International Journal of Database Theory and Application
Vol.6, No.5 (2013), pp.133-140
http://dx.doi.org/10.14257/ijdta.2013.6.5.12

Analysis and Research of Enterprise Technology Competent Advantage on Text Mining and Correspondence Analysis

Xingang Wang[1] and Song Liu[2]

[1]School of Information, Qilu University of Technology
University Road 3501,Changqing District, Jinan, Shandong, China
[2]School of Information, Qilu University of Technology
University Road 3501, Changqing District, Jinan, Shandong, China
[1]wxg@spu.edu.cn, [2]liusong@spu.edu.cn

Abstract

As correspondence analysis based on structured data ignores the important information contained in the text field of patent documents easily, this article combines text mining algorithms and corresponding analysis methods, and then uses the method to analyze enterprise technology competent advantage. This article describes the procedure of analyzing enterprise competitive advantage in detail. It takes the State Intellectual Property Office patent database as data source, analyses 210 authorized patents in optical communication and the top twenty applicants, gets the correspondence analysis figure.

Keywords: *correspondence analysis, text mining, enterprise technology competent advantage*

1. Introduction

With the development and operation of modern enterprise, if an enterprise wants to approach international markets or to satisfy domestic demand, the enterprise must set up its own core competencies, especially the core competencies based on technology.

With comprehensive analysis in all aspects, there are a lot of factors affecting competitiveness of enterprise, including interaction capability of customer service, organizational capability of sales and marketing, integration capability and comprehensive management ability of finance, technical force and the capability of new product research and development. Analyzed from technical perspective, core technology and the capability of production and new product research and development are core technical competencies. To analyze the technical competitive advantage of enterprise is to analyze the corresponding advantage products and technology.

This article takes the data of patent literature as analysis object to analysis technical competition capability and to obtain the corresponding dominant products and technology of enterprise. It is not enough to only analyze patent record items, such as time, inventor, applicant organizations and the corresponding number of patents, because the key technology and products of enterprise cannot be obtained through the analysis, especially cannot obtain the dominant products and technology of enterprise from analysis. We must comprehensively analyze patent text field data by text mining, and analyze the correspondence between enterprise and its dominant products and technology.

In the 1960s, Jean Paul Benzerci set up the correspondence analysis. A kind of interdependence variable statistical analysis technique did not become widespread until 1980s in occident. It majorly analyzes multidimensional frequency table of denominate variable or

ISSN: 2005-4270 IJDTA

sequencing variable, explores differences between various categories in the same variable and correspondence of different categories in different variable [1, 2]. Guo Zhigang summarized four merits of correspondence analysis: The more categories of named variables, the more obvious the analysis advantage will be; It can change named variables or sequencing variables into interval variable; It can reveal the link between categories of the row variables and column variables; It can display the link between the variable categories in graph intuitively[3].

The using conditions of the correspondence analysis include several points: the variable is a named variable or a sequencing variable, the categories of row variables and column variables are independent, the cross-tabulation consists of column variables and row variables which cannot be 0 or negatives [4].

As correspondence analysis is multi-dimensional graphical analysis, it has the graphical advantage. It also calculates singular root and characteristic root during analyzing, and quantifies the interpretive degree of total information amount and improves the scientific result. It can integrate the sample information and variable information, is widely used in many fields, such as economy, agriculture, education, health care and so on [5-8] .The method is also applied in literature analysis and patent analysis abroad [9, 11]. However, these applications are based on structured data, they utilize the existing data in the database, such as class number, organization names, area and time, and ignore the important information contained in the text field of patent documents.

In conclusion, we can analyze the correspondence between enterprise and products and technologies through the correspondence analysis, obtain the corresponding superior products and technology, and then obtain the technical competitive advantage of enterprise.

2. Detail Description and Realization of Enterprise's Technical Competitive Advantage Analysis

Referring to a variety of correspondence analysis application examples, this text combines text mining algorithms and corresponding analysis methods, takes technology of patents profiles and product key words as an analysis factor, and takes patent application enterprise as another analysis factor, obtains enterprise's most competitive advantage product technology through correspondence analysis. The algorithm process is as follows.

Step one: Retrieve and Filter Patent Data

In view of the Chinese patent database, carry on the research in the field F for retrieval of relevant technology in order to obtain the field collection D, represented by $D = \{d_1, d_2, ..., d_n\}$, and establish a Chinese patent subject database.

