Tulips R
University of Tsukuba Library

<Workshop>A Graph-based Method for Automatic Generation of Multilingual Keyword Clusters and Its Applications

著者	Aizawa Akiko, Kando Noriko, Kageura Kyo
journal or publication title	ディジタル図書館
number	12
page range	95-104
year	1998-09
URL	http://hdl.handle.net/2241/103012

A Graph-based Method for Automatic Generation of Multilingual Keyword Clusters and Its Applications

Akiko AIZAWA, Noriko KANDO and Kyo KAGEURA

National Center for Science Information Systems
3-29-1 Otsuka, Bunkyo-ku, Tokyo, 112-8640 Japan
Tel:+81-3-3942-6994,6968 ; Fax:+81-3-5395-7064
E-Mail: akiko@rd.nacsis.ac.jp

Abstract We are investigating an approach to automatic generation of Japanese-English bilingual keyword clusters using the keyword lists assigned to academic papers by the authors. The bilingual clusters generated by our graph-based method contain keywords with similar meanings from both languages and could be valuable linguistic resources in various information retrieval (IR) applications. In this paper, we first present an overview of our clustering method and then show several experimental results to evaluate the generated clusters. We also discuss the limitation and possible extensions of the current implementation.

keywords cross-lingual information retrieval, automatic extraction of thesaurus, bilingual corpus, graph theory, academic paper database

1 Introduction

The explosive growth of online documents has increased the need for IR systems that cross language boundaries. For instance, since the first workshop on cross-lingual information retrieval (CLIR) in 1996 (Grefenstette, Smeaton & Sheridan, 1996), the topic has been one of the most actively pursued in IR research field. One especially interesting and important research to this direction is the *automatic* generation of domain-dependent multilingual thesaurus; this not only help searchers of multilingual scientific databases but also are useful in monolingual search since technical terms are often imported either in their original forms or as acronyms, transliterated, and then translated (Kando 1997).

Presently, we are investigating an approach to automatic generation of Japanese-English bilingual keyword clusters using the keyword lists assigned to academic papers by the authors (Aizawa & Kageura, 1998). The bilingual clusters generated by our graph-based method contains keywords with similar meanings from both languages and could be valuable linguistic resources in various IR applications.

The basic idea of our clustering strategy is that we apply graph theoretic method instead of statistical ones which seem dominant in corpus-based approaches; the bilingual keyword pairs constitute a tangled graph of Japanese and English keywords; as such, the clustering problem can be regarded as a problem of partitioning the original keyword graph by eliminating wrongly generated links from the graph; then, the problem can be transformed into the *minimum cut problem* in the graph theory.

Applying graph theoretic method has several advantages. First, low-frequency key-

words can be treated properly by utilizing topological features of the graph. Second, the clusters contain not only Japanese-English pairs, but also Japanese-Japanese and English-English pairs. Thus, they are usable for not only CLIR but also query expansion in monolingual IR. Third, the whole process can be achieved with reasonable computational cost.

The keyword data in view of bilingual corpus also has several advantages. In conventional corpus-based approaches for CLIR (Dunning & Davis, 1993; Landauer & Littman, 1990; Carbonell, 1997), the lack of readily available parallel corpora has been a bottleneck. On the other hand, our corpus is available for a great many subject domains, with rich variations that may not be listed in the standard dictionaries but are nevertheless meaningful and useful in IR. Another advantage of our keyword corpus is that keyword pairs can be easily extracted without expensive natural language processing for segmentation and alignment.

In the following, we first briefly summarize the outline of our clustering method and then report several experimental results to evaluate the generated clusters; comparison with standard dictionaries, analyzing errors in the clustering, and the performance with bilingual and monolingual IR. We also discuss the limitation and possible extensions of the current implementation.

2 Overview of Multilingual Keyword Cluster Generation Procedure

Our procedure for generating multilingual keyword clusters is composed of the following stages: (1)extraction of Japanese-English keyword pairs from basic corpus, (2)simple normalization, (3)generation of initial keyword graph, (4)screening obvious non-errors, (5)detection of possible correspondence errors, (6)detection of possible homonymous words, (7)partitioning keyword clusters, and (8)output of final clustering results. Each stage is described briefly in the following.

