Classification of Japanese Documents and Ranking of Representative Documents Using Characteristic of Frequencies of Words

Jun Kimura
JustSystems Corp., 6-8-1 Nishinjuku, Shinjuku-ku, Tokyo 163-6017

Yasunari Yoshitomi, and Masayoshi Tabuse
Graduate School of Life and Environmental Sciences, Kyoto Prefectural University,
1-5 Nakaragi-cho, Shimogamo, Sakyo-ku, Kyoto 606-8522, Japan
E-mail:{yoshitomi, tabuse}@kpu.ac.jp
http://www2.kpu.ac.jp/ningen/infsys/English_index.html

Abstract

We have developed a method for classification of Japanese documents and ranking of representative documents using characteristic of frequencies of nouns. The representative document is defined as the document whose feature vector is the closest to the center of gravity of the class in the feature vector space among all documents belonging to the class. The ranking of the representative documents is decided in the descending order of the number of documents belonging to the class.

Keywords: Document classification, Extraction of representative document, Clustering, and Frequency of nouns.

1. Introduction

Recently, Web pages on the Internet have been increasing, resulting in that it is very difficult to read through all of Web pages in which we are interested. However, as a fact, there are too many similar Web pages among them. For efficiently acquiring useful Web pages, it is necessary to select only Web pages having important and independent contents with which we can understand essential parts on an event adequately. A Web page has some kinds of media, such as document, image, and sound.

We focus on selecting Web page on the Internet according to characteristics of document on the page. Although the classification of documents has received considerable attention in document analysis research,[1-8] there is no research for selecting a representative document in a class of documents, followed by ranking several representative documents in order of importance or in any meaning useful for us, to our best knowledge.

In the present study, we have developed a method for classification of Japanese documents and ranking of representative documents using characteristic of frequencies of nouns.

2. Proposed Method

2.1. *Extraction of nouns*

Firstly, all nouns in a document are extracted using MeCab[9] with which the document is resolved into several morphemes (Fig. 1).

```
インターネットを利用する。
インターネット　名詞,一般,*,*,*,*,インターネット,インターネット,インターネット
を　　　助詞,格助詞,一般,*,*,*,を,ヲ,ヲ
利用　　名詞,サ変接続,*,*,*,*,利用,リヨウ,リヨー
する　　動詞,自立,*,*,サ変・スル,基本形,する,スル,スル
。　　　記号,句点,*,*,*,*,。,。,。
EOS
```

Fig. 1. Output of MeCab.

2.2. *Connection of nouns having a meaning as a set*

Some nouns directly connecting each other are treated as one noun in the case that they have a meaning in assuming one noun. For example, 2014 and 年 in Japanese has a meaning as a set of 2014 年. 年 in Japanese means year in English.

2.3. *Addition of negative attribution*

When a sentence expresses a negative meaning with use of 'not', the extracted nouns in the sentence are treated as having a negative attribution. In other words, a noun can have either positive or negative attribution. The noun having a negative attribution is treated as being different from the noun with a positive attribution in making a feature vector for the document where the noun exists.

2.4. *Feature vector generation*

After every noun composing of only one of a hiragana, which is the rounded Japanese phonetic syllabary, or a katakana, which is the angular Japanese syllabary, or a symbol is erased, a feature vector having a relative frequency of each noun as each element is generated for each document. The relative frequency is defined as the ration of frequency of the noun to that of all nouns in the document except nouns erased using the above criterion.

2.5. *Document classification and extraction of representative one in each class*

For clustering, we use Ward method. The representative document is defined as the document whose feature vector is the closest to the center of gravity of the class in the feature vector space among all documents belonging to the class.

2.6. *Ranking of representative documents*

The first-rank document is defined as the document whose feature vector is the closest to the center of gravity of all documents in the feature vector space. In this case, the number of class is one. Afterward, the number of class is increased one by one, and then the ranking from the second-rank for the representative documents is decided in the descending order of the number of documents belonging to the class for each number of classes. The maximum number J of classes in the stepwise clustering is given beforehand. Though a document can be selected more than once in the ranking process, the only first selection for the document is accepted.

