ATLANTIS PRESS

Advances in Intelligent Systems Research (AISR), volume 155

2018 International Conference on Advanced Control, Automation and Artificial Intelligence (ACAAI 2018)

Analysis of Scientific and Technical Literature in the Big Data

Wen Zeng[1,*], Hui Li[2] and Na Qi[1]

[1]Institute of Scientific and Technical Information of China

[2]Beijing Institute of Science and Technology Information

*Corresponding author

***Abstract*—Compared with the others data, scientific and technical literature is multi-source and multiple types. Its content is more emphasis on technical and correlation. In order to analyze and evaluate it, the paper got the value of correlation based on VSM. And it introduced the value of correlation into evaluation indexes of scientific and technical papers and patents in China. Experimental results showed that the method was reasonable and it could improve the traditional evaluation method of scientific and technical literature. The work in this paper will provide a good foundation for the future research.**

Keywords- analysis; scientific and technical literature;big data

I. INTRODUCTION

With the arrival of the era of big data, data has permeated into all sides of human life[1]. We are not only users but also producers of data. With the development of scientific and technical, the growth rate of information increases and Internet improves efficiency of information sharing which brings convenience to users. Under this circumstance, users' information demand increases the rapid growth of data amount. What's more, data formats have diversified greatly so as to express uses' need vividly. Information is displayed to users in not only text but also video and audio format. The content of scientific and technical literature is the most technical which contains valuable information. The amount of scientific and technical literature and its users are huge and grow rapidly. As a result, process and analysis of multisource scientific and technical literature has become one of the most pressing issues among the study of scientific and technical literature. The characteristic of multisource scientific and technical literature, which includes scientific and technical literature periodicals, patents, scientific and technical literature policy, scientific reports, thesis etc, is the diversity of data sources, document type, document content and date format. The main features of scientific and technical literature are:1)Difference of data structure: data format of scientific and technical literature involves structured rational database, object-oriented database, semi-structured HTML/XML, free text, multimedia data etc. The differences of data format among these data sources lead to discrepancy and conflict during the data processing procedure.2)Different data standard: collected data file of scientific and technical literature may be displayed in various formats.3)Semantic ambiguity: Though attributes of one entity may be the same, ambiguity also exists because of difference of emphasis to solve the problem.4)Distributed feature: data is stored in different physical memorizer though Internet connects them logically.5)Magnanimity and instantaneity. The chief problem about scientific and technical literature analysis is the acquisition and processing based on the characteristic of them. The second is to analyze the content of scientific and technical literature and find valuable technology and knowledge.

II. RELATED WORK

In general, processing methods for scientific and technical literature only implements some basic operation, such as document scanning, indexing and OCR, to support services of Web information retrieval and reading. Intelligent data reprocessing to scientific and technical literature is not achieved. The following problems need to be solved:1) There are inevitably some processing or input errors in the stored scientific and technical literature. It's difficult to recognize and process these errors manually in the huge amount of dataset so it's necessary to employ intelligent preprocessing technology.2) Different method of data processing leads to the diversity of data storage structure, data description method etc. Mapping and normalization methods of different structure and description need to be done.3)For foreign databases, data needs to be exported from disk or website. The exported data format also should be transferred and data structure should be processed as needed.4)More detailed information should be extracted from scientific and technical literature for data mining and analysis. It includes not only title, keywords, abstract, but also includes name disambiguation, organization disambiguation, quotation and full text. Current digital database cannot provide some information we described above, so it's necessary to reprocess existing data.5)Mining knowledge from huge scientific and technical literature, finding correlation among documents, analyzing and forecasting technology trends are hot topics in the field of information science whose foundation is scientific and technical literature resources. Though lots of foreign researcher studied data mining, data analysis and achieved exciting result, most of their study was based on high-quality and limited amount of data. In China, practical survey about data collection at library shows that there is a big gap between the quality of experiment data and factual data. The result of current technology is unpredictable if it is applied to factual data. What's more, it's important to analyze the correlation among different type of scientific and technical literature because its value of technology or research is higher than that of the literature in single domain or type.

III. ANALYSIS OF SCIENTIFIC AND TECHNICAL LITERATURE

The analysis of scientific and technical literature focus on finding valuable information and solving some problems. At present, the analysis of scientific and technical papers or patents has been preliminary implemented. For example, the analysis of scientific and technical policies is based on clustering analysis of scientific and technical policy documents. By word segmentation and results of statistical calculation, scientific and technical policies are refined and clustered according to the meaning of policies. Based on the clustering results, we can establish relationship between policies and enterprises by label word, provide related scientific and technical policy by clustering for users.