2.1 Extraction of Japanese-English Keyword Pairs

The basic data used in the current study is the Japanese and English keywords assigned by the authors to their papers, extracted from the NACSIS Academic Conference Database (NACSIS, 1997). We selected 28,122 papers related to the field of computer science. Of the papers selected, 26,060 (about 93 %) have the same number of Japanese and English keywords. An example is :

Japanese: 多言語検索 / 用語クラスタ / グラフ理論 / NACSIS データベース
English: cross-lingual information retrieval / keyword cluster / graph theory / NACSIS Database.

Since the Japanese and English keyword pairs generally maintain good one-to-one correspondences, we mechanically extract total 112,364 Japanese and English keyword pairs (60,186 different ones) and use them as a basic bilingual keyword corpus.

Table 1 shows some examples of the extracted keyword pairs. The most general English translation for each Japanese keyword is marked with '*'.

2.2 Simple Normalization

After the extraction of bilingual keyword pairs, simple normalization is applied to deal with notation variation problem such as *cross-lingual* and *Cross Lingual*. Also, acronyms are detected and marked so that they can be tested for homonyms at later stage.

2.3 Generation of Initial Keyword Graph

The initial graph expression of a bilingual keyword corpus is easily derived by representing Japanese and English keywords as nodes and their translation pairs as links. The frequency of the keyword pair appeared in the

Table 1: Example of Japanese-English keyword correspondences extracted from the database.

Japanese keywords	English keywords	frequency
キーワード	information retrieval	1
キーワード	keyword	39*
テキスト検索	information retrieval	1
テキスト検索	text retrieval	6*
テキスト検索	text search	3
検索指示語	keyword	1
広域情報検索	information retrieval	1
情報検索	information gathering	4
情報検索	information retreival	1
情報検索	information retrieval	320*
情報検索	information search	5
情報収集	information gathering	6*
情報収集	information retrieval	1
文献検索	bibliographic search	1*
文献検索	document retrieval	11
文書検索	document retrieval	19*
文書検索	text retrieval	1

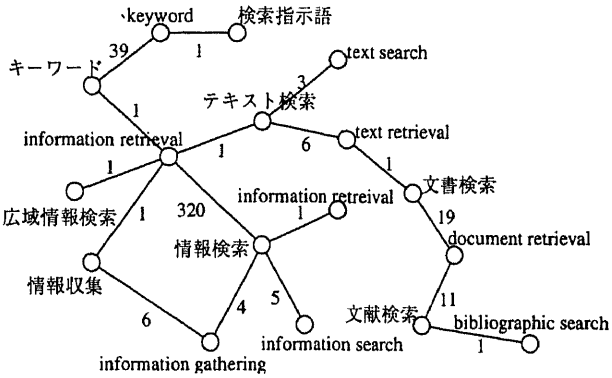

Figure 1: Initial keyword graph generated from keyword pairs in Table 1.

corpus is expressed as the capacity of the corresponding links. Figure 1 shows the initial keyword graph generated from the keyword pairs shown in Table 1.

The global keyword graph is composed of numbers of disjoint sub-graphs, which we define as *bilingual keyword clusters*. In case of the above example, the whole nodes belong to the same keyword cluster at the initial stage and thus are considered to have similar meanings.

2.4 Screening Obvious Non-Errors

This stage currently includes (1)recognition of Japanese and English pairs with identical notations, (2)checking keyword pairs with sufficient frequencies, and (3)detecting *minor examples*. The last case applies when a keyword pair stands for the unique translation of either of the Japanese or English counterpart. Links which satisfy the above conditions are considered to be correct and maintained automatically to reduce the computation cost at later stages.

2.5 Detection of Possible Correspondence Errors

The initial keyword graph constructed from 60,186 different translation pairs contains a huge keyword cluster with as many as 20,659 nodes (about 34%) in it. Since our subject is restricted to the specific academic field *computer science*, the major cause for this is the existence of improper translation pairs which connects otherwise disjoined keyword clusters into one (This is in contrast to general cases where the homonyms are much more common). A typical example found in Figure 1 is the link <キーワード (*keyword*), information retrieval >.