3. Calculation Environment

The development of system and the experiment for evaluation of the proposed method were performed in the following environment for computation: personal computer; DELL OPTIPLEX 780(CPU: Intel Core2 Duo CPU E8400 3.00GHz, RAM: 4.00GB), OS; Microsoft Windows 7 Professional, Development language; Python 2.7.3.

4. Experiments and Discussion

4.1. *Document classification*

Firstly, we evaluated the performance of document calcification by the proposed method. We gathered 20 documents on politics (document nos. 1-10) and horse racing (document nos. 11-20) from Yahoo! Japan News[10] in January 2013, and then the number of clusters was set to be two, resulting in that our system gave the outputs shown in Table. 1. The clusters of C_1 and C_2 were composed of the documents on politics and horse racing, respectively (Table 1). Accordingly, the document calcification by the proposed method was perfect.

Table 1. Nos. of documents belonging to each cluster.

Cluster C_1	Cluster C_2
1, 2, 3, 4, 5, 6, 7, 8, 9, 10	11, 12, 13, 14, 15, 16, 17, 18, 19, 20

4.2. *Extraction of representative documents*

4.2.1. *Experiment I*

Next, we evaluated the performance of extraction of representative document by the proposed method. We gathered top 20 documents obtained by a retrieval from Google News[11] and those from Yahoo! Japan News using the retrieval keyword of '大阪府 高校' in Japanese, which is Osaka Prefecture high school in English, on 22 January 2013. The name of document obtained was set to be the same as the rank by each retrieval, and then all documents were categorized.

The name of category was decided to be the content name when more than two documents having the similar content each other existed, and otherwise the document was assigned to be a category of 'Others'. The categorization was manually performed through our understanding for each document, while the clustering was performed by the proposed method. Therefore, it was not guaranteed for the clustering result to correspond with the document group structure given by the manual categorization.

(a) Google News

Table 2 shows the document group structure when we used Google News in our experiment. There were five kinds of categories (Table. 2). Table 3 shows the ranking of representative documents given by the proposed method when using $J = 4$.

Table 2. Document group structure I.

Category	Rugby	Board of education	
Document No.	1, 9, 12, 18, 20	2, 4, 5, 11, 16, 17, 19	
Category	Skating	Distress accident	Others
Document No.	3, 14	6, 7, 8, 15	10, 13

Table 3. Result I.

Ranking of representative documents expressed by Nos.
6, 1, 4, 3

The four representative documents were successfully extracted one by one from all categories except the category of 'Others' in the order of 'Distress accident', 'Rugby', 'Board of education', and 'Skating' (Table 3).

(b) Yahoo! Japan News

Table 4 shows the document group structure when we used Yahoo! Japan News in our experiment. There were five kinds of categories (Table. 4). Table 5 shows the ranking of representative documents given by the proposed method when using $J = 4$.

The four representative documents were successfully extracted one by one from all categories except the category of 'Rugby' in the order of 'Board of education', 'Center exam', 'Distress accident', and 'Others' (Table 5).

Table 4. Document group structure II.

Category	Rugby	Board of education	
Document No.	15, 19	2, 3, 4, 11, 14, 16, 17, 18, 20	
Category	Distress accident	Center exam.	Others
Document No.	8, 12, 13	9, 10	1, 5, 6, 7

Table 5. Result II.

Ranking of representative documents expressed by Nos.
18, 10, 8, 5

4.2.2. *Experiment II*

We gathered top 20 documents obtained by a retrieval from Google News and those from Yahoo! Japan News using the retrieval keyword of 'Microsoft' on 22 January 2013. The name of document obtained was set to be the same as the rank by each retrieval, and then all documents were categorized in the same manner as those in the section 4.2.1.

(a) Google News

Table 6 shows the document group structure when we used Google News in our experiment. There were four kinds of categories (Table. 6). Table 7 shows the ranking of representative documents given by the proposed method when using $J = 4$. The six representative documents were extracted from all categories in the order of 'Others', 'Others', 'Windows 8', 'MS Essentials', 'Surface' and 'Others' (Table 7).

Table 6. Document group structure III.

