Discovering Relationship between Topics of Conferences by Filtering, Extracting and Clustering

MINE, Tsunenori LU, Shimiao AMAMIYA, Makoto
Graduate School 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 that obtains the relationship between the topics of international conferences. The system at first gathers Web pages related to the query 'call for paper' with a search engine and filters out irrelevant pages to the query with SVM. Next it extracts the topics of conferences and clusters them. Each cluster holds the relationship between topics and between topics and conferences. The topics clustered are shown through an Explorer like graphical user interface. The preliminary experimental results promises that the method works not only for obtaining the relationship between topics of conferences, but also for discovering the relationship between information entities users are interested in.

1 Introduction

The Internet being spreading rapidly brings us a big revolution on our circumstances about information technologies and information environments. Spreading World Wide Web(WWW) especially gives us a lot of knowledge and useful accessing means to electronic information entities. In this circumstances, a Web mining or text mining to discover new knowledge from huge Web pages has become more and more attractive.

This paper presents DIREC(DIscovering Relationship between data by filtering, Extracting and Clustering), which is a text mining system that obtains the relationship between topics extracted from Web pages or E-mail messages of the call for papers(CFPs) of international conferences. DIREC system at first gathers Web pages related to the query 'call for paper' with a search engine and filters out irrelevant pages to the query with SVMs. Next it extracts the topics of conferences and clusters them. Each cluster holds the relationship between topics and between topics and conferences. The topics clustered are shown through an Explorer like graphical user interface. The preliminary experimental results show that DIREC system works not only for obtaining the relationship between topics of conferences, but also for discovering the relationship between information entities users are interested in.

In what follows, section 2 describes the overview of DIREC system and section 3 shows the functions of DIREC system. Section 4 discusses the experimental results.

2 The Overview of DIREC System

Figure 1 shows the overview of DIREC system. The system consists of 5 modules cascaded in the following order:

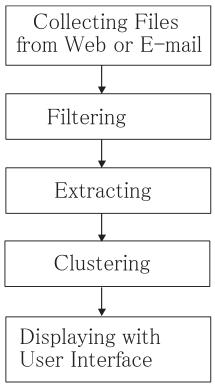

Figure 1: The Overview of DIREC System

1. **Collecting files**: This module collects files related to a query with search engines and by crawling web sites specified already.
2. **Filtering files**: This module filters out the files irrelevant to the category specified already from the files gathered by the collecting module, using Support Vector Machines(SVMs)[1, 2].
3. **Extracting information entities**: This module extracts information entities such as keywords or topics from the files selected by the filtering module.

4. **Clustering information entities**: This module clusters the information entities extracted by the extracting module to obtain the relationship between the entities extracted.
5. **Displaying results**: The obtained relationships between the entities at the clustering module are displayed through the graphical user interface that has the function of retrieving information entities.

3 Functions of DIREC system

3.1 Filtering CFPs Files with SVM

Since the files collected with search engines and received by e-mail often include a lot of irrelevant files, we need to filter them out. As a filtering method, we employ SVM[1, 2] because SVM should perform well for text categorization[3].

In order to apply SVM to a text categorization problem, text documents are transformed into text features vectors. Text features vector x_i of the ith document is defined as follows:

$$x_i = \{w_j | w_j = tf_{i,j} * log\frac{N}{n_j}, 1 \le j \le m\} \qquad (1)$$

Where w_j is the weight of the jth word that appears in the ith document, $tf_{i,j}$ is the number of occurrences that jth word appears in the ith document, N is the number of all documents in a collection, n_j is the number of documents that jth word appears in the collection and m is the number of word types in the ith document. If $n_j = 0$, then we take $\log\frac{N}{n_j} = 0$.

The ith training vector is defined as $\{\mathbf{x}_i, y_i\}$, $(i = 1, \ldots, l,\ y_i \in \{-1, 1\})$

Where $\mathbf{x}_i$ is the ith feature vector defined in the equation 1 and y_i is a scalar that if the ith document in a training documents set belongs to a target category, i.e., the ith document is positive data, y_i is set to 1, and otherwise, i.e., that is negative data, y_i is set to -1.