The detection algorithm employed in our procedure is based on a simple principle that *a set of links which decompose a connected keyword cluster into disjoint subclusters when they are removed from the original cluster are the candidates of improper translations.* In the conventional graph theory, such a link set is called an *edge cut* and the edge cut with the minimal total capacity among all the edge cuts obtained for a graph is called a *minimum edge cut.* Minimum edge cut problem is one of the most principal prob-

lems in the graph theory and there exist number of algorithms which guarantee sufficient performance for our purpose.

2.6 Detection of Possible Homonyms

Though homonyms do not seem to occur so frequently in a specific scientific domain, we still have observed several cases such as < ATM (*Asynchronous Transfer Mode*)> and < ATM (*Automatic Teller Machine*) >.

The detection of possibly homonymous keywords can also be done utilizing the topological feature of the keyword cluster. It can be assumed that *homonymous nodes are the ones which decompose the cluster when the node and all the edges starting from the node are removed.* Thus, the problem is transformed again to the well-known node cut problem of the graph theory. Since most of the homonyms we have observed are acronyms, we presently consider only keywords composed of English capital characters and symbols as candidates as node cuts.

2.7 Partitioning Keyword Clusters

The minimum edge cut of a keyword cluster does not always represent imprecise translation. Removing correct pairs inevitably causes oversplitting, i.e. generating more than one clusters with similar meanings. On the other hand, the distinction between corresponding keyword pairs and associated but not corresponding ones depends on the application and is difficult even for human experts. For example, the keyword pair < テキスト検索 (*text retrieval*), information retrieval > may be improper in view of strict terminological definition but not necessarily be incorrect for searchers of academic paper databases.

Our current implementation employs a simple stopping criteria: partitioning occurs only when (1)the total capacity of the minimum edge cut is equal or less than N_α, and also (2)each of the newly generated clusters contains at least one nodes with greater than N_β frequencies. We presently set $N_\alpha = N_\beta (= N)$ and use the same value for all the clusters.

Once candidates for improper translations are obtained, partitioning is done automatically by removing the links on the original graph expression. Homonyms can be similarly processed by splitting the nodes into different clusters. The detection and deletion stages described so far is applied for each keyword cluster recursively until no more pairs can be removed.

2.8 Final Clustering Results

Figure 2 shows the result of the partitioning of the keyword cluster given in Figure 1.

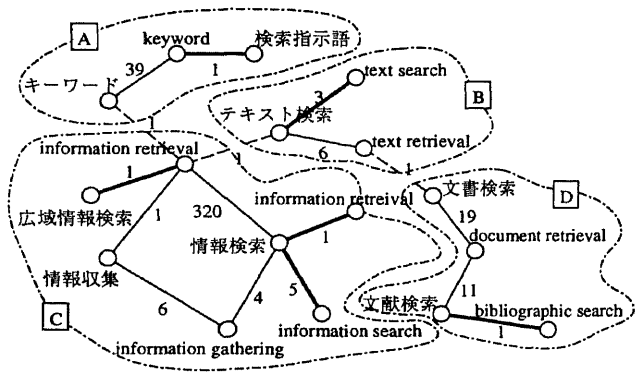

Figure 2: Example of partition of keyword cluster.

As a result of the detection of correspondence errors, three keyword pairs <キーワード (*keyword*), information retrieval>, <テキスト検索 (*text retrieval*), information retrieval>, <文書検索 (*document retrieval*), text retrieval>, are removed, and four clusters A, B, C, and D are newly created. The bold lines shows that the links are marked as unremovable at screening stage. This follows that such pairs as <情報検索 (*information retrieval*), information retreival> (spelling error), <検索指示語 (*keyword*), keyword> (rare case), and <広域情報検索 (*wide-area information retrieval*), information retrieval>