Vector Space Model (VSM) has been widely used in the area of information retrieval, text classification and document clustering in the recent 30 years. In VSM, if every word in the document is used to express a text vector, the vector is high-dimensional and very spare. According to information theory, IDF is cross entropy of probability distribution of words under specified condition, and TF can enhance the words' weight to reflect the characteristic of words. So in this paper, the paper's author extracted terms from scientific and technical literature rather than words so that the extracted features can reflect the content of document and the dimension of document vector can be reduced. The paper's author had studied the method about extract terms[2][3]from scientific and technical literature and the method can be found in paper[4].In the paper, the correlation calculation for scientific and technical literature was realized based on VSM. In fact, that is, the correlation calculation for scientific and technical periodical papers and patents is realized by calculating similarity of term vectors. Because the type, structure and length of scientific and technical periodical paper and patent are different, so the paper's author proposed formula (1) as follows:

$$w_i = \sum_{t \in D,P}\left[\log\frac{M}{df_{-t-M}}\right]\cdot\frac{(k_1+1)df_{-t-m}}{k_1\left((1-b)+b\times\frac{L_d}{L_{d-avg}}\right)+tf_{dt}}\times\left[\log\frac{N}{df_{-t-N}}\right]\cdot\frac{(k_1'+1)df_{-t-N}}{k_1\left((1-b')+b\times\frac{L_p}{L_{p-avg}}\right)+tf_{ps}} \quad (1)$$

And,

df_{-t-m} : the frequency of term t in scientific and technical papers.

df_{-t-N} : the frequency of term t in scientific and technical patents.

L_{d-avg} : the average length of scientific and technical papers.

L_d : the length of single scientific and technical paper, that is, length of a scientific and technical paper containing term t.

L_p : the length of single scientific and technical patent, that is, length of a scientific and technical patent containing term t.

L_{p-avg} : the average length of scientific and technical patents.

IV. ANALYSIS AND EVALUATION ABOUT SCIENTIFIC AND TECHNICAL LITERATURE

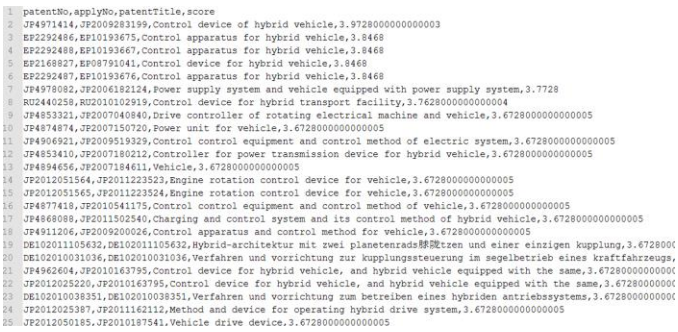
Scientific and technical paper is a kind of academic literatures. It is the important output of research work about natural and social science. It can reflect basic research, applied research work, development trend of various disciplines and evaluate the work of researchers. It has important and practical significance. Patent contains the higher business value and more useful, technical and innovative knowledge than other types of scientific and technical literature. What's more, there are many patent databases at home and abroad which contain patent documents in formative data format. It's important to evaluate patent documents for patent information analysis and strategic research. Evaluation indexes for scientific and technical paper can be divided into three categories, including maternal literature, literature and social assessment information. To be more specific, evaluation indexes of scientific and technical paper include:1) Indexes of maternal literature: type of maternal literature, impact factor of maternal literature.2) Indexes of literature: financing type, citation frequency of literature, download frequency of literature, title of author.3) Social assessment: prize winning, online citation, value of correlation between paper and patent.Because patents contain technical, legal and economic information, patents should be evaluated from three aspects. To be more specific, evaluation indexes of patent include: 1) Technical indexes of patent literature: type of patent, number of category, citation indicator, the number of patent claim, application distribution of patent, patent owner, value of correlation between paper and patent.2) Economical indexes of patent literature: scale of patent family, exploitation of patent. 3) Legal indexes of patent literature: patent litigation, patent life, and patent extension.

