Bidirectional Extraction of Phrases for Expanding Queries in Academic Paper Retrieval

Yuzana Win
Graduate School of Engineering
Nagasaki University
Nagasaki, Japan

Tomonari Masada
Graduate School of Engineering
Nagasaki University
Nagasaki, Japan

***Abstract*—This paper proposes a new method for query expansion based on bidirectional extraction of phrases as word *n*-grams from research paper titles. The proposed method aims to extract information relevant to users' needs and interests and thus to provide a useful system for technical paper retrieval. The outcome of proposed method are the trigrams as phrases that can be used for query expansion. *First*, word trigrams are extracted from research paper titles. *Second*, a co-occurrence graph of the extracted trigrams is constructed. To construct the co-occurrence graph, the direction of edges is considered in two ways: *forward* and *reverse*. In the forward and reverse co-occurrence graphs, the trigrams point to other trigrams appearing after and before them in a paper title, respectively. *Third*, Jaccard similarity is computed between trigrams as the weight of the graph edge. *Fourth*, the weighted version of PageRank is applied. Consequently, the following two types of phrases can be obtained as the trigrams associated with the higher PageRank scores. The trigrams of the one type, which are obtained from the forward co-occurrence graph, can form a more specific query when users add a technical word or words before them. Those of the other type, obtained from the reverse co-occurrence graph, can form a more specific query when users add a technical word or words after them. The extraction of phrases is evaluated as additional features in the paper title classification task using SVM. The experimental results show that the classification accuracy is improved than the accuracy achieved when the standard TF-IDF text features are only used. Moreover, the trigrams extracted by the proposed method can be utilized to expand query words in research paper retrieval.**

Keywords*—word n-grams; Jaccard similarity; PageRank; TF-IDF; query expansion; information retrieval; feature extraction

I. Introduction

In these days, it is an important but complex task to get valuable information by searching the Web. With the rapid increase of information, users often perceive the difficulty of accessing the rich information resource effectively and of obtaining the information associated with their needs accurately. When users want to find the information relevant to their needs, they are required to find appropriate query words or phrases. However, the search results may not be relevant due to the inability of the queries to represent the needs accurately. Especially in academic paper retrieval, in many cases, users also want to find the papers focusing on specific and precise research topics, not general and vague topics. It can be considerably difficult for users to formulate a query for retrieving the papers discussing clear and specific topics. If the query contains only a single word, the search result consists of papers discussing a wide range of topics. That is, while the recall is high, the precision is low. If the query contains too many words, users may get only a limited number of academic papers as a search result. That is, while the precision is high, the recall is low. To overcome the above problems, the solution of this paper is to provide users with help in extracting from a large text set phrases that can be used to expand a less specific query. By expanding queries with the extracted phrases, users may get a search result containing a sufficient number of papers talking about specific research topics.

This paper proposes a new method for extracting important phrases as word *n*-grams from research paper titles. The extracted phrases are expected to be fruitful in query expansion for academic information retrieval. The proposed method is special in the following sense. The method extracts two types of phrases, each of which realizes a different query expansion, i.e., the expansion to the left and the expansion to the right. For example, the proposed method gives "a framework for" and "in sensor networks" as its outcome. The phrase "a framework for" can expand queries like "clustering", "classification", etc., to the left and give more specific queries like "a framework for clustering", "a framework for classification", etc. The phrase "in sensor networks" can expand queries like "clustering", "classification", etc., to the right and give more specific queries like "clustering in sensor networks", "classification in sensor networks", etc.

A brief explanation of the proposed method is given as follows. *First*, the proposed method extracts word trigrams as phrases that can be used for query expansion from a large number of research paper titles. There are two reasons why we focus on trigrams. The one reason is that, while word *n*-grams will be useful for text analysis, longer *n*-grams may cause data sparseness problem. Because the *n*-grams longer than three may be too long to obtain a sufficient large number of technical papers as a search result. The other is that unigrams and bigrams are too short to make a single word query express a specific and precise topic. *Second*, the proposed method builds a co-occurrence graph of the extracted trigrams. To construct the co-occurrence graph, the extracted word trigrams are used as nodes and the co-occurrence relations of trigrams appearing in the same paper titles as edges. Here, both the forward and reverse directions of edges are considered. In the forward co-occurrence graph, the trigram points to other trigrams appearing after it in a paper title. In the reverse co-occurrence graph, the trigram points to other trigrams appearing before it in a paper title. *Third*, the proposed method evaluates the Jaccard similarity for all co-occurring pair of trigrams and utilizes the similarity as the edge weight. And *fourth*, the proposed method applies a weighted version of PageRank on the forward and reverse co-occurrence graphs. As a result, we can get the top-ranked trigrams with reference to PageRank scores. Many of the top-ranked trigrams given from these two co-occurrence graphs can be regarded as important phrases. Details will be explained later.