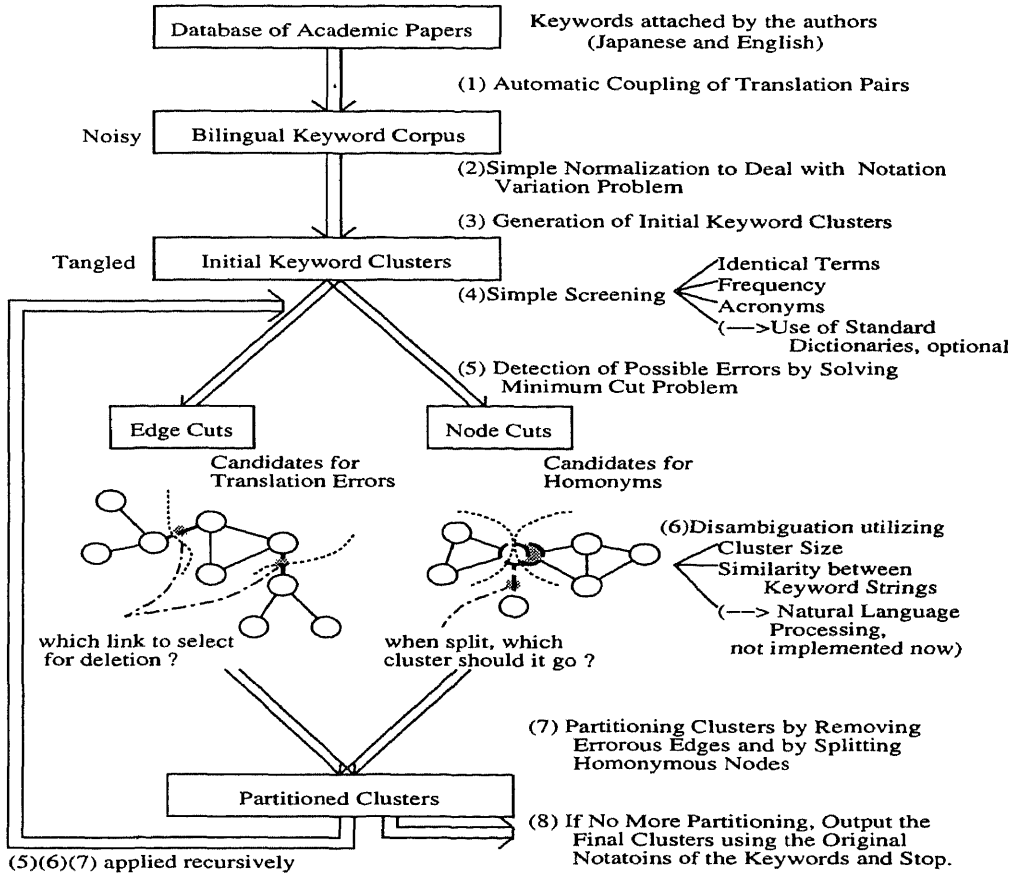

Figure 3: Overview of the proposed keyword clustering procedure.

(related but not equivalent pair) are maintained even after the partitioning.

Applying the method to the whole corpus with $N = 5$, 1,469 of the total 60,186 keyword pairs were deleted, generating the total 27,918 keyword clusters. The number of keyword pairs included in the biggest cluster was reduced to 159 from 20,659 of the initial graph. The overall procedure described in this section is illustrated in Figure 3.

3 Experimental Results

3.1 Comparison and integration with standard dictionaries

In Table 2, the keyword corpus we use in our study is compared with total 22,690 different term pairs extracted from four dictionaries and handbooks of the field of computer science (Aiso, 1993; Japan Society for Artificial Intelligence 1990; Ralston, 1983; Shapiro, 1987).

The comparison shows that the number of common elements between the corpus data and the standard technical dictionaries is relatively small, i.e. many of the keyword pairs assigned by the authors are not listed in the standardized technical dictionaries. This may partly be because the keyword data contains some noises such as spelling errors or notation variations [1]. The noises themselves are useful in IR task since they may also be

[1]The number of common elements are increased to 3,074 from 2,066 after simple normalization to deal with notation variation.

Table 2: Comparison of the technical dictionaries and the bilingual keyword data.

	dictionary terms	corpus pairs	common for both
Japanese words	20,636	37,170	3,966
English words	19,562	49,918	2,814
different translation pairs	22,690	60,186	2,066
average number of translations per word(Jpn)	1.10	1.62	–
average number of translations per word(Eng)	1.16	1.21	–
maximum number of translations per word(Jpn)	7	86	–
maximum number of translations per word(Eng)	6	29	–
number of English acronyms (Jpn)	844	1,007	212
number of English acronyms (Eng)	451	1,233	114
identical Japanese and English pairs	57	1,336	18

found in queries of databases.

