Document Clustering Using Reference Titles[*]

인용문헌 표제를 이용한 문헌 클러스터링에 관한 연구

Sang Hee Choi[**]

ABSTRACT

Titles have been regarded as having effective clustering features, but they sometimes fail to represent the topic of a document and result in poorly generated document clusters. This study aims to improve the performance of document clustering with titles by suggesting titles in the citation bibliography as a clustering feature. Titles of original literature, titles in the citation bibliography, and an aggregation of both titles were adapted to measure the performance of clustering. Each feature was combined with three hierarchical clustering methods, within group average linkage, complete linkage, and Ward's method in the clustering experiment. The best practice case of this experiment was clustering document with features from both titles by within-groups average method.

초 록

본 연구에서는 원문헌의 표제가 문헌클러스터링에서 문헌의 주제를 나타내는데 효과적인 자질로 인식되고 있지만 동의어나 유사어를 포함하여 문헌의 주제를 대표하는데 한계가 있음을 인지하고 인용문헌의 표제로 클러스터링 자질을 확대하는 방안을 제시하였다. 문헌 클러스터링의 자질로 원 문헌의 표제 용어와 인용문헌의 표제 용어, 두 종류의 표제 용어를 혼합하여 적용하여 인용문헌의 표제가 클러스터링 성능을 향상시키는 정도를 측정하였다. 각 자질별로 계층적 클러스터링 기법 3개, within group average linkage, complete linkage, Ward 기법을 결합하여 클러스터를 생성하는 성능을 비교, 분석하였는데 원문헌과 인용문헌 표제어를 혼합하여 within group average linkage 기법으로 클러스터링 한 경우가 가장 좋은 결과를 나타내었다.

Keywords: document clustering, clustering feature, clustering method, title, citation
문헌클러스터링, 클러스터링 자질, 클러스터링 기법, 표제, 인용,

[*] This work was supported by research funds from the Catholic University of Daegu.

[**] Department of Library Science, Catholic University of Daegu(shchoi@cu.ac.kr)

▪ Received : 27 May 2010 ▪ Revised : 4 June 2010 ▪ Accepted : 17 June 2010
▪ Journal of the Korean Society for Information Management, 27(2): 241-252, 2010.
〔DOI:10.3743/KOSIM.2010.27.2.241〕

1. Introduction

Document clustering models have been developed in many information retrieval areas to help users. Document clustering enables users to find information by subject category and to recognize valuable patterns in information. Most existing document clustering methods are based on lexical analysis, such as the bag-of-words model. The bag-of-words model is considered effective for grasping the content of a document. Vocabulary from the title, the abstract, and the body of a document are usually used as clustering features in terms of representing the contents of the documents. There are, however, some issues in adapting these components to cluster documents.

Title word is a good feature for revealing the topic of a document, especially considering that these words are picked by authors to present key concepts in their work. In particular, the title words of scientific publications are usually regarded as very effective for presenting the topic of a document.

A title usually consists of 5-10 words except stop words, so authors choose the best terms for their titles among many synonyms. The nature of title words leads to drawbacks in clustering documents. As documents are clustered by the portion of lexically equivalent words, the documents could be poorly clustered with few words from titles without a variety of synonyms. In contrast, if the clustering system uses words from the full text of a document as the clustering feature, the portion of lexically equivalent words could increase and it could give a solution to the lack of clustering features. Even though it provides sufficient features to measure the similarity of documents' contents, too many words that are irrelevant to a document's topic could scatter the documents themselves. Compared with title words, many terms in full text words are not effective to represent the subject of a document. Therefore, simply adding words in the text to title words would not improve the performance of clustering.

Since ignoring synonyms in relevant documents and increasing irrelevant terms both result in poor performance, it is more important to find proper synonyms and relevant terms rather than to increase the number of irrelevant words to the topics of documents.

The title words of cited works could be a good source for extracting synonyms and relevant terms to represent the topic of a document without adding irrelevant terms in clustering documents. Cited papers are usually considered to be closely related to the subject of the original paper, which has put them in its references.

This study attempts to examine the power of reference title terms to represent the topic of the original document in clustering documents and suggests an approach to improve the performance of clustering documents with them. To identify the effectiveness of title words clearly, scientific articles in the Scopus database were selected as documents to cluster. To narrow the subject areas of documents, author affiliation was limited to a specific institute, the Catholic University of Daegu.