Step two: Select Products and Technology Keywords

For the Chinese patent text document collections D obtained by the first step, select products and technology keywords. Selection method is mainly divided into the following main steps.

(1)Preliminary processing feature vector

1) According to pre-established Chinese patent participle stop using dictionary, and Chinese patent term lexical rules by which words are formed, and Stop using rules, and Ambiguity using rules, process the text segmentation and mark the rules of the property of a certain word.

2) Process word frequency statistics for processed key items, and then calculate feature

weighting according to formula TF-IDF, as show below.

$$W(t,\vec{d}) = \frac{tf(t,\vec{d}) \times \log(N / n_t + 0.01)}{\sqrt{\sum_{t\in\vec{d}} \left[tf(t,\vec{d}) \times \log(N / n_t + 0.01)\right]^2}} \quad \textbf{(1)}$$

Among them, $W(t,\vec{d})$ presents weight of t in text $\vec{d}$, $tf(t,\vec{d})$ presents word frequency of *t* in text $\vec{d}$, *N* presents the totality of training text, n_t presents text number of training text sets which include *t*, the denominator presents the normalization factor.

3) Build matrix representation of documents/characteristic item, stored to the two-dimensional tables of database, then form matrix X of original documents/characteristic items, represented by X=$[w_{ij}]_{m \times n}$, is shown as below.

Table 1. Matrix X of Original Documents/Characteristic Items

Documents Characteristic items	D_1	D_2	……	D_n
T_1	W_{11}	W_{12}	……	W_{1n}
T_2	W_{21}	W_{22}	……	W_{2n}
……	……	……	……	……
T_m	W_{m1}	W_{m2}	……	W_{mn}

4) Characteristic item T_i of original documents/characteristic items matrix X is stored in candidate term set *C*.

5) According to the Zipf, feature words which are less than 3 times or more than 1000 times are removed from candidate set *C*, form original documents/characteristic items matrix X' .

(2)Process the semantic concept space- extract technological terms

1) Extract technological terms according to C-Value

C-Value[12] is proposed by Frantzi & Anaiadou, is the statistical method of extracting parameter C-Value by new terms, As show below.

$$C-Value(a) = \begin{cases} \log_2 |a|.f(a) & if\ a\ not\ in\ any\ candicate\ term \\ \log_2 |a|.(f(a) - \frac{1}{c(a)}\sum_{i=1}^{c(a)} f(b_i)) & if\ a\ in\ other\ candicate\ terms \end{cases} \quad \textbf{(2)}$$

Frantz's C-Value mainly considers the relationship between simple terms and complex terms, between terms and terms context. It got an excellent result in experiment and application.

We consult the methods of Chen [13], then calculate the weight of characteristic items of technological terms using TF-IDF and C-Value. We consider not only terms' measure value, but also general importance measure of a word in corpus, as show below.

$$CPCV - IDF = C - Value \times IDF = C - Value(a) \times \log(N / n_t + 0.01) \quad \textbf{(3)}$$

$CPCV - IDF$ is the combination of C-Value and IDF.

Calculate the weight of TF-IDF and C-Value and CPCV-IDF at the same time. Calculate the weight of TF-IDF by formula (1), calculate C-Value by formula (2), calculate CPCV-IDF by formula (3). CPCV-IDF and C-Value are evaluated with respect to a collection of documents.TF-IDF is evaluated with respect to a single document of Chinese patent.

2) After extracting terms, program will automatically remove the words which obviously do not conform to the semantic rules by the rules of word-formation of Chinese patent terms.

3) After above extracting terms, CPCV - IDF term measure values calculated using the formula (3) are inverted. According to the number of patent documents in Chinese text mining, manually set the threshold value of measure values of a term candidate word of science and technology term every time, then filter out candidates which are less than the threshold in the candidate term list of science and technology term. According to the needs of analysis threshold can be set to 3, 5 or 10.

4) Draw the candidate term list of science and technology, the program automatically removes parts of candidate terms which are nested by longer candidate terms, and then fake term and obviously not technology term words are removed by artificial. At the same time, must carry on post-processing short term embedded in long term of the candidate term list of science and technology, then short terms which are below a certain threshold value and nested in the long term will be filtered, their weights will be recalculated. So-called filtering threshold is the maximum value of the weight of weight of the short term and long term when post-processing of filtering short words. If short words' threshold is below filtering threshold, the short words will be filtered, otherwise be kept.