Our first paper [18] describes a method for exploring technical phrase frames by extracting word *n*-grams. However, this paper introduces a new approach that applies weighted PageRank algorithm on the forward and reverse co-occurrence graphs of trigrams. The distinction between these two types of co-occurrence graphs does not appear in [18]. As a result, the two types of top-ranked trigrams are obtained. The performance of the extracted trigrams are evaluated as additional features in paper title classification using SVM. This evaluation is also not included in [18].

The remainder of this paper is divided into four sections. Section 2 describes the related work. Section 3 explains the proposed method. Section 4 contains the results of the evaluation experiment. The final section concludes the paper with discussion on future work.

II. Related Work

The extraction of important word sequences, e.g. keyphrases and key sentences, is relevant to our problem. There are two types of extraction, i.e., supervised [2], [6], [7], [9] and unsupervised methods [1], [3], [4], [8], [10], [11]. Natural language processing techniques [12], [13], [14] have also been used for keyphrase extraction.

Mihalcea [15] proposed an unsupervised method for automatic sentence extraction using graph-based ranking algorithms. The author used a text graph to represent the interconnection of words or other text entities with meaningful relations, ranked the entire sentences in weighted graphs manner, sorted in reversed order of their scores and selected the top ranked sentences for summary. The author evaluated the method in text summarization task. The experimental results show that graph-based ranking algorithms (HITS and PageRank) are useful for sentence extraction when applied to graphs extracted from texts.

Litvak et al. [17] analyzed two graph-based approaches, i.e., unsupervised and supervised ones, which enhance to extract keywords to be used in summarizing documents. The researchers built a graph to represent the co-occurrence in a window of a fixed number of words. They used HITS algorithm to get the top-ranked keyword and identified the keywords in order to generate the summarization. As a result, they argued that if a large number of summarized documents were available then supervised classification was the most accurate to identify the keywords in a document graph. Unless the number of summarized documents are large, unsupervised classification is better to extract the keywords in a graph.

Wan et al. [16] proposed CollabRank, a collaborative approach to single-document keyphrase extraction from multiple documents. They implemented the CollabRank to obtain document clusters by using the clustering algorithm. They used the graph-based ranking algorithm to extract the keyphrases within each document cluster. They built a graph based on all candidate words in the documents of the given cluster and evaluated the candidate phrases in the document based on the scores of the words contained in the phrases. Finally, they chose a few phrases with highest scores as the keyphrases of the document.

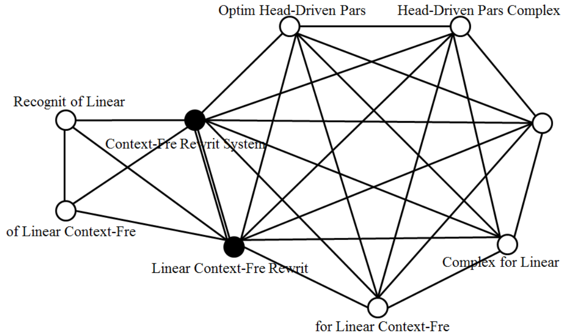


Fig. 1. A small portion of the co-occurrence graph

Contribution. This paper proposes a method that applies weighted PageRank algorithm on the forward and reverse co-occurrence graphs of trigrams. Consequently, the method can extract two different types of trigrams that can be used for query expansion: 1) Many of the trigrams obtained from the forward co-occurrence graph can form a more specific query when users add a word *before* them (e.g. "**clustering** for web search"); 2) Many of the trigrams obtained from the reverse co-occurrence graph can form a more specific query when users add a word *after* them (e.g. "automatic extraction of **clustering**"). This kind of bidirectional nature of extraction was not achieved by any of the PageRank-type methods described above.

