals, in order to continuously move forward. However, because this paper is mainly concerning the data analysis methodology, the open data issue itself has not been explored enough and shall be discussed further in future work.

Acknowledgment

This research partly supported by Natural Science Foundation of China (Nos. 71874018, 71471019, 71772017).

References

[1] S. Auer, et al., "DBpedia: A Nucleus for a Web of Open Data," The Semantic Web: the sixth International Semantic Web Conference, 2nd Asian Semantic Web Conference, ISWC + ASWC, Nov. 2007, pp. 722-735, doi: 10.1007/978-3-540-76298-052.

[2] S. S. Dawes, T. A. Pardo and A. M. Cresswell, “Designing electronic government information access programs: a holistic approach,” Government Information Quarterly, vol. 21, 2004, pp. 3-23, doi:10.1016/j.giq.2003.11.001.

[3] P. Conradie and S. Choenni, “Exploring process barriers to release public sector information in local government,” The twelfth International Conference on Theory and Practice of Electronic Governance Icegov(ICEGOV), ACM, Oct. 2012, pp. 5-13, doi:10.1145/2463728.2463731.

[4] A. S. Abrahams, J. Jiao, G. A. Wang and W. G. Fan, “Vehicle defect discovery from social media,” Decision Support Systems, vol. 54, Dec. 2012, pp. 87-97, doi: 10.1016/j.dss.2012.04.005.

[5] A. S. Abrahams, W. Fan, G. A. Wang and W. G. Fan, “An integrated text analytic framework for product defect discovery,” Production & Operations Management, vol. 24, Sept. 2014, pp. 975-990, doi: 10.1111/poms.12303.

[6] M. Winkler, A. S. Abrahams, R. Gruss and J. P. Ehsanib, “Toy safety surveillance from online reviews," Decision Support Systems, vol. 90, Oct. 2016, pp. 23-32, doi:10.1016/j.dss.2016.06.016.

[7] D. Law, R. Gruss and A. S. Abrahams, “Automated defect discovery for dishwasher appliances from online consumer reviews,” Expert Systems with Applications, vol. 67, Jan. 2017, pp. 84-94, doi:10.1016/j.eswa.2016.08.069.

[8] H. Kanayama and T. Nasukawa, “Fully Automatic Lexicon Expansion for DomainOriented Sentiment Analysis,” Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing, ACL, Jan. 2006, pp. 355-363, doi:10.3115/1610075.1610125.

[9] C. Flick, “ ROUGE: A Package for Automatic Evaluation of summaries," Proceedings of the Workshop on Text Summarization Branches Out(WAS), Jan. 2004, pp. 10, doi:doi:http://dx.doi.org/.

ISBN: 978-1-61208-757-3