3.2 Extracting Topics of Conferences

It is useful to present only the keywords or phrases that users are interested in after extracting them from a lot of documents or data retrieved. As such keywords, we decide to extract the topics of conferences because one of the most important factors that users decide to submit their research papers to a conference seems to be whether or not the conference includes a topic they are interested in. The topics extraction consists of the following two steps:

1. Finding topics blocks which contain a set of topics
2. Extracting topics from the blocks

3.2.1 Topics Block Extraction

In order to extract a topic block, we assumed that topics are itemized and there exist some keywords such as ' topics', ' area', ' issue', ' aspect' and 'theme' that appear frequently at the previous of the block. We call such keywords as topic markers.

The topics itemization style depends on a file style. When itemized in an HTML file, a list tag `<li>` is attached at the head of topics line. When itemized in a text file, a mark shape appears there. Where the mark shape stands for a string being composed of one or more symbols such as space, '$*$', '$\bullet$', '$\circ$', or ' $-$'.

3.2.2 Topics Extraction

When extracting topics from a topic block, we used punctuation marks, conjunctions(e.g. 'and', 'or'), and prepositional particles(e.g. for', ' from') as a delimiter between topics. Then we eliminated stopped words from the set of topics extracted, and transformed their plural forms to singlular forms.

3.3 Clustering Topics

In order to obtain the relationship between topics, we cluster them. The clustering procedure consists of the following 3 steps:

1. Calculating the similarity between every two topics
2. Creating base clusters based on the results of the similarity calculation at the 1st step. The base cluster will be defined at the next section.
3. Combining base clusters with a common topic.

3.3.1 Similarity Calculation between Topics

The similarity between topics is calculated by the inclusive rate of a set of files where each of the topics appears.

Let T_m be the set of files where topic m appears, and $|T_m|$ be the number of files included in T_m. $|T_m \cap T_n|$ stands for the number of common files between T_m and T_n.

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

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

$$|T_m \cap T_n| / |T_n| > TH_n \qquad (3)$$

Both values of TH_m and TH_n are decided empirically. Calculating the similarity between every two topics, we make a cluster that includes a topic and its relative topics each of whose similarity with the topic is 1. We call such cluster **base cluster**.

This method is basically same as Zamir and Etzioni's method to calculate the similarity between base clusters[5].

3.3.2 Creating and Combining Base Clusters

First, we calculate the similarity values between every two topics and create base clusters based on the values. Next, we combine base clusters with one or more common topics.

Figure 2 shows a cluster composed of two base clusters that are connected with 3 common topics, each of which is connected with a dotted line. A small circle stands for a topic and a big circle containing several topics stands for a base cluster. The topic in the center of a base cluster is the central topic of the cluster and other topics are the relative topics of the topic.

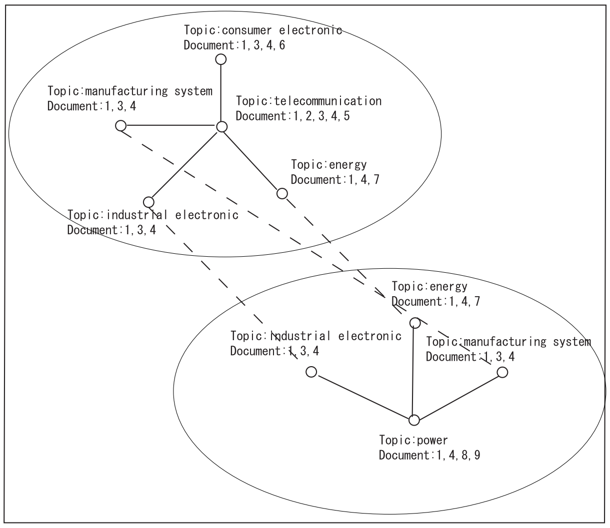

Figure 2: An Example of Clustering

3.4 Displaying Clusters

The User Interface(UI) module provides an Explore like display interface with a retrieval function.

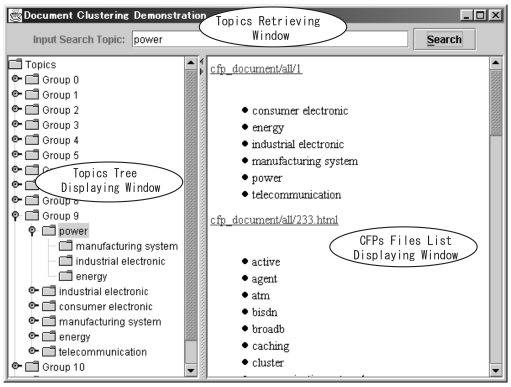

Figure 3: The User Interface

As shown in figure 3, it consists of three windows: Topics Retrieving(TR) window, Topics Tree Displaying(TTD) window and CFPs files List Displaying(CLD) window. The TR window located in the top side of the UI module accepts a keyword to search a topic node in the TTD window. The found topic is marked in the TTD window. The CLD window shows a list of the names of CFPs files, whose one of the topics is a leaf of the topics tree displayed, with the links to the files.