This study measured and compared the performance of clustering with title words which were extracted under three different conditions. Clustering features were chosen from two types of titles to detect the subjects of documents. One was the title words of scientific articles; the other was the title words in the references of the articles. Each feature was independently adapted to generate clusters and the aggregated feature which integrated two features was also used to measure the improvement by reference titles. In addition, three hierarchical clustering methods were applied to cluster documents and the best practice case with three different title features was identified.

2. Previous Research

Document clustering has been adapted to a variety of areas such as web results clustering, topic discovery, automatic summarization, and research trend analysis. Automatic document clustering provides an effective solution for information-overloaded users. Users can visualize the search space or search results using labeled clusters of documents which have been classified into topical categories (Aljaber, Stokes, and Bailey 2009).

Document clustering also enables researchers to discover important patterns in a specific domain. Yang discovered the historical vein and forecasted the future possible tendency of machine learning research areas by literature cluster analysis (Yang, Liao, Wu, and Yin 2009). In recent research, clustering scientific literature has also been used to identify country core competencies (Kostoff, Cortes, et al, 2007).

Most clustering models rely on the vector space model to measure the similarity between documents. In general, the vector is the bag of words from a document. To cluster documents, titles and full text are commonly used as clustering features, but they often provide poor clustering results (Staff 2008). The list of frequencies of words from a document might represent the content of the document to some extent, but the semantics might be misinterpreted. Even though word stemming can help to infer the semantic equivalence of lexical variants, synonyms in a particular domain may be regarded as different terms. This problem may produce poor clustering results due to the low similarity of documents that use different terms for the same concepts (Tong, Dinakarpandian, and Lee 2009). Therefore, it led several researchers to attempt to select more effective words for clustering features

To improve the accuracy of document clustering, citations could be an informative source. A recent study of document clustering referred to citation semantics as Citonomy. According to this research, Citonomy can infer some of the semantics of a paper based on the references it cites. The rationale behind using the references in document clustering is that documents on the same topic will have similar semantics. Citonomy is based on clustering the list of references of each document followed by document level clustering (Tong, Dinakarpandian, and Lee 2009). Aljaber also used citation context to provide relevant synonymous and related terms to increase the effectiveness of lexical representation (Aljaber, Stokes, and Bailey 2009). This research suggested a link-based clustering method which determines the similarity between documents using the number of co-citations. In another study using citation context for document clustering, sparse citation graph analysis was adapted in measuring the significance of individual words. The weight of clustering features was based on the underlying link structure of the document collection to generate distinct document clusters (Bolelli, Ertekin, and Giles 2006).

3. Research Overview

3.1 Data Collection and Subject Categories of Scopus

The data for this study were collected in the Scopus database to compare the performance of each clustering feature, titles and reference titles. Scopus provides scientific articles with titles and references and also categorizes the articles with pre-defined subject categories. It allows the assigning of multiple subject categories to one article if it has multidisciplinary content. Scopus subject categories were used as the criteria to evaluate clustering performance.

To build suitable data collection for the examination of clustering performance, data in Scopus were extracted through three steps. First, 1,427 articles were searched where author affiliation was the Catholic University of Daegu. Scopus provides an affiliation index which allows access to data produced by authors from a certain institute. This study limited author affiliation to one institute to collect a more cohesive data group in terms of subject. If author affiliation was limited to a university, a certain group of authors was selected which enabled the gathering of multiple research results produced by the authors who had done research on the same topics. Therefore, data searched by affiliation could be categorized into more cohesive groups in terms of subject and it could give a clearer view in evaluating the performance of clustering documents by subject.

Second, among the 1,427 articles searched, a filtering process was applied to articles which were assigned to multiple subject categories. To compare automatically produced clusters with pre-defined subject categories by Scopus, two criteria were adapted to select articles for the clustering experiment. One was the number of assigned categories; the other was the size of subject categories. To compare pre-defined subject categories with auto-generated subject clusters, articles assigned to multiple categories were excluded. In addition, since overly small categories, for example those containing only one or two documents, could adversely affect clustering performance, categories with fewer than 10 documents were omitted from this clustering study. 500 articles within a single category were re-selected.

In the third stage of data collection building, titles and reference titles of 500 articles were parsed and term frequencies were calculated. Stop words and terms with a frequency below 2 were removed, which were usually not able to affect the clustering process. As a result, 118 articles were filtered in the selected data and 382 articles which had terms with a frequency over 2 remained and were applied to this study.

