Support Method for Reference of Documents based on Correspondence Analysis

Makoto SUZUKI , Tomohiro YOSHIKAWA , Takeshi FURUHASHI

Graduate School of Engineering, Nagoya University

Furo-cho, Chikusa-ku, Nagoya 466-8603, Japan

Abstract: The opportunity that we want to search target document(s) among a large amount of electronic documents such as WEB information and documents inputted by document scanner will be more increasing in near future. We focus on the method that applies the distribution pattern of the keyword frequency in documents to the classification of the documents. In this paper, we propose the method that uses the correspondence analysis for the classification of documents based on the similarity of the keyword pattern among them. We apply the proposed method to the data of patents of Nagoya University, and report the result of the accuracy for the classification. The experimental result shows that we can classify the related patents to the inputted keyword by a user.

Keywords: *Classification of Documents, Correspondence Analysis, Pattern of Keyword Frequency, Patent, Document Search*

I. INTRODUCTION

According to the spread of mobile devices and tablet PC, we have a lot of opportunity to search target document(s) among a large amount of electronic documents, for example WEB information and documents inputted by document scanner[1][2][3][4], which will be increasing more and more in near future. In the conventional document search by inputted keyword(s), we cannot search the target or related documents which do not include the keyword(s). Thus we have paid attention to the method that applies the distribution pattern of the keyword frequency in documents to the document search.

In this paper, we propose the method that uses the correspondence analysis[5][6][7][8] for the classification and the reference of documents. In this method, the distribution pattern of the keyword frequency is generated based on those of the documents including the inputted keyword, and the similarity of documents is calculated based on the similarity of the keyword pattern between them. We apply the proposed method to the data of patents of Nagoya University, and report the result of the document search. The experiment is done under the several conditions with the deletion and the addition of keywords. The experimental result shows that we can classify the related patents to the inputted keyword by a user.

II. CORRESPONDENCE ANALYSIS

The correspondence analysis is the analytical method that is originated by Jean-Paul Benzecri[5][6][7][8]. This chapter explains the algorithm of the correspondence analysis using the matrix data with $m*n$, in which each column represents the samples (the number of samples: m) and each row does the categories (the number of categories: n).

$$A=\begin{pmatrix} \sigma_{11} & \sigma_{12} & \cdots & \cdots & \sigma_{1n} \\ \sigma_{21} & \ddots & & & \vdots \\ \vdots & & \sigma_{ij} & & \vdots \\ \vdots & & & \ddots & \vdots \\ \sigma_{m1} & \cdots & \cdots & \cdots & \sigma_{mn} \end{pmatrix} \quad (1)$$

where $\sigma_{ij}=k$ (the number of responded times of sample i to category j)

$$B=\begin{pmatrix} b_1 & & & & 0 \\ & \ddots & & & \\ & & b_i & & \\ & & & \ddots & \\ 0 & & & & b_m \end{pmatrix} \quad (2)$$

$$C = \begin{pmatrix} c_1 & & & 0 \\ & \ddots & & \\ & & c_j & \\ & & & \ddots \\ 0 & & & c_n \end{pmatrix} \quad (3)$$

where $b_i = \sum_{j=1}^{n} \sigma_{ij}, c_j = \sum_{i=1}^{m} \sigma_{ij}$

A sample score and category score are denoted x and y, respectively. By maximizing the correlation coefficient between x and y under the conditions that the average is 0 and the variance is 1, the following Eigen equation is derived.

$$Hl = \lambda^2 l \quad (4)$$

where $H = B^{-\frac{1}{2}} A C^{-1} A^T B^{-\frac{1}{2}}$

We can obtain the Eigen values λ^2 and the corresponding Eigen vector $\boldsymbol{l}$, and we can calculate the sample score x. By using the Eigen vector $l^{(h)}$ corresponding to the h-th Eigen value $\lambda^{2\,(h)}$ excepting the condition that the Eigen value is 1, the h-th sample score $x^{(h)}$ is shown below.

$$x^{(h)} = B^{-\frac{1}{2}} l^{(h)} \times \lambda^{(h)} \quad (5)$$

In this paper, we analyzed the data multiplying the weight $\lambda^{(h)}$ to each axis of the h-th category score $x^{(h)}$ in the score space.

III. PROPOSED METHOD

A. *Calculation of Synthesized Score*

We obtain the synthesized score x_r when a user is interested in the category r using the values of sample score aquired by the correspondence analysis. By adding the values of sample score of the data in which the number of category r are more than zero, the value of the synthesized score x_r along the h-th axis is calculated as shown below.

$$x_r^{(h)} = \sum_{i=1}^{m} x_i^{(h)} \quad \{i | \sigma_{ir} > 0\} \quad (6)$$

It is thought that the synthesized score denote the major direction of score in category r.