4 Experimental Results

4.1 Results of Filtering

First, we made experiments to measure the filtering performance with SVM^{light}[4, 3]. The training data and test data were obtained with Google[1], Yahoo[2] and DIKU[3]. In order to obtain positive data, i.e., the data belonging to the CFPs category, the collecting file module sent the query 'call for paper' to Google and accessed DIKU's page. To obtain negative data, we sent 16 queries[4] to Google and Yahoo and gathered 20 pages each for avoiding deviation.

The criteria to classify the data collected to positive data is that the data contains the following information: a conference name, topics, program committee, submission deadline, submission information other than date.

As a kernel function of SVM, we employ a linear function based on preliminary experimental results for evaluating the performance of three kernel functions:linear, RBF and Polynomial functions. The number of files in the training set is 604(positive 213, negative 391) and that in the test data is 253(positive 79, negative 174). Table 1 shows the experimental result of filtering.

Table 1: Experimental Results of Filtering with SVMs

Pos.	Neg.	Accuracy	Precision	Recall
79	174	91.07%	84.71%	90.00%

Where Pos. and Neg. stand for the number of positive data and negative data, respectively. Accuracy, Precision and Recall are defined by the following equations.

$$Accuracy = \frac{\#\ of\ files\ classified\ correctly}{\#\ of\ test\ data} \qquad (4)$$

[1] http://www.google.com

[2] http://www.yahoo.com

[3] http://www.diku.dk/research-groups/topps/Conferences.html

[4] They are travel, program, course, staff, information system, conference, computer science, research, game, sport, finance, education, music, weather and book.

$$Precision = \frac{\#\ of\ files\ classified\ correctly\ in\ Positive\ data}{\#\ of\ files\ classified\ as\ Positive\ data} \quad (5)$$

$$Recall = \frac{\#\ of\ files\ classified\ correctly\ in\ Positive\ data}{\#\ of\ Positive\ data} \quad (6)$$

From this result, we made sure that SVM would work enough for the filtering task of DIREC system.

4.2 Results of Extracting Topics

First, sending the query 'call for paper' to goo[5], the Web page collecting module gathered 951 web page files. After filtering out irrelevant pages with SVM^{light}, the filtering module judged 485 files among them as Positive data. The experimental results of extracting topics are shown in table2. Although the rate of the files whose topics were extracted is only 56.5% of the whole, it is 93.8% of the whole files with an itemization style. It is only 18 files with an itemization style that any topic was not extracted. The accuracy of topics extracted from 274 files is 79.2% with complete exactness and 88.7% with partially exactness.

Table 2: Experimental Results of Topics Extraction

The number of Files	Extracted	not-Extracted	
485	274	214	
The number of Files	Correct	Wrong	Partially Correct[6]
274	217	31	26

4.3 Results of Clustering

For the experiments of clustering, we used the topics extracted from 274 Web pages as shown in table 2 with adding the topics extracted from the 124 E-mail messages that were also filtered by the filtering modules for the messages received by one of the authors.

Table 3 shows the relationship between the number of CFPs files that a topic appears(# of CFPs) and the number of topics that appear in X numbers of CFPs files($X \geq 2$). When considering the topics that appear two or more CFPs files, the number of them are totally 969.

Table 3: The Number of Topics According to the Number of Occurrences that the Topics Appeared

# of CFPs	# of Topics	# of CFPs	# of Topics
2	488	5	56
3	197	6	45
4	93	≥ 7	90

The clustering procedure is described as follows:

1. Extracting topics from the CFPs files collected.
2. Calculating the similarity between every two topics among the whole topics extracted and obtaining a set of topics relative to each topic. Where both TH_m and TH_n described in section 3.3.1 were set to 0.5 based on preliminary experimental results.
3. Creating base clusters each of which consists of a pair of a topic and its relative topics. All topics with no relative topics were classified to a special cluster.
4. Creating another cluster by combining base clusters with a common topic. If a base cluster containing only two topics is not combined with any other base clusters, the base cluster is also classified to the special cluster.