There were 11 subject categories in the final selected data. The titles of each category are presented in Table 1. All subject categories were scientific areas and their size varied from 10 to 62.

3.2 Clustering Features and Methods

Clustering features were the key factors for gathering similar documents into a cluster. Title words have been used as a key feature in many clustering experiments to identify the subjects of documents. With scientific papers, there are two kinds of titles, both of which can represent their subjects. One is the title of the article, which consists of subject words picked by the author. The other is the titles of papers in the article's reference list. The basic assumption concerning the relationship between an article and its references is that the topics of cited articles are closely related to that of the citing article. Thus, title words from referenced articles could also be features to represent the topic of the article which the references belong to.

To compare the effects of clustering features in gathering similar documents, the following three features from the titles in selected articles were applied.

Feature 1: Title words from original article titles.
Feature 2: Title words from the titles of articles cited in the bibliography
Feature 3: An aggregation of features 1 and 2

In the rest of this paper, feature 1 is referred to as ‘original titles’, feature 2 as ‘reference titles’, and feature 3 as ‘aggregated titles’.

Three agglomerative hierarchical clustering methods, within-group average linkage, complete linkage, and Ward's method, were applied to generate document clusters. The clustering performance of each method was discussed in association with the three clustering features outlined above.

In total, nine cases were applied in this study combining the three clustering features with the three clustering methods. They are as follows.

<Table 1> Pre-defined subject category titles of data collection

Pre-Defined Subject Category Titles(by Scopus)	Number of Documents
Agricultural and Biological Sciences	42
Biochemistry, Genetics and Molecular Biology	44
Chemistry	16
Computer Science	73
Dentistry	10
Engineering	45
Materials Science	20
Mathematics	28
Medicine	61
Pharmacology, Toxicology and Pharmaceutics	28
Physics and Astronomy	15
Total	382

Case 1 : original titles with within group average clustering method

Case 2 : reference titles with within group average clustering method

Case 3 : aggregated titles with within group average clustering method

Case 4 : original titles with complete linkage method

Case 5 : reference titles with complete linkage method

Case 6 : aggregated titles with complete linkage method

Case 7 : original titles with Ward's method

Case 8 : reference titles with Ward's method

Case 9 : aggregated titles with Ward's method

3.3 Measures for Clustering Performance

To evaluate clustering performance, prerequisites such as manually pre-defined clusters and the number of clusters must be generated. According to Chung and Lee, who compared the existing measure of clustering performance and developed a new unbiased measure, WACS, the performance of clustering should be measured by the same measure with the same number of clusters. (Chung and Lee 2002) This study used eight measures adapted in Chung and Lee's research (2002) and in Kim and Lee's (2001). Each measure is explained below.

D: Total number of documents

m: Number of manually pre-defined clusters

n: Number of automatically generated clusters

Mi: Manually pre-defined cluster i

Ci: Automatically generated cluster I

1) Chi-square

$$x^2 = \sum_{i=1}^{i=R}\sum_{j=1}^{j=C}\frac{(O_{ij}-E_{ij})^2}{E_{ij}}$$

R: Number of clusters from clustering result A

C :Number of clusters from clustering result B

O_{ij} : Observed frequency

E_{ij} : Expected frequency

2) Entropy

$$\text{Entropy(Ci)} = \sum_{i=1}^{n} -p_i \log_2 p_i$$

pi: ratio of Cluster C's documents assigned cluster pi

3) F measure, precision and recall

$$\text{Clustering Precision} = \frac{1}{D}\sum_{j=1}^{n}\sum_{i=1}^{m}\frac{|Mi\cap C_j|}{|C_j|}$$

$$\text{Clustering Recall} = \frac{1}{D}\sum_{j=1}^{n}\sum_{i=1}^{m}\frac{|Mi\cap C_j|}{|M_i|}$$

$$\text{Clustering F(p,r)} = \frac{2pr}{p+r}$$

4) WACS

$$\text{Clustering Recall} = \frac{1}{D}\sum_{j=1}^{n}\sum_{i=1}^{m}\frac{2|Mi\cap C_j|^2}{|M_i|+|C_j|}$$

5) Mutual Information(MI)

$$\text{I(M:C)}=\frac{1}{D}\sum_i\sum_j|M_i\cap C_j|\log_2\frac{D|M_i\cap C_j|}{|M_i||C_j|}$$

<Table 2> Analysis of clustering result by number of document pairs

		Clustering Result B	
		same cluster	different cluster
Clustering Result A	same cluster	a	b
	different cluster	c	d

6) CSIM

1: number of document pairs assigned to the same cluster.