B. *Evaluation of Similarity between Synthesized Score and Sample*

The similarity between the synthesized score x_r in eq. (6) and each sample is calculated as shown below.

$$Sim(x_i, x_r) = \frac{x_i \cdot x_r}{|x_i||x_r|} \quad (7)$$

We can consider that the sample which has larger value of cosine similarity is closer to the synthesized score in the score space. Thus this method sorts the samples by the cosine similarity, and we can obtain the samples (documents) related to the category r which the user is interested in.

IV. APPLICATION TO PATENT DATA

A. *Patent Data*

This paper studied the validity of the proposed method by applying it to the patent data of Nagoya University applied from December 28, 2002 to May 25, 2010. The total number of patents was 1000. In this experiment, the summary area of each patent was used and the noise terms surrounded by "[]" such as "[purpose]" and "[effect]" were deleted .

B. *Extraction of Category Data*

Nouns were extracted from the summary area of each patent using the morphological analysis tool Chasen[9]. Though many extracted nouns were common words which did not refrect the contents of the patent such as "thing" and "this invention," they were not removed here. This experiment used the top 2000 nouns in the frequency of appearance.

C. *Extraction of Patents based on "Gene"*

We extracted patents related to "gene" from the 1000 patents using the proposed method. "Gene," "foreign gene" and "manifestation of gene" were grouped as "gene," and there were 59 patents with these categories in 1000 patents.

The result of the extracted patents sorted by the cosine similarity is shown in Table 1. The orange colored patents contained "gene" in the summary area. In Table 1, the patents that contained the keyword, "gene," are ranked in the higher level with high cosine similarity. It also shows "JP 2005-516799" as the yellow colored 56th patent even without "gene" in the summary. Actually, this patent contains "gene" in the title, which shows the proposed method can extract the patents related to the keyword even it is not directly contained.

D. *Evaluation of Potential of Proposed Method*

The result of the correspondence analysis is not affected by only the specific category information (in this case,

whether each patent has "gene" or not in the summary area) but the overall distribution of categories (the pattern of the appearance frequency of all nouns). This feature could give the stable extraction of documents comparing with the simple keyword search. Thus the stability of the proposed method was evaluated with the following procedure. To compare the proposed method with the conventional method, "*F* measure" was used for the evaluation.

E is the number of the correct patents in the extracted ones, *N* is the number of all extracted patents and *M* is the number of the all correct patents. Using the precision *P* and the recall *R*, *F* measure is calculated as shown below.

$$F = \frac{2PR}{P+R}$$
$$P = \frac{E}{N}, \quad R = \frac{E}{M} \tag{8}$$

We evaluated the proposed method in the next three conditions. We deleted or added the keyword from some of documents to evaluate the extraction performance for the related documents. In all conditions, the number of the correct documents was 59.

Condition1: Same as Table 1.

Condition2: The keywords, "gene," were deleted from randomly selected 29 patents in 59 patents containing "gene". Then the synthesized score was generated with the remaining 30 patents.

Conditon3: The keyword, "gene," was added into randomly selected 51 patents other than the 59 patents. Then the synthesized score was generated with the total 110 patents.

The results of condition1 - condition3 are shown in Fig.1(a) – (c), respectively. In these figures, the horizontal axis denotes the number of the extracted documents and the vertical axis does *F* measure. And the value of the conventional method, "keyword search," is shown by the expectation value. In Fig.1, the proposed method was equivalent to keyword search in F measure. On the other hand, the proposed method was superior to keyword search in Fig.1(b) and Fig.1(c). These results show the potential of the proposed method in the extraction of documents by the correlation of the keyword pattern.

VI. CONCLUSION

This paper proposed the extraction method of documents which a user wanted to search based on a keyword using the correspondence analysis. In this method, the synthesized score is calculated based on the distribution pattern of the keyword frequency of the documents including the inputted keyword, and the similarity of documents is calculated based on the cosine similarity between them. This paper applied the proposed method to the patents of Nagoya University, and reported the result of the extraction of documents. The experiment was done under the several conditions with the deletion and the addition of keywords, and it showed that the proposed method was superior to the conventional keyword search method in every condition.

Reference

[1] Y. Uchida, T. Yoshikawa, T. Furuhashi, E. Hirao and H. Iguchi, "Visualization of Time Series Variation of Web User Review based on Evaluation Keywords," Fuzzy Theory and Intelligent Informatics,Vol.22 No.3, 2010, pp.377-389

[2] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent Dirichlet allocation," Journal of Machine Learning Research,3, 2003, pp.993–1022

[3] V. Mauricio and P. Roberto: Dimensionality reduction by minimizing nearest-neighbor classification error, Pattern Recogn. Lett, Vol.32 No.4, 2011, pp.633-639

[4] Y. Miyazaki and I. Kohno: Document Navigation Using relevant keywords for Surveying and Refining Search Results, IPSJ SIG Technical Reports, 2008, pp.7-12

[5] Benzecri, J.p., " L'Analyse des Donnees(Tome 2), L'Anakyse des Correspondences ", Dunod,1973

[6] M. O. Hill, Correspondence analysis: A neglected multivariate method, Applied Statistics, Vol.23, No.3, 1974, pp.340-354.