(3) Process the semantic concept space- process the theme concept dictionary

Use the Chinese patent theme concept dictionary, including the synonym dictionary and the inclusion dictionary, the terms of corresponding semantic meaning will be mapped to the same subjects and keywords, corresponding subjects and keywords will become the characteristic items of Chinese patent document's feature vector, then the characteristic items is mapped to semantic concept space.

An example of the Chinese patent theme concept dictionary is shown below Table 2.

Table 2. An Example of the Chinese Patent Theme Concept Dictionary Main Dictionary

Characteristic items	Synonym dictionary	Inclusion dictionary	……
Fiber-optic network	1	0	……
Wavelength division multiplexing	1	0	
Teacher	1	1	
……	……	……	……

Synonym Dictionary

Characteristic items	Synonym word 1	Synonym word 2	……
Fiber-optic network	optical networks	Networks of Fiber Communications	……
Wavelength division multiplexing	Wave Division Multiplexing	WDM	……
Teacher	Schoolmaster	Master	……
……	……	……	……

Inclusion Dictionary

Characteristic items	Inclusion word 1	Inclusion word 2	……
School	University	Middle school	……
Teacher	Professor	Instructor	……
……	……	……	……

According to semantic concept dictionary and requirement of text mining, filter relevant synonym words and inclusion words. Recalculate the term measure value of CPCV-IDF and the weight of TF-IDF. Finally obtain the key products of this field and the list of properties keywords.

Step three: Form the Standardized Results Matrix

We select main enterprises those apply for patents in the domain from Chinese patents data obtained in step one, take enterprises in the field as an analytic factor, take products and attribute keywords in the field as another analytic factor. Based on the data relationships of applicants and products or technology keywords in Chinese patent data, we can establish a cross summary table of column data and row data, then standardize the data in the table, finally we can get the summary results matrix.

Step four: Form the 2-dimensinoal Factor Scattergram based on the Correspondence Analysis Algorithm

The correspondence analysis algorithm is to change the cross-tabulation which consists of column variable and row variable into a scatter diagram, and then display related category information in the table in the form of the spatial location relationship of splashes.

The algorithm uses a data conversion shown as below.

$$z_{ij} = \frac{x_{ij} - x_{i.}x_{.j} / \sum_{i=1}^{n}\sum_{j=1}^{m} x_{ij}}{\sqrt{x_{i.}x_{.j}}} (i = 1, 2...n; j = 1, 2...m) \quad \textbf{(4)}$$

It changes $X = (x)_{n\times m}$ the original data matrix which has n samples and m variables into another matrix $Z = (z)_{n\times m}$, analyzes the covariance matrix of correlation between variables with $R = Z'Z$ and the covariance matrix of correlation between samples with $Q = ZZ'$,in addition, R and Q have the same nonzero characteristic root $\lambda_1 \geq \lambda_2 \geq ... \geq \lambda_p$, their corresponding eigenvector is U_i shown as below.

$$U_i = \left(u_{1i}, u_{2i}...u_{ni}\right)', \; V_i = \left(v_{1i}, v_{2i}...v_{ni}\right)' \quad \textbf{(5)}$$

Due to the particularity of transformation, they have close relations.

Analyze the two covariance matrixes R and Q, and then extract the two most important common factors R_1、R_2 and Q_1、Q_2 separately. As the particularity of transformation, R_1 and Q_1 are essentially equivalent, R_2 and Q_2 are essentially equivalent too. So use dimension 1 as the unified logo of R_1 and $Q_{1,}$ and use dimension 2 as the unified logo of R_2 and $Q_{2.}$ The loads of column variable and row variable in dimension 1 and dimension 2 are shown as below.

$$\begin{pmatrix} u_{11}\sqrt{\lambda_1} & u_{12}\sqrt{\lambda_2} \\ u_{21}\sqrt{\lambda_1} & u_{22}\sqrt{\lambda_2} \\ \vdots & \vdots \\ u_{n1}\sqrt{\lambda_1} & u_{n2}\sqrt{\lambda_2} \end{pmatrix} \begin{pmatrix} v_{11}\sqrt{\lambda_1} & v_{12}\sqrt{\lambda_2} \\ v_{21}\sqrt{\lambda_1} & v_{22}\sqrt{\lambda_2} \\ \vdots & \vdots \\ v_{n1}\sqrt{\lambda_1} & v_{n2}\sqrt{\lambda_2} \end{pmatrix} \quad \textbf{(6)}$$

Then form the 2-dimensinoal factor scattergram in the same coordinate system, with the figure it is easy to study the relationship between the variables and the samples.