0: number of document pairs assigned to a different cluster.

$$\text{CSIM}(C_A,\ C_B) = \frac{2a}{2a+b+c}$$

4. Analysis of Clustering

4.1 Clustering Performance of Features

4.1.1 Clustering performance of features by within group average method

When documents were clustered using the within-group average method, aggregated titles showed the best performance of clustering by eight measures (Table 3). In comparison with the performance of unitary features (original titles and reference titles), reference titles were superior to original titles in all cases. This revealed that reference titles could be a better feature than original titles for representing the topic of documents. This result also led to the interpretation that the improvement by aggregated titles could be caused by reference titles.

4.1.2 Clustering performance of features by Ward's method

The most effective feature combined with Ward's method was also aggregated title. Except for the case of recall, seven measures proved that aggregated titles gathered documents more similarly to pre-de-

<Table 3> Clustering performance of features by within group average method

measures	Ori.-Titles	Ref.-Titles	Aggregated-Titles
Chi-square	282.011	346.371	453.393
Entropy	2.340	2.100	1.760
F measure	0.222	0.297	0.359
Precision	0.236	0.279	0.347
Recall	0.209	0.318	0.371
WACS	0.205	0.269	0.332
MI	0.817	1.000	1.324
CSIM	0.187	0.278	0.330

* Underlined score is the best case.

fined subject categories by Scopus(Table 4). Even though reference titles scored better than aggregated titles in recall measurement, the gap between the two scores was slight enough (0.001) for the difference to be ignored. Therefore, it is reasonable to conclude that both features resulted in the same performance.

Original titles provided the poorest performance in clustering documents among other features in all measures like the within-group average method's clustering results.

4.1.3 Clustering performance of features by complete linkage method

In most cases, complete linkage presented that aggregated titles were more effective for clustering documents based on their topics than the other two features. By six measures except recall and CSIM, the clustering result with aggregated titles was picked as the best case(Table 5). However, unlike the other methods, complete linkage seemed to work effectively with original titles, which were regarded as clustering documents most poorly among the clustering features.

In case of recall, the original title provided superior performance to the other features. The difference from other features' scores was also large and affected the F measure. The F measure was calculated with precision and recall, so the big difference of the

<Table 4> Clustering performance of features by Ward's method

measures	Ori.-Titles	Ref.-Titles	Aggregated-Titles
Chi-square	279.217	321.906	367.494
Entropy	2.522	2.387	2.309
F measure	0.310	0.355	0.381
Precision	0.250	0.269	0.300
Recall	0.409	0.524	0.523
WACS	0.208	0.245	0.265
MI	0.697	0.838	0.931
CSIM	0.202	0.223	0.236

<Table 5> Clustering performance of features by complete linkage method

measures	Ori.-Titles	Ref.-Titles	Aggregated-Titles
Chi-square	170.097	290.262	350.886
Entropy	2.762	2.422	2.278
F measure	0.312	0.300	0.361
Precision	0.199	0.263	0.293
Recall	0.722	0.348	0.471
WACS	0.219	0.218	0.246
MI	0.449	0.790	0.904
CSIM	0.231	0.201	0.225

recall score had an effect on the F score of original titles. The F score became higher than that of reference titles due to the recall score.

Original titles also showed the best CSIM score. In terms of interpreting CSIM scores, the score of original titles was closer to that of aggregated titles than the score of reference titles. This indicated that original titles had a greater effect on the performance of aggregated titles than reference titles.

In addition, WACS evaluated original titles and reference titles had similar capability to cluster documents. In previous clustering results attained by other methods, reference title proved to be a better feature than original title by all measures.

4.2 Overall Performance of Clustering Features and Methods

The most effective feature for document clustering was aggregated titles. Seven of eight measures produced the best score with aggregated titles. Only with the complete linkage method, clustering recall showed the best score with original titles. In general, aggregated titles were most effective for gathering similar documents like the pre-defined subject categories of SCOPUS(Table 6).