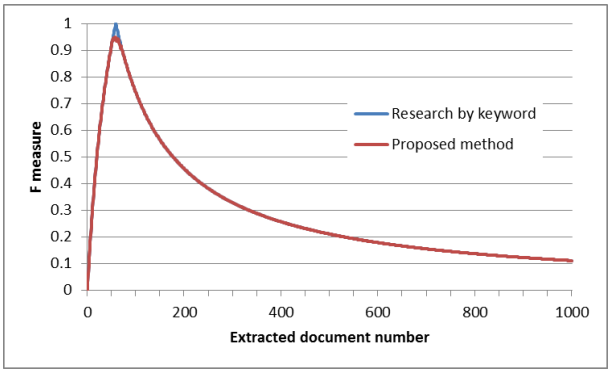
[7] C. Hayashi (1985). Analysis of multivariate data [IV]. The Journal of the Institute of Electronics and Communication Engineers, Vol.68, No.7, pp.779-786 (in Japan).

[8] Donna L. Hoffman & George R. Franke, Correspondence Analysis: Graphical Representation of Categorical Data in Marketing Research. Journal of Marketing Research, Vol.XXIII, 1986, pp.213-227.

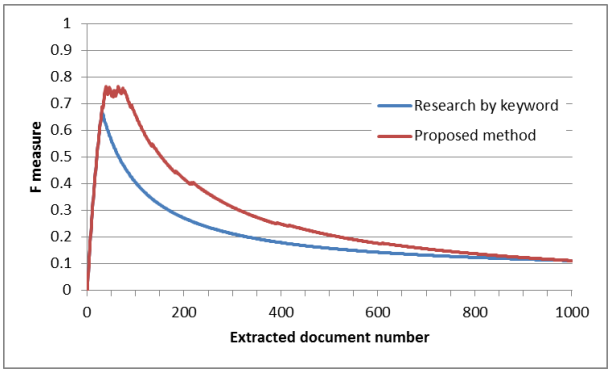
[9] http://chasen-legacy.sourceforge.jp/ (in Japanese)

Table 1　Extraction by “gene”

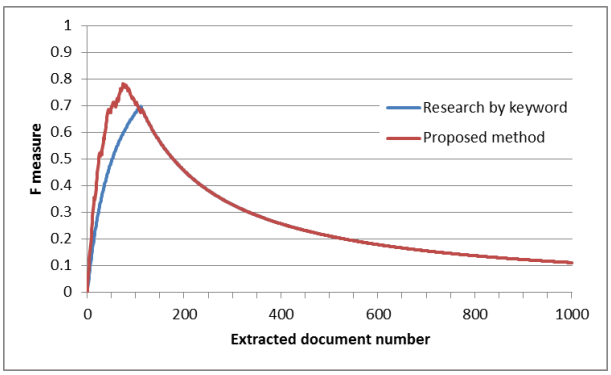
Order	Application Number	Cosine Similarity
1	JP 2007-544160	0.403
2	JP 2008-104962	0.398
3	JP 2010-54484	0.398
4	JP 2007-507135	0.388
5	JP 2008-522403	0.379
6	JP 2007-505883	0.367
7	JP 2007-198207	0.359
8	JP 2007-6588	0.354
9	JP 2007-521153	0.352
10	JP 2009-53481	0.349
11-49	・・・	・・・
50	JP 2008-539830	0.167
51	JP 2008-78421	0.166
52	JP 2006-266918	0.162
53	JP 2007-514646	0.152
54	JP 2007-61038	0.144
55	JP 2006-550765	0.139
56	JP 2005-516799	0.130
57	JP 2005-29809	0.128
58	JP 2006-328869	0.124
59	JP 2006-54414	0.119
60	JP 2005-149903	0.116
61	JP 2006-537712	0.110
62	JP 2005-379867	0.108
63	JP 2010-514521	0.094
64	JP 2009-48643	0.094
65	JP 2007-546426	0.088
66	JP 2007-520192	0.083
67	JP 2007-110320	0.083
68	JP 2008-502635	0.081
69	JP 2009-195728	0.078
70	JP 2008-542195	0.077



(a) Without deletion or addition of the keyword (Condition1)



(b) Deletion of the keyword (Condition2)



(c) Addition of the keyword (Condition3)

Figure 1. Evaluation of F Measure (“gene”)