Though evaluation indexes are different for different types of scientific and technical literatures, evaluation method is consistent. It's necessary to determine weights of different indexes. To evaluate paper and patent comprehensively, the paper took experts' advice and adoped some objective method such as Analytic Hierarchy Process (AHP)[5]. AHP was a classical method put forward by T.L.Saaty, a professor at University of Pittsburgh, based on network system theory and multi-objective integrated method. It considered a multi-objective decision problem as a system, divided goal into several sub-goals or criterions, and then grouped into several levels of multi-indicators. Through qualitative index and fuzzy quantification to calculate hierarchical single order and total order, as the goal, the decision system of multiple optimization can be established.

V. EXPERIMENT ANALYSIS

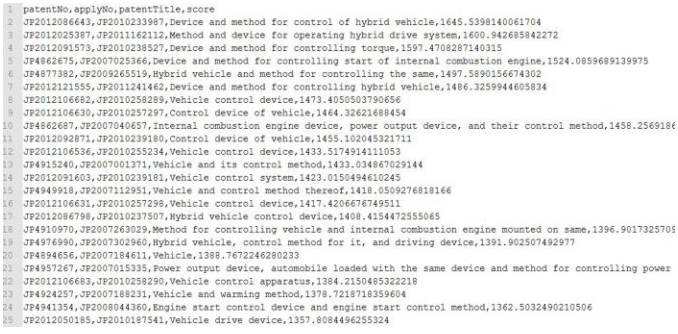
In order to validate our idea and method, we took new energy vehicles as an example, we processed data as follows: the number of papers is 7750, and the number of patents is 1263. Through the calculation of correlation between scientific and technical papers and patents mentioned above, the value

of correlation was used as one of indexes to participate in the evaluation of scientific and technical papers and patents. We evaluate two methods respectively. They are: unadopted the value of correlation index, adopted the value of correlation index. The calculated values of two methods were gotten in descending orders. If the calculated value of a paper or patent is higher, it shows that the importance of this paper or patent is greater. The results are shown in FIGURE1 and FIGURE 2.



```
patentNo,applyNo,patentTitle,score
JP4971414,JP2009283199,Control device of hybrid vehicle,3.9728000000000003
EP2292486,EP10193675,Control apparatus for hybrid vehicle,3.8468
EP2292488,EP10193667,Control apparatus for hybrid vehicle,3.8468
EP2168827,EP08791041,Control device for hybrid vehicle,3.8468
EP2292487,EP10193676,Control apparatus for hybrid vehicle,3.8468
JP4978082,JP2006182124,Power supply system and vehicle equipped with power supply system,3.7728
RU2440258,RU2010102919,Control device for hybrid transport facility,3.7628000000000004
JP4853321,JP2007040840,Drive controller of rotating electrical machine and vehicle,3.6728000000000005
JP4874874,JP2007150720,Power unit for vehicle,3.6728000000000005
JP4906921,JP2009519329,Control control equipment and control method of electric system,3.6728000000000005
JP4853410,JP2007180212,Controller for power transmission device for hybrid vehicle,3.6728000000000005
JP4894656,JP2007184611,Vehicle,3.6728000000000005
JP2012051564,JP2011223523,Engine rotation control device for vehicle,3.6728000000000005
JP2012051565,JP2011223524,Engine rotation control device for vehicle,3.6728000000000005
JP4877418,JP2010541175,Control control equipment and control method of vehicle,3.6728000000000005
JP4868088,JP2011502540,Charging and control system and its control method of hybrid vehicle,3.6728000000000005
JP4911206,JP2009200026,Control apparatus and control method for vehicle,3.6728000000000005
DE102011105632,DE102011105632,Hybrid-architektur mit zwei planetenradsätzen und einer einzigen kupplung,3.6728000
DE102010031036,DE102010031036,Verfahren und vorrichtung zur kupplungssteuerung im segelbetrieb eines kraftfahrzeugs,
JP4962604,JP2010163795,Control device for hybrid vehicle, and hybrid vehicle equipped with the same,3.67280000000000
JP2012025220,JP2010163795,Control device for hybrid vehicle, and hybrid vehicle equipped with the same,3.67280000000
DE102010038351,DE102010038351,Verfahren und vorrichtung zum betreiben eines hybriden antriebssystems,3.67280000000000
JP2012025387,JP2011162112,Method and device for operating hybrid drive system,3.6728000000000005
JP2012050185,JP2010187541,Vehicle drive device,3.6728000000000005
```