Another important point shown in the table is that the Japanese keywords often contains English acronyms in both standard dictionaries and in keyword corpus. The fact supports our assumption that bilingual keyword clusters should be valuable not only for CLIR but also in Japanese monolingual search.

Table 3 shows how the ratio of common pairs between the dictionaries and the corpus changes against the frequencies they appear in the corpus. We can observe that the more frequent a keyword pair appears in the corpus, the higher is the probability it is also referred in the standard dictionaries. For example, among 68 pairs with more than 100 frequencies, only 10 are *not* referred in the standard dictionaries. They are mostly acronyms or terms representing relatively new technologies: i.e. { CAI, CASE, CSCW, LOTOS, OSI, WWW, agent, multi media, genetic algorithm, reuse }. The average frequency of the common pairs in the corpus is 5.9 before normalization and 11.1 after normalization.

In clnclusion, it can be expected that the keyword corpus reflects particular views and concepts of the authors, which can not be covered with the standard dictionaries. Upon a extraction of basic keyword corpus, words from standard dictionaries can easily be integrated (with frequencies set to the infinite) to introduce more generality in the original corpus data. The effect of such integration is examined in later through IR evaluations.

Table 3: The ratio of common terms against the frequency.

frequency more than N	num. of keyword pairs	num. of common pairs	ratio
1	51062	3074	0.06
2	10990	1999	0.18
5	2916	1062	0.36
10	1216	606	0.50
20	538	326	0.61
50	176	127	0.72
100	68	58	0.85
200	24	24	1.00

3.2 Analyzing clustering result

The noise in the keyword corpus can be categorized into either of the following groups (the examples shown for each group are the English correspondings to the Japanese term "情報検索 (*information retrieval*)"):

Table 4: Errors detected at the first iteration of partitioning.

total errors detected	951
(1) spelling error	0
(2) expressional variations	0
(3) related keywords	598
(4) obvious errors	326
(5) correct	27

(1) spelling errors
example: information retreival

(2) expressional variations
example: information retrieving

(3) related keywords
example: information seeking

(4) obvious errors
example: keyword

When generating keyword clusters, it is advantageous to maintain spelling errors and expressional variations in view of recall, while it is important to eliminate obvious errors in view of precision. The treatment of related keywords should depend on the database and also the context of the search.

We analyze manually the errors detected at the first iteration of the cluster partitioning, i.e. the minimum cut links with the capacity equal to 1. The result is summarized in Table 4. Of the total 951 pairs detected as errors, 326 are obvious errors, 598 are related but not exactly corresponding, and 27 are detected wrongly though they actually are the correct ones. It is remarkable that regardless of the considerable number of spelling errors and expressional variations included in the database, non of them are detected as errors. This may be because such errors or variations occur most likely with low frequency and seldom are repeated in different clusters.

Among 598 related pairs, 136 have hierarchical relationship while the rest are the ones simply correlated. Among 27 detection errors, 10 are caused by homonymous keywords that are not acronyms, 8 by mis-processed acronyms, and 9 by minor examples that errors actually occur more frequently than correct pairs in the corpus.

The result shows that the major improvement can be expected by refining conditions for links elimination after minimum cuts detection stage. Our present implementation use only graph-topological conditions, N_α and N_β, common for all the clusters. Better results may be obtained by changing the value depending on the cluster size or by applying natural language processing. Since the number of candidates are greatly reduced by minimum cuts detection stage, the disambiguation of errors and non-errors can be more time consuming.

In the following, a few examples of the generated keyword clusters are shown where the number in the parenthesis indicates the frequency of the keyword. Example 1 shows the keywords in the biggest cluster with their frequencies ≥ 3. We can observe the closely related Japanese and English keywords are clustered together.