4.4 Discussion about Clustering Results

Table 5 depicts the clustering results. Only the remarkable topics, that have more relative topics than others, are shown in the table. The cluster no.2 stands for the special cluster described in the previous section.

Since most of the CFPs files used in this experiment have the relation to a computer science field, most of the topics extracted were classified in the cluster no.0, which consequently can be regarded as a computer science field. Although each of other clusters, however, has only a few topics, we can see that they have their own properties. That is, as shown in table 4, the clusters no. 3, 6, 7 and 10 clearly stands for a test, bioscience, electronics and electric field, respectively.

Table 5: Clustering Results

Cluster No.	# of Topics	Cluster No.	# of Topics
0	603	7	3
1	3	8	3
2	326	9	6
3	3	10	3
4	4	11	3
5	3	12	3
6	8		

The pairs of a central topic of a base cluster and its relative topics shown in table 6 say that the relationship between them are suited to our intuition. For example, the topic ' network architecture' has a strong relevance to the topics ' security', ' protocol', 'service', 'wireless', and the topic ' natural language interface' has a strong relevance to the topics ' information retrieval', 'ontology', 'machine translation', ' information extraction', and so forth. Furthermore, since the topics are also connected with the conference, we can easily get the conference names relevant to a topic. For example, from the topic 'fault modeling', we obtained some conference names such as 'IEEE European Test Workshop' or ' 1st IEEE Latin-American Test Workshop', and knew its remarkable topics such

[5]http://www.goo.ne.jp

Table 4: Clusters Example

Cluster No.	Topics Example	Cluster No.	Topics Example
0	electronic commerce, telecommunication, tracking, distributed database, information retrieval	7	power, industrial electronic, energy, consumer electronic, manufacturing system
1	corba, java, dcom	8	maj, theme, future of
2	**[isolated topics]**	9	panel, presenter, tutorial
3	defect, thermal testing, fault modeling	10	magnetic, proces, monitoring
4	hypertext, text corpora, statistical model	11	interpretation, composition, artificial
5	movement, film, media	12	power electronic, modelling, diagnostic
6	biological, psychological, lo angele, evolution, chemical, ucla		

Table 6: Example of Pairs of Central Topic and its Relative Topics

network architecture	natural language interface	magnetic	java
security network protocol service wireless	information retrieval ontology machine translation information extraction requirement engineering	process monitoring	dcom corba
pattern recognition	**parallel application**	**electronic commerce**	**decision tree**
multimedia image process neural network computer vision image analysis	distributed digital library agent mobile computing the internet multimedia technology internet service domain specific language collaboration technology	security browser replication payment virtual marketplace rational information agent authorization auction	planning neural network reinforcement learning problem solving

as 'test', 'reliability', 'defect', 'thermal testing', ' simulation' and ' process control' even though we would not be familiar to the field.

5 Conclusion and Future Works

This paper discussed DIREC system that obtains the relationships between topics of international conferences by filtering Web pages, extracting and clustering topics.

Experimental results said that DIREC system created clusters each of which clearly corresponded to each different research field, although most of topics were assigned to one cluster corresponding to the computer science field. We have to do further experiments with huge size of data.

We are combining DIREC system with the conference information management system[6] so that DIREC finds another relationship between other named entities appeared in CFPs, and also applying DIREC system to the Research and Education Report database in Kyushu University [7] although we can not describe them due to the page limitation.

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

[7]Most of the contents of the database are currently written in Japanese. http://www.ofc.kyushu-u.ac.jp/kyokandb/

References

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

[2] Christopher J.C. Burges. A tutorial on support vector machines for pattern recognition. *Data Mining and Knowledge Discovery*, 2(2):121–167, 1998.

[3] Thorsten Joachims. Text categorization with support vector machine:learning with many relevant features. In *Proceedings of Tenth European Conference on Machine Learning(ECML-98)*, pages 137–142, 1998.

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

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

[6] Tsunenori Mine, Makoto Amamiya, and Teruko Mitamura. Conference information management system : Towards a personal assistant system. In *Proceedings of The First Asia-Pacific Conference on Web Intelligence (WI-2001)*, pages 247–253, 10 2001.

[7] Guoqiang Zhong, Satoshi Amamiya, Ken'ichi Takahashi, Tsunenori Mine, and Makoto Amamiya. The design and application of kodama system. *IEICE Transactions*, page to appear, 2002.