Category	Windows 8	MS Essentials
Document No.	3, 5, 14	2, 6, 12
Category	Surface	Others
Document No.	9, 11, 15	1, 4, 7, 8, 10, 13, 16, 17, 18, 19, 20

Table 7. Result III.

Ranking of representative documents expressed by Nos.
18, 4, 3, 6, 11, 20

(b) Yahoo! Japan News

There were two kinds of categories (Table. 8). Table 9 shows the ranking for the document group shown in Table 8 by the proposed method.

Table 8. Document group structure IV.

Category	Cannon ITS	Others
Document No.	6, 9, 14	1, 2, 3, 4, 5, 7, 8, 10, 11, 12, 13, 15, 16, 17, 18, 19, 20

Table 9. Result IV.

Ranking of representative documents expressed by Nos.
2, 9, 15, 13, 20, 11

The six documents were extracted in the order of categories of 'Others', 'Cannon ITS', and four sets of 'Others' (Table. 9). In this document group, almost all documents belonged to the category of 'Others'. However, one document was extracted from the category of 'Cannon ITS' in the second order.

4.3. *Discussion*

4.3.1. *Document group structure dependency*

When the document group had a distinct structure such as that in the sections 4.2.1 (a) & (b) and 4.2.2 (a), the performance of the proposed method was almost sound in the meaning that the documents can be extracted one by one from all categories except 'Others'. On the other hands, when the document group had a scattered structure such as that in the section 4.2.2 (b), it might not be meaningful to try to cover almost all contents by extracting the representative documents using the proposed method.

4.3.2. *Performance improvement*

It might be necessary to apply the proposed method to many document-groups for finding out assignments of the proposed method. It might be effective to use a thesaurus for reducing the dimension of feature vector space, potentially resulting in extracting more appropriately representative documents and/or reducing the calculation cost.

4.3.3. *Definition of representative document*

In the present study, the representative document was geometrically defined in the feature vector space. It is necessary to investigate the validity of the definition through questionnaires. In the investigation, other definitions on the representative document might be on the discussion.

5. Conclusion

We have developed a method for classification of Japanese documents and ranking of representative documents using characteristic of frequencies of nouns. The experiments where Web pages were collected and used for evaluating the efficiency of the proposed method proved the usefulness of the proposed method.

References

1. F. Can and E. A. Ozkarahan, Computation of term/document discrimination values by use of the cover coefficient concept, *J. Amer. Soc. for Inf. Sci.*, **38**(3) (1987) 171-183.
2. A. Kawai, An automatic document classification method based on a semantic category frequency analysis (in Japanese), *J. IPSJ* **33**(9) (1992) 1114-1122.
3. N. Yuasa, T. Ueda, and F. Togawa, Classifying articles using lexical co-occurrence in large document databases (in Japanese), *J. IPSJ* **36**(8) (1995) 1819-1827.
4. K. Hatano, R. Sano, Y. Duan, and K. Tanaka, A classification view mechanism for web documents based on self-organizing maps and search engines (in Japanese), *J. IPSJ* **40**(SIG 3(TOD1)) (1999) 47-59.
5. H. Takamura and Y. Matsumoto, Co-clustering for text categorization (in Japanese), *J. IPSJ* **44**(2) (2003) 443-450.
6. H. Takamura and Y. Matsumoto, Constructive induction and text categorization with SVMs (in Japanese), *J. IPSJ* **44**(SIG 3(TOD 17)) (2003) 1-10.
7. R. Kusaya, T. Yamamura, H. Kudo, T. Matsumoto, Y. Takeuchi, and N. Ohnishi, Measuring similarity between documents using term frequency (in Japanese), *Trans IEICE* **J87-D-Ⅱ**(2) (2004) 661-672.
8. Y. Bamba, K. Shinzato, T. Shibata, and S. Kurohashi, Web information observation using keyword distillation based clustering (in Japanese), *J. IPSJ* **50**(4) (2009) 1399-1409.
9. MeCab, http://mecab.sourceforge.net/
10. Yahoo! JAPAN News, http://headlines.yahoo.co.jp/hl
11. Google News, https://news.google.co.jp/