Step five: Analyze the Corresponding Analysis Figure and Find out the Enterprise's Corresponding Products Advantages and Technology Advantages

The analytical methods mainly judge advantages by the distance of points. Distance of points means the difference of variables, the distance of variables which have close relationship is shorter. According to distance between enterprises' corresponding points and productive technology keywords, judge the enterprises' corresponding products advantages and technology advantages.

3. Empirical Research

To the State Intellectual Property Office patent database (www.sipo.gov.cn) as the data source, create the Chinese patent database about communication technology. Retrieves the type of optical communication in Chinese patent, limit time for 1985.1.1-2007.1.31.Ritrieved data is shown below Table 3.

Table 3. Table of Optical Communication in Chinese Patent

	Invention	Utility model	Total
Unauthorized	386	0	386
Authorized	140	70	210
Total	526	70	596

We use the 210 patents in Chinese patent data of optical communications technology in the empirical study. Take the top twenty applicants order by the number of patent applications as analysis objects, analysis enterprise's technology competent advantage in the field of optical communication with the correspondence analysis described in this section. Use Figure 1 as an example, this figure is the correspondence analysis figure of enterprise's productive technology competent advantage in the field of optical communication, made with the correspondence analysis described in this section.

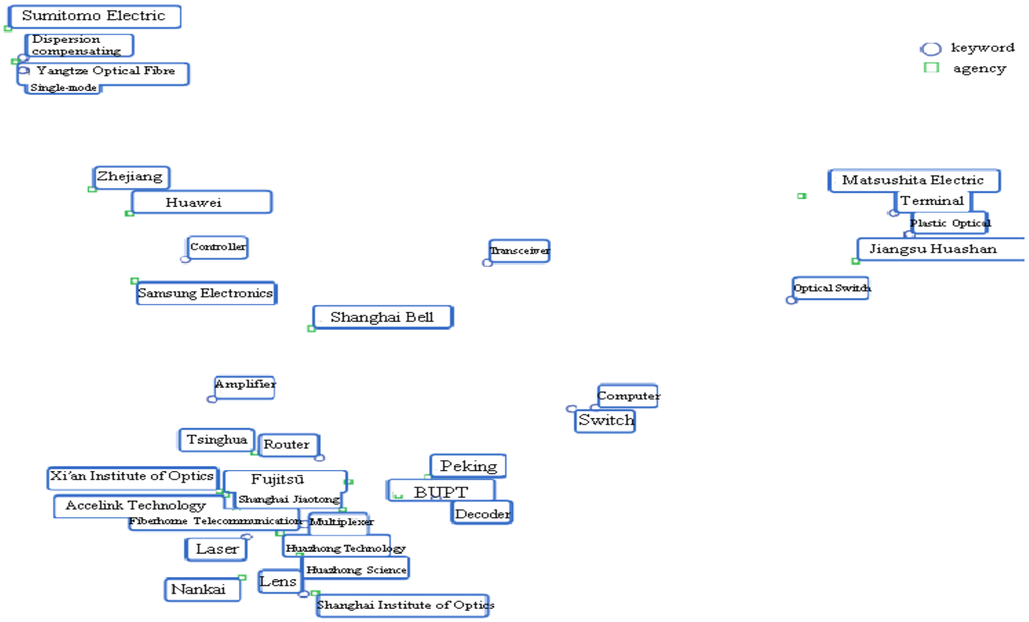


Figure 1. The Correspondence Analysis Figure of Enterprises' Productive Technology Competent Advantage in the Field of Optical Communication

As shown in Figure 1, the results as follows:

(1)Sumitomo Electric Industries and Yangtze Optical Fibre and Cable Company Ltd in the upper left corner have technology competent advantage in "dispersion compensating fiber" and "single-mode fiber".

(2)Matsushita Electric Industrial and Jiangsu Huashan photoelectric co., LTD in the upper right corner have technology competent advantage in "terminal equipment" and "plastic Optical Fiber".