The original title was the least effective clustering feature. It ranked lower than the other two features by most measures except recall. Compared with the other measures, recall showed inconsistency in evaluating the performance of clustering features.

The performance of reference title was ranked between aggregated title and original title. In many cases, reference title improved the clustering performance of original titles. Consequently, the improvement of clustering performance using aggregated titles could result not from original titles, but from reference titles. This indicates that titles from citations could more effectively represent the topic of literature rather than its own title.

In terms of clustering method, the within-group average method was the most suitable method for document clustering with reference titles and aggregated titles. It was ranked the highest by 6 meas-

〈Table 6〉 Ranking by performance of clustering features and methods

clustering methods	Within Group Average(WGAV)			Ward's(WARD)			Complete Linkage(COMP)		
measures	ORI.	REF.	AGG.	ORI.	REF.	AGG.	ORI.	REF.	AGG.
Chi-square	7	4	1	8	5	2	9	6	3
Entropy	5	2	1	8	6	4	9	7	3
F measure	9	8	3	6	4	1	5	7	2
Precision	8	4	1	7	5	2	9	6	3
Recall	9	8	6	5	2	3	1	7	4
WACS	9	2	1	8	5	3	6	7	4
MI	6	2	1	8	5	3	9	7	4
CSIM	9	2	1	7	6	3	4	8	5

* gray cell: the best case, ** ORI: Original titles, *** REF: Reference titles, **** AGG: Aggregated titles

ures with aggregated title. With original titles, however, it was ranked the lowest by 4 measures. The poor performance of original titles in within-group average clustering was notably improved by adding reference titles as a clustering feature. While the performance of original titles was ranked 9th, the performance of aggregated titles, which was composed of original titles and reference titles, was ranked 1st in the evaluation of clustering performance by WACS and CSIM. Therefore, it was quite clear that reference titles contributed to the large improvement of the performance in within-group average clustering.

In contrast, the complete linkage method seemed to prefer original titles as clustering features. Except for the complete linkage method, the other two clustering methods showed better clustering performance with reference titles than with original titles in all cases evaluated by 8 measures. Only in the complete linkage cases, 4 of 8 cases showed better clustering performance with original titles than with reference titles. With recall in particular, clustering performance with original titles was the best of all. Proceeding from this fact, it could be logically assumed that reference titles did not worked so effectively with the complete linkage method as they did with the within-group average method.

