A Text Mining System DIREC: Discovering Relationships between Keywords by Filtering, Extracting and Clustering

MINE, Tsunenori   LU, Shimiao   AMAMIYA, Makoto
Faculty of Information Science and Electrical Engineering,
Kyushu University
6-1 Kasuga-koen, Kasuga, Fukuoka, Japan, 816-8580
{mine, shimiao, amamiya}@al.is.kyushu-u.ac.jp

Abstract. This paper presents a text mining system DIREC that obtains the relationships between keywords. After gathering Web pages related to a query with search engines and filtering out the pages irrelevant to the query, DIREC extracts prespecified keywords from the pages left after filtering, and clusters them. Each cluster includes keywords related to each other. The clustered keywords are shown through an Explorer-like graphical user interface. We discuss the experimental results applying the DIREC system to the Research and Education Report Database of the Faculty at Kyushu University.

1 Introduction

The rapid spread of the Internet has brought about a big revolution in information technologies and information environments. The development of the World Wide Web (WWW) especially makes available a lot of knowledge and useful means for accessing electronic information entities. In these circumstances, Web mining or text mining to discover new knowledge from the vast number of Web pages has become more and more attractive. Although a lot of work for obtaining topics or keywords from documents have been performed, there are only few work for finding the inter-relationships between the extracted topics or keywords.

This paper presents a text mining system DIREC(DIscovering Relationships between keywords by filtering, Extracting and Clustering). The DIREC system, first, gathers Web pages related to a query with search engines and filters out the pages irrelevant to the query with support vector machines(SVMs)[5, 1]. Next, it extracts the keywords specified by a user and clusters them. Each cluster embodies the relationships between keywords and between a keyword and other named entities such as persons. It shows the clustered keywords through an Explorer-like graphical user interface. In order to evaluate the performance of the DIREC system, we applied it to the two sets: call for paper(CFP) files of international conferences[3] and the Research and Education Report Database of the Faculty at Kyushu University[4]. The former set is used for evaluating DIREC's following three functions : filtering, extracting and clustering. The later is for evaluating the robustness of the clustering function with a lot of noisy keywords left after morphological analysis and for trying other document set than CFP files. The results confirmed that the DIREC system worked well for obtaining the relationships between topics of conferences, and for discovering the relationships between research keywords.

In what follows, section 2 describes an overview of the DIREC system and section 3 discusses DIREC's clustering performance only for the research keywords due to the limitation of space.

2 The DIREC System

The DIREC system consists of 5 modules: the collecting file module, the filtering file module, the extracting information entity module, the clustering information entity module and the user interface module. This section describes only the clustering information entity module to explain the experimental results shown in section 3. For the other functions, please see [3].

2.1 Clustering Keywords

The clustering procedure consists of the following 3 steps:

1. Calculating the similarity between every pair of keywords.
2. Creating base clusters based on the results of the similarity calculation in the 1st step.

   (The base cluster will be defined at the next section.)
3. Combining base clusters whose similarity value is over the pre-determined threshold.

2.2 Calculating the Similarity between Keywords

The similarity between keywords is calculated by the proportion of their inclusion in a set of files where each of the keywords appears.

Let T_m be the set of files where keyword m appears, and $|T_m|$ be the number of files included in T_m. $|T_m \cap T_n|$ represents the number of files T_m and T_n have in common.

We set the similarity between two keywords to 1 if one of the following conditions is satisfied and 0 otherwise.

$$|T_m \cap T_n| / |T_m| > TH \qquad (1)$$

$$|T_m \cap T_n| / |T_n| > TH \qquad (2)$$

The values of TH is decided empirically. Calculating the similarity between every pair of keywords, we make a cluster that includes a keyword and its related keywords, each of whose similarity with the keyword is 1. We call such cluster a **base cluster**.

This method is basically the same as Zamir and Etzioni's method for calculating the similarity between base clusters[6].

3 Discussion of Clustering Results to the Faculty Members' Research Keywords

Research keywords of the faculty were extracted from the Research and Education Report Database of the Faculty at Kyushu University(http://www.ofc.kyushu-u.ac.jp/kyokandb/)[4]. We extracted keywords directly from this database without filtering and extraction phrase. Instead of that, we applied Japanese morphological analyzer 'Chasen'[2] to find keywords matched partially with one another because most of them consist of Japanese compound nouns that had no space delimiter. For example, if the following three keywords: MARUCHI / EJENT (Multi Agent), MARUCHI / EJENT / SHISUTEMU (Multi Agent System), MOBAIRU / EJENT (Mobile Agent) occurred, they should be included in the same cluster because of the common word EJENT(Agent).

After applying morphological analysis to 10972 original research keyword types that came from 1937 faculty members, we obtained 17366 keyword types including original ones. Among them, we empirically selected only the keywords whose number of occurrences in the faculty members' research reports was more than 1 and less than 16 because most of the keywords frequently appeared in the reports were general words. With these keywords, we evaluated both the change of the number of isolated keywords that did not belong to any base clusters and that of the number of clusters created according to the change of the threashold TH_1 and TH_2. Where TH_1 and TH_2 are the threashold value for the proportion of a set of files where each of keywords appears and that of keywords that were shared among base clusters, respectively. From the evaluations, we aimed to confirm whether or not TH in section 2.2, which is here TH_1 , depended on the target documents and to investigate the change of the number of clusters according to TH_2.