Example 1: *Frequent keywords in the largest cluster.*
Japanese: 並列処理 *(740)*, 並列 *(62)*, 並列化 *(56)*, 並列計算 *(29)*, 並行処理 *(19)*, 並列性 *(10)*, 並列システム *(6)*, 多重処理 *(6)*, 並行システム *(5)*, マルチプロセス *(5)*, 並列プロセス *(4)*, 並列度 *(3)*, 多重プロセス *(3)*, マルチプロセシング *(3)*, パラレル処理 *(3)*
English: *parallel processing(672), parallel(74), parallelization(44), concurrent processing (20), parallel computing(18), parallel computation(18), parallelism(14), parallel process(8), multi process(8), concurrent system(7), parallel system(6), parallel processings(6), multi processing(6), multiprocessing(5), multiprocess(5), concurrent(5), future(4), parallelize(4), parallel processes(4), parallel operation(4), paralell processing(4)*

Example 2 contains two clusters which are originally a single cluster but separated since each of the clusters has sufficient frequency in the target corpus. Again, we show only the keywords with their frequencies ≥ 3 on

account of the limited space.

Example 2 : *Associated clusters devided into two.*
(1) CLUSTER 1
Japanese: 知的 *cai(118), ITS(67),* 知的教育システム *(31), ICAI(11),* 知的教授システム *(8),* 指導方略 *(7),* 教授戦略 *(7),* 問題演習 *(4),* 定式化 *(4),* 教育戦略 *(4),* 対象理解 *(3),* 教育効果 *(3)*
English: *ITS(68), intelligent cai(65), intelligent tutoring system(45), ICAI(30), intelligent educational system(9), tutoring strategy(8), teaching strategy(8), IES(7), teaching paradigm(3), object understanding(3), intelligent tutoring systems(3)*
(2) CLUSTER 2
Japanese: *CAI(156),* 教育支援システム *(27),* 教育支援 *(21),* 学習支援 *(13),* 学習支援システム *(8), cai*システム *(7), CAL(6),* 表層構造 *(4),* 派生語 *(4),* 学習支援環境 *(4),* コンピュタ支援教育 *(3)*
English: *CAI(169), computer assisted instruction(27), computer aided instruction(8), CAL(7), education support system(6), derivative(4), cai system(4), surface structure(3), learning support(3), education suport(3), computer assisted learning(3)*

Example 3 shows another but smaller cluster with all the Japanese and English keywords included. It can be observed that minor translation examples are integrated into more frequent cases.

Example 3 : *An output including minor keywords.*
Japanese: *CAD(246),* 設計支援 *(63),* 計算機援用設計 *(14),* 計算機支援設計 *(9),* コンピュタ支援設計 *(8),* コンピュタ援用設計 *(4),* コンパイルエラ *(4),* 設計工学 *(3),* キャド *(3),* 設計エンジニアリング *(2),* 計算機設計支援 *(2),* ソフトウェア設計支援 *(2),* 知能 *(1),* 自動組立 *(1),* 三面図理解 *(1),* 航空機主翼設計 *(1),* 計算援用設計 *(1),* 演算器シェア *(1),* コンピュタデザイン *(1)*
English: *CAD(225), computer aided design(70), design support(36), design engineering(6), software design support(4), design assistance(2), cad(2), support to plan(1), support method for design(1), software design suppo'rt(1), prulog(1), planning aids(1), operator resource sharing(1), design = support(1), design suppore system(1), design suppont(1), design suppert(1), design suport(1), design assist(1), desighn support(1), computer cided design(1), computer aidid design(1), computer aided disign(1), computer aided desigh(1), compile time errors(1), compile error(1), co'mputer aided design(1), assisted design(1), DAD(1)*

3.3 Query Expansion in Cross-lingual Information Retrieval

Lastly, We examine the effectiveness of the generated clusters on the two retrieval tasks:

(1) CLIR: Japanese queries retrieving documents from an English collection (J-E task), and

(2) monolingual IR: Japanese queries retrieving documents from a Japanese collection (J-J task).

The search performance is tested against the test version of the NACSIS Test Collection 1 (Kando et al, 1998) for the J-E task E collection, which contains 186,809 documents, and for the J-J task, J collection, 338,668 documents.