(3) Huawei Technologies Co., Zhejiang University and Samsung Electronics Co., Ltd in the middle have technology competent advantage in "controller".

(4)In addition, other enterprises and product keywords distribute in the lower left corner densely, represented by Shanghai Jiaotong University、 Tsinghua University and Fujitsū Kabushiki-gaisha and so on, they have intense competition with each other in optical communication products represented by "multiplexer", "router", "laser", "lens" and so on.

4. Conclusion

This article takes 210 authorized patents in optical communication and the top twenty applicants as analytic target, use the method which combines text mining algorithms and corresponding analysis to analyze enterprise technology competent advantage, obtain the correspondence analysis figure. Compare with corresponding analysis, the result coming from the method which combines text mining algorithms and corresponding analysis is more accurate.

References

[1] W. Shunyu, "SPSS Correspondence analysis application in foreign language learning requirement analysis", Chongqing University of Posts and Telecommunications journal (Social Sciences Edition), vol. 5, **(2004)**, pp. 153-155.

[2] D. Bartholomew, F. Steele, I. Moustaki and J. Galbraith, "The Ananlysis and Interpretation of M ultivariate Data for Social Scientists", London:Chapman&Hall, vol. 81, **(2002)**.

[3] G. Zhigang, "Social statistical analysis methods-SPSS software application", Beijing: China Renmin University Press, **(1999)**.

[4] X. Yu and X. Ren, "Multivariate statistical analysis", Beijing: China Statistics Press, **(1999)**.

[5] Y. Liu and F. Xie, "Applied research in patent information analysis in the field of flash memory technology", The Defense Technology Management Academic Council, **(2006)**, pp. 45-47.

[6] H. Xianping and Q. Zhou, "Correspondence analysis method in the application of the environmental pollution", Journal of Yibin University, vol. 12, no. 12, **(2012)**, pp. 36-38.

[7] Z. Yuxin and Z. Jinzong, "Comprehensive effectiveness evaluation of land-use in China based on correspondence analysis", Progress in Geography, vol. 29, no. 4, **(2010)**, pp. 478-482.

[8] G. Yunlong and L. Yuan, "Correspondence analysis in marketing research of segmentation for Cerato sedan-take the market in Jiangsu for example", Application of Statistics and Management, vol. 28, no. 4, **(2009)**, pp. 685-690.

[9] S. Bhattacharya, C. Pal and J. Arora, "Inside the frontier areas of research in physics: A micro level analysis", Scientometrics, vol. 47, no. 1, **(2000)**, pp. 131-142.

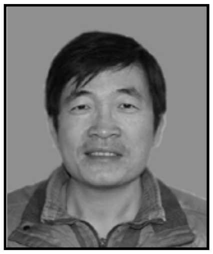
[10] J.-C. Dore, T. Ojasoo, Y. Okubo, T. Durand and G. Dudognon, "Correspondence factor analysis of the publication patterns of 48 nations over the period 1981-1992", Journal of the American Society for the Information of Science, vol. 47, no. 8, **(1996)**, pp. 588-602.

[11] J. Christophe, C. Dutheuil and J. Miquel, "Multidimensional analysis of trends in patent activity", Scientometrics, vol. 47, no. 3, **(2000)**, pp. 131-142.

[12] K. Frantzi, S. Ananiadou and H. Mima, "Automatic recognition of multi-word terms: the C-value NC-value method", International Journal of Digital Libraries, vol. 3, no. 2, **(2000)**, pp. 117-132.

[13] C. Shiji and W. Xiaomei, "The Topic Recognition Research of Paper Cluster by Combining C-value and TF-IDF", Journal of intelligence, vol. 28, no. 6, **(2009)**, pp. 821-826.

Authors



Xingang Wang, he received his M.Sc. in Software Engineering (2005) from Shandong University. Now he is associate professor of informatics at Information Department, Qilu University of Technology. Since 2011 he is AC Member of YOCSEF. His current research interests include different aspects of Data Mining and Information Integration.

Song Liu, he received his M.Sc. in Computer Software (2000) from Shandong University and PhD in Management Science and Engineering (2011) from Beijing Institute of Technology. Now he is associate professor of informatics at Information Department, Qilu University of Technology. Since 2012 he is AC Member of YOCSEF. His current research interests include different aspects of Data Mining and Knowledge Management.