Considering the TH_1 through 0.4 to 0.7, the number of clusters of both 0.5 and 0.6 was almost the same, and when TH_1 was 0.7, the number of isolated keywords became more 200 than the case that TH_1 was 0.6. From these results, TH_1 should be more than and equal to 0.5. This result was the same as the clustering topics of international conferences[3]. On the contrary, determining the optimum value of TH_2 seemed to be more difficult. It might be done by making use of the contents of the clusters. Although we only checked cluster 'Artificial Intelligence'[1], the keywords belonging to the cluster almost seemed to match the name of the cluster. Table 1 shows an example of base clusters with keyword **agent**.

4 Conclusion and Future Work

This paper discussed the DIREC system that obtained the relationships between information entities by filtering files, extracting and clustering keywords. We performed experiments with the Research and Education Report Database of the Faculty at Kyushu University. The main objectives of the experiments were to investigate the change of the number of isolated topics that did not belong to any base cluster according to the threshold for the similarity between keywords, and the change of the number of clusters according to the proportion of the common keywords among base clusters. The experimental results showed that the threshold value to the similarity between keywords was almost the same as the case of CFP files[3]. On the contrary, the appropriate threshold for combining base clusters should be investigated furthermore. We are doing further experiments with huge data sets and also investigating how to calculate the similarity between keywords.

[1] We employed the keyword occurring most frequently in a cluster as the name of the cluster

Table 1: An Example of Base Clusters including Keyword ’Agent’ when $TH_1 = 6$

Keword	Related Keywords	Faculty Members
Agent	Intelligence, Intelligent, Multi, Parallel Distributed, Discovery Science, Information Retrieval, Autonomous Distributed, Flow, Architecture, Massively Parallel, Knowledge Acquisition, Dialogue System, Control System, Education System, Reinforcement Learning, Inductive Reasoning, Machine Discovery, Machine Learning, Self-teaching Support, Soft Computing	1779, 1374, 377, 276
Machine Learning	Algorithm, Logic, Intelligence, Probability, Graph, Mining, Data Mining, Discovery, Complexity, Reasoning, Complexity Theory, Approximated, Genome Information, **Agent**, Discovery Science, Parallel Algorithm, Distributed Algorithm, Approximated Algorithm, Inductive Reasoning, Machine Discovery, Graph Algorithm	1413, 1374
Autonomous Distributed	Algorithm, Understanding, Intelligence, Intelligent, Robot, Graph, Vision, Distributed, Multi, Real Time, Discription, **Agent**, Parallel Distributed, Flow, Architecture, Parallel Algorithm, Distributed Algorithm, Massively Parallel, Dialogue System, Computational Geometry, Graph Algorithm, Algorithm Engineering	878, 277, 276

In the future, we will implement the DIREC system as a multi-agent system[7].

Acknowledgments

This research was partly supported by a Grant-in-Aid for Scientific Research, Encouragement of Young Scientists(A), No. 13780291 from the Japan Society for the Promotion of Science, and by a Grant for a R&D activities in the information and communications area from the Telecommunication Advancement Organization of Japan.

References

[1] Thorsten Joachims. Svm light: http://svmlight.joachims.org.

[2] Yuji Matsumoto, Akira Kitauchi, Tatsuo Yamashita, Yoshitaka Hirano, Hiroshi Matsuda, Kazuma Takaoka, and Masayuki Asahara. Morphological analysis system chasen version 2.2.8 manual. Technical report, Nara Institute of Science and Technology, 2001.

[3] Tsunenori Mine, Shimiao Lu, and Makoto Amamiya. Discovering relationship between topics of conferences by filtering, extracting and clustering. In *the 3rd International Workshop on Natural Language and Information Systems(NLIS2002)*, to appear, 2002.

[4] Yusuke Nonaka, Sozo Inoue, Katsuhiko Hatano, Tsutomu Harada, Yoshinari Nomura, Mizuho Iwaihara, Tsunenori Mine, and Kazuo Ushijima. Development and operation of a document database for university research and education activities. *Systems and Computers in Japan*, to appear, 2002.

[5] Vladimir N. Vapnik. *Statistical Learning Theory*. JOHN WILEY & SONS, INC., 1998.

[6] Oren Zamir and Oren Etzioni. Web document clustering : A feasiblity demonstration. In *Proceedings of the 21th Intl. ACM SIGIR Conference*, pages 46–54, 1998.

[7] Guoqiang Zhong, Satoshi Amamiya, Keníchi Takahashi, Tsunenori Mine, and Makoto Amamiya. The design and application of kodama system. *IEICE Transactions on Information and Systems*, E85-D(4):637–646, 4 2002.