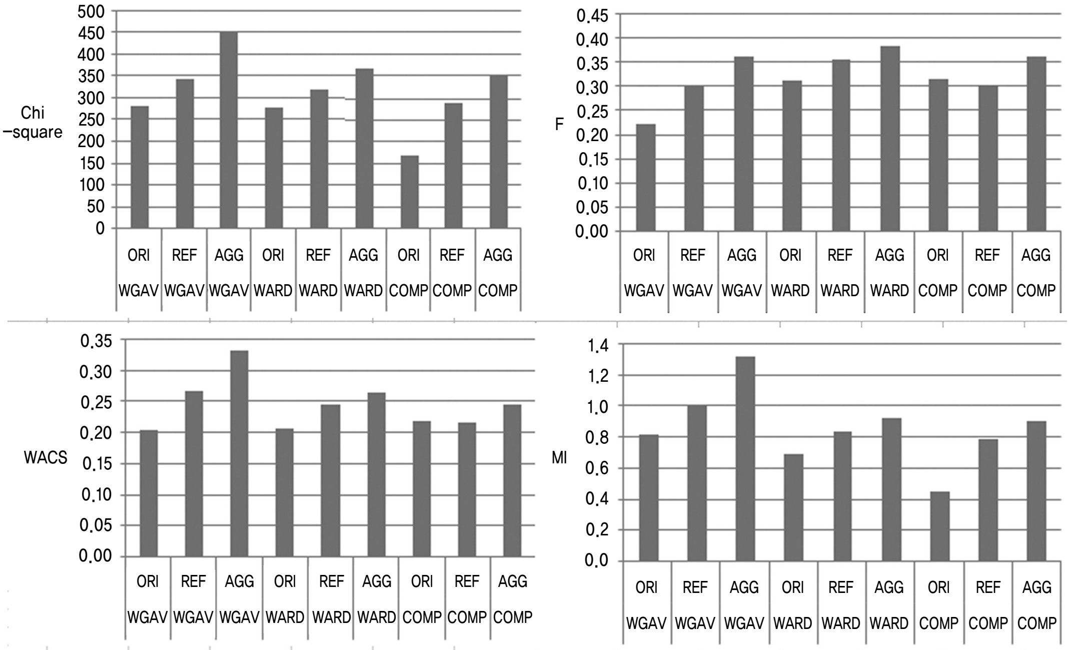
5. Conclusion

This study reviewed the capability of referenced titles to represent the topic of literature in document clustering. The conclusion to be drawn here is that reference titles showed better performance in clustering documents than original titles, and aggregating original titles with reference titles is an effective solution for generating clustering features. Reference titles proved to be a better feature for document clustering in most cases, even though complete linkage showed that original titles had clustered documents better than reference titles.

The improvement made by reference titles can be clearly seen in Figure 1. In the 4 measures, aggregated titles produced the best clustering results regardless of which of the three clustering methods was used. This suggests that adding reference titles to clustering features could improve the performance of clustering documents with original titles. Compared Tong's study, which used semantic factors of references such as citation locality and contextual information of citation, this study suggests simplified way to improve clustering performance using reference titles and it will help to reduce time to select clustering features.

Another finding of this study is that some clustering methods could have preference for clustering features. Proceeding from what has been said above, complete linkage prefers original titles to reference titles, although reference titles have showed better clustering performance in all cases of the other clustering methods. Conversely, reference titles had worked well, especially with the within-group average method, and greatly improved clustering performance.

The results of this study provide a useful foundation for utilizing citation titles to cluster documents or to detect the topic of documents, but there remains the need for more investigation and research. This study used only the title part of references. More research should be done to explore the capability of other parts of the citation text, such as journal titles or author names, as clustering features.



* ORI: orignal titles, ** REF: reference titles, *** AGG: aggregated titles

<Figure 1> Clustering performance with three title features

References

Chung, Young-mee, and Jae Yun Lee. 2001. "Development of an unbiased measure for clustering performance." *Proceedings of the 7th conference of Korean Society for Information Management,* 23-24 August, 2001, [KISTI, Seoul], 167-172.

Guo, Qinglin, and Ming Zhang. 2009. "Multi-document automatic abstracting based on text clustering and semantic analysis." *Knowledge-based systems,* 22(6): 482-485.

Hudes, Mark L., Joyce C. McCann, and Bruce N. Ames. 2009. "Unusual clustering of coefficients of variation in published articles from a medical biochemistry department in

India." *The FASEB Journal,* 23(3): 706-708.

Kim, Jun-Ha and Jae Yun Lee. 2000. "A Comparative study on performance evaluation of document clustering results." *Proceedings of the 7th conference of Korean Society for Information Management,* 24-25 August, 2000, [Ewha Womans Univ., Seoul], 45-50.

Kostoff, Ronald N. J. Antonio del Río, Héctor D. Cortés, Charles Smith, Andrew Smith, Caroline Wagner, Loet Leydesdorff, George Karypis, Guido Malpohl, and Rene Tshiteya 2007. "Clustering methodologies for identifying country core competencies." *Journal of Information Science,* 33(1): 21-40.

Kuo, June-Jei, and Hsin-Hsi Chen. 2007. "Cross-document event clustering using knowledge mining from co-reference chains." *Information Processing and Management,* 43(2): 327-343.

Staff, Chris. 2008. "Bookmark category web page classification using four indexing and clustering approaches." *Lecture notes in computer science,* Vol.5149: 345-348.

Tong, Tuanjie, Deendayal, Dinakarpandian, and Yugyung Lee. 2009. "Literature clustering using citation semantics." *Proceedings of the 42nd Hawaii international conference on system sciences.* 5-9 January 2009, [HICS; Waikola, HI], 1-10.

Zhang, Lin, Frizo Janssens, Liming Liang, and Wolfgang Glänzel. "Journal cross-citation analysis for validation and improvement of journal-based subject classification in bibliometric research." *Scientometrics,* 82(3): 687-706.

Zhao, Yueyang, Lei Cui and Hua Yang. 2009. "Evaluating reliability of co-citation clustering analysis in representing the research history of subject." *Scientometrics,* 80(1): 91-102.

Zhu, Shanfeng, Ichigaku Takigawa, Jia Zeng and Hiroshi Mamitsuka. 2009. "Field independent probabilistic model for clusteing multi-field documents." *Information Processing and Management,* 45(5): 555-570.