We indexed Japanese terms by character (uni-gram), and English terms by word. English terms appeared in Japanese texts were also indexed by word. Queries are submitted as Japanese natural language sentences. They were initially segmented into words using a Japanese morphological analyzer, Chasen v1.5 (Matsumoto, et al., 1997). Words and phrases were then automatically selected asn query terms using several patterns defined over part-of-speech tags.

Each query term was translated or expanded using the bilingual keyword clusters reported here. We treated terms in the cluster, containing the query term, as synonyms. In order to examine the effect of standard dictionary integration and also partitioning parameter N, we tested these strategies listed below; K3: keyword clusters obtained with $N = 3$, KD3: keyword and dictionary term clusters obtained with $N = 3$, KD10: keyword and dictionary term clusters obtained with $N = 10$, D3: dictionary terms in KD3, and D10: dictionary tems in KD10. The average numbers of translated or expanded terms per a query term are shown in Table 5.

The search engine is OpenText 6, which can handle both English and Japanese char-

Table 5: Average number of expanded terms per a query.

	K3	KD3	KD10	D3	D10
J-E	6.81	4.48	3.66	0.84	0.77
J-J	10.7	7.01	5.72	1.4	1.3

acters. The documents in the returned set are ranked using OpenText's "RankMode Relevance1".

The retrieval results for the J-E task are summarized in Figure 4 where each plot represents the search effectiveness compared with the *baseline*, which is obtained by applying the original Japanese query terms to the Japanese correspondings of the target English documents.

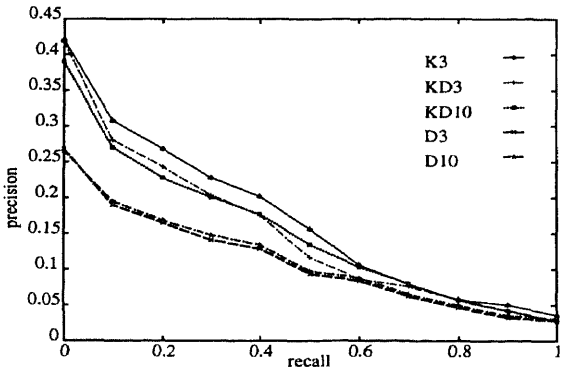

Figure 4: Retrieval results of J-E task.

For the J-E task, the keyword clusters (K3) showed the highest effectiveness, then followed by the keyword and dictionary term clusters (KD3 and KD10). The clusters consisted of dictionary terms only (D3 and D10) achieved less than 65% of the K3. The number of translated terms seems to be an important index to estimate the search effectiveness in the J-E task. No significant difference was found regarding the minimum edge cut levels of 3 and 10 in the keyword and dictionary clusters (KD3, KD10), but some improvement is shown in the dictionary clusters (D3, D10).

The retrieval results for the J-J task are summarized in Figure 5 where each plot represents the 11 point average search effectiveness over the *baseline*, which in this case is the performance without query expansion.

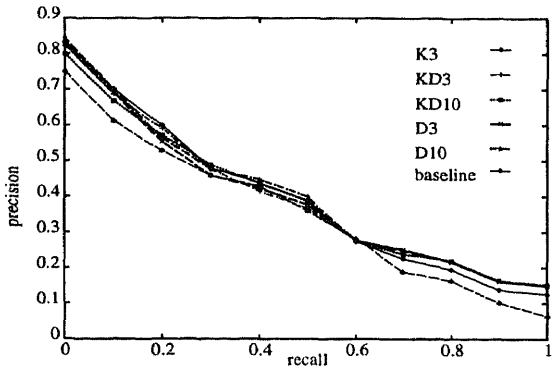

Figure 5: Retrieval results of J-J task.

For the J-J task, all the strategies showed the improvement of search effectiveness of 11.3-13.9% in the average over the baseline. Though the average numbers of expanded terms were significantly small in the dictionary clusters (D3, D10), they showed as same as, or even more improvement of search effectiveness over the baseline. Again, there were no significant difference according the clustering parameter in the keyword and dictionary clusters (KD3 and KD10); Some improvement is observed in the dictionary only clusters (D3 and K10).