FIGURE I. UNADOPTED THE VALUE OF CORRELATION INDEX



```
patentNo,applyNo,patentTitle,score
JP2012086643,JP2010233987,Device and method for control of hybrid vehicle,1645.5398140061704
JP2012025387,JP2011162112,Method and device for operating hybrid drive system,1600.942685842272
JP2012091573,JP2010238527,Device and method for controlling torque,1597.4708287140315
JP4862675,JP2007025366,Device and method for controlling start of internal combustion engine,1524.0859689139975
JP4877382,JP2009265519,Hybrid vehicle and method for controlling the same,1497.5890156674302
JP2012121555,JP2011241462,Device and method for controlling hybrid vehicle,1486.3259944605834
JP2012106682,JP2010258289,Vehicle control device,1473.4050503790656
JP2012106630,JP2010257297,Control device of vehicle,1464.32621688454
JP4862687,JP2007040657,Internal combustion engine device, power output device, and their control method,1458.2569186
JP2012092871,JP2010239180,Control device of vehicle,1455.102045321711
JP2012106536,JP2010255234,Vehicle control device,1433.5174914111053
JP4915240,JP2007001371,Vehicle and its control method,1433.034867029144
JP2012091603,JP2010239181,Vehicle control system,1423.0150494610245
JP4949918,JP2007112951,Vehicle and control method thereof,1418.0509276818166
JP2012106631,JP2010257298,Vehicle control device,1417.4206676749511
JP2012086798,JP2010237507,Hybrid vehicle control device,1408.4154472555065
JP4910970,JP2007263029,Method for controlling vehicle and internal combustion engine mounted on same,1396.901732570
JP4976990,JP2007302960,Hybrid vehicle, control method for it, and driving device,1391.902507492977
JP4894656,JP2007184611,Vehicle,1388.767224628023
JP4957267,JP2007015335,Power output device, automobile loaded with the same device and method for controlling power
JP2012106683,JP2010258290,Vehicle control apparatus,1384.2150485322218
JP4924257,JP2007188231,Vehicle and warming method,1378.7218718359604
JP4941354,JP2008044360,Engine start control device and engine start control method,1362.5032490210506
JP2012050185,JP2010187541,Vehicle drive device,1357.8084496255324
```

FIGURE II. ADOPTED THE VALUE OF CORRELATION INDEX

To evaluate the effectiveness about automatic evaluation for scientific and technical papers and patents, the experimental results were validated and analyzed by artificial assessment. The participants are the researchers who engage in research of new energy vehicles. Accuracy calculation formula is as follows.

$$\Pr ecision = \frac{Correct\_number}{Total\_number} \times 100\% \tag{2}$$

We can find that the method that adopted the value of correlation index is better than unadopted the value of correlation index. Specific compared results about the evaluation of scientific and technical papers and patents are shown in TABLE I.

TABLE I. EXPERIMENT RESULT

Name	***Unadopted the value of correlation index***	***Adopted the value of correlation index***
papers	About 62.5%	About 70.8%
Patents	About 52.8%	About 65..2%

VI. CONCLUSION

The paper proposed analysis methods about scientific and technical literature. Experimental results showed that the methods were reasonable, and had effectiveness on the evaluation of scientific and technical papers and patents. The research work in this paper provided a research foundation for the future research about scientific and technical literatures. The shortcoming of this paper was that the imperfection of the data quality. The research quality will be certainly improved in the future.

ACKNOWLEDGMENT

This article was supported by the National Social Science Fund Project: Research on Information Analysis Method and Integrated Tools Based on Fact-type Scientific and Technical Big Data. Grant number: 14BTQ038. And it was also supported by ISTIC Innovation Fund Project. Grant number: MS2017-09.

REFERENCES

[1] C.Wang,"Soft computing in big data intelligent transportation systems",Applied Soft Computing, 2016, vol.38,pp.1099-1108.

[2] I. H.Witten, "Kea: practical automatic keyphrase extraction", in Proceedings of the 4th ACM Conference on Digital Libraries (DL'99). New York: ACM Press, 1999,pp.254-255.

[3] P.Qu, H.Wang,"Patent term extraction for information analysis ",Library and Information Services, 2013, vol.57, pp.130-135.

[4] W.Zeng,"The Research and Analysis on Automatic Extraction Technology of Scientific and technical Literature Term", XianDai TuShu Qingbao Jishu, 2014,vol.1,pp.51-55.

[5] M.A. Atkinson, B.Karpak, O.Bayazit, "A Case Study Using the Analytic Hierarchy Process for IT Outsourcing Decision Making", International journal of information systems and supply chain management, 2015,vol.8,no.1,pp.60-84.