Based on the results, we can temporarily conclude that the bilingual clusters generated by our method showed significant improvement of search effectiveness both in CLIR and monolingual IR. In CLIR, the bilingual clusters of author-assigned keywords (K3) were more effective than the ones with dictionary terms, while in monolingual IR, adding dictionary terms to the clusters (KD10) showed improvement. It is also observed that the clustering parameter N affects the performance in some cases, suggesting that fur-

ther improvement can be expected by refining clustering conditions. The IR experiments described here will be explained more in detail (Kando & Aizawa, 1998).

4 Discussions

The keyword clusters generated by our method can also be utilized in automatic indexing such as LSI (latent semantic indexing) to reduce the dimension of the frequency matrix by term clustering. Also, the clusters generated by our graph-based method but using only dictionary terms are utilized in domain map visualization task to assist interactive document retrieval (Aihara & Takasu, 1998).

Though our present implementation use only graph-topological information, we are now looking for the possibility of incorporating natural language processing. For example, only keyword level correspondences are considered so far but given that many keywords are complex, we can expect better performance by utilizing morpheme (or word) level correspondences. This will also give a way to weight the bilingual pairs according to their centrality within a cluster and to control the granularity of the generated cluster.

Acknowledgement

The research reported here is a part of the research project "A Study on Ubiquitous Information System for Utilization of Highly Distributed Information Resources", granted by the Japan Society for the Promotion of Science.

References

[1] Grefenstette, G., Smeaton, A. and Sheridan, P. (eds.) (1996) *Workshop on Cross-Linguistic Information Retrieval.*

[2] Kando, N. (1997) *Cross-linguistic scholarly information transfer and database services in Japan.* Presented at the Annual Meeting of the American Society for Information Science.

[3] Aizawa, A. and Kageura, Kyo (1998) *An Approach to the Automatic Generation of Multilingual Keyword Clusters.* COMPTERM'98.

[4] Dunning, T. and Davis, M. (1993) *Multilingual information retrieval.* Technical report MCCS-93-252, Computer Research Laboratory, New Mexico State University.

[5] Landauer, T.K. and Littman, M.L. (1990) *Fully automatic cross-language document retrieval.* In Proceedings of the Sixth Conference on Electronic Text Research, p.31-38.

[6] Carbonell, J.G., Yang, Y., Frederking, R.E., Brown, R.D., Geng, Y. and Lee, D. (1997) *Translingual information retrieval: a comparative evaluation.* In Proceedings of the 15th International Joint Conference on Artificial Intelligence (IJCAI-97), p.708-714.

[7] NACSIS (1997) *Introduction to the National Center for Science Information Systems.* NACSIS.

[8] Aiso, H. (ed.) (1993) *Joho Syori Yogo Daijiten.* Tokyo: Ohm.

[9] Japan Society for Artificial Intelligence (ed.) (1990) *Jinko Tinou Handobukku.* Tokyo: Ohm.

[10] Ralston, A. (ed.) (1983) *Encyclopedia of Computer Science and Engineering.* Amsterdam: Van Nostrand Reinhold. [Toujou, A. (trans. ed.) *Compyu-ta Daihyakka.* Tokyo: Asakura. 1987.]

[11] Shapiro, S (ed.) (1987) *Encyclopedia of Artificial Intelligence.* New York: John Wiley. [Ohsuga, S. (trans. ed.) *Jinko Tinou Daijiten.* Tokyo: Maruzen. 1991.]

[12] Kando,N., Koyama,T., Oyama,K., Kageura,K., Yoshioka,M., Nozue,T., Matsumura,A. and Kuriyama,K. (1998) *NTCIR : NACSIS Test Collection Project [Poster].* the 20th Annual BCS-IRSG Colloquium on Information Retrieval Research.

[13] Matsumoto, Y. et al. (1997) *Japanese Morphological Analyzer Chasen 1.5.* NAIST.

[14] Kando, N. and Aizawa, A. (1998) *Cross-Lingual Information Retrieval using Automatically Generated Multilingual Keyword Clusters* IRAL'98 (submitted).

[15] Aihara, K. and Takasu, A. (1998) *Domain Visualization Based on Authorized Documents* SCI '98/ISAS '98.