III. The Proposed Method

In this section, the four steps of the proposed method are explained.

A. Word Trigrams

First, the proposed method extracts trigrams from a large set of research paper titles after applying stemming. For example, the proposed method extracts from the paper title "Recognition of Linear Context-Free Rewriting Systems" the following trigrams: "Recognit of Linear","of Linear Context-Fre", "Linear Context-Fre Rewrit", and "Context-Fre Rewrit System". Word trigrams are extracted by using the natural language toolkit for python (NLTK).

B. Co-occurrence Graph

The next step of the proposed method is to construct a co-occurrence graph of the extracted trigrams. In order to build the co-occurrence graph, the extracted word trigrams are used as nodes. When two trigrams appear in the same title, they are connected by an edge. Fig. 1 shows a small portion of the co-occurrence graph. This portion is obtained from the

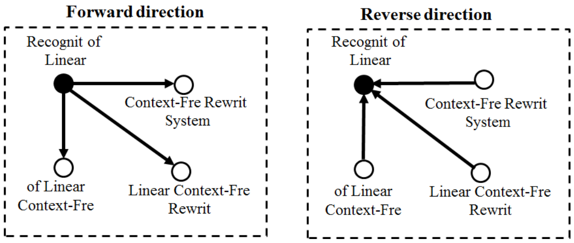


Fig. 2. Co-occurring pairs of trigrams according to forward and reverse directions

TABLE I. DATA SETS

Fields	*Venue*
NLP	ACL, EACL, COLING, CICLing, NAACL, IJCNLP, EMNLP, NLDB, TSD
DM	SIGMOD, VLDB, PODS, SIGIR, WWW, KDD, ICDE, ISWC, CIDR, ICDM, ICDT, EDBT, SDM, CIKM, ER, ICIS, SSTD, WebDB, SSDBM, CAiSE, ECIS, PAKDD
ALG	STOC, FOCS, ICALP, STACS, ISAAC, MFCS, FSTTCS, FCT, COCOON, CSR, WoLLIC
PRG	POPL, PLDI, ECOOP, OOPSLA, ISMM, ICLP, ICFP, CGO, ESOP, FOSSACS, CP, CC, LOPSTR, FLOPS, HOPL, AOSD

following two paper titles: “Recognition of Linear Context-Free Rewriting Systems” and “Optimal Head-Driven Parsing Complexity for Linear Context-Free Rewriting Systems”.

Further, the direction of edges is specified according to the order of trigrams. The direction of edges is determined in two ways: *forward* and *reverse* directions, as shown in Fig. 2. On the left panel of Fig. 2, the trigram “Recognit of Linear” points to the trigrams “of Linear Context-Free”, “Linear Context-Fre Rewrit”, and “Context-Fre Rewrit System”, because the latter three trigrams appear *after* the trigram “Recognit of Linear” in the paper title “Recognition of Linear Context-Free Rewriting Systems”. This direction is called *forward* direction. In contrast, on the right panel of Fig. 2, the same trigram “Recongnit of Linear” is pointed by the other three trigrams. In this case, each trigram points to the trigrams appearing *before* it. This direction is called *reverse* direction. According to the forward and reverse directions of edges, the two co-occurrence graphs, i.e., forward co-occurrence graph and reverse co-occurrence graph, can be constructed.

C. Jaccard Similarity

In the third step, the Jaccard similarity is evaluated for all co-occurring pairs of trigrams and the similarity is utilized as the edge weight. Let (t_1,t_2) denote a pair of trigrams whose similarity is to be calculated. Let $S(t_i)$ denote the set of paper titles that contain the trigram t_i. The Jaccard similarity is computed between two trigrams t_1 and t_2 as follows:

$$sim(t_1, t_2) = \frac{|S(t_1) \cap S(t_2)|}{|S(t_1) \cup S(t_2)|} \qquad (1)$$

After assigning the Jaccard similarity to each edge, a weighted version of PageRank algorithm is applied. The survey paper [5] analyzed many binary similarity measures. There are two reasons why we compute the Jaccard similarity. The first one is that it is simple to compute. The second one is that the Jaccard similarity is measured with the exclusion of *negative matches* [5]. In our approach, negative matches are related to the research paper titles where both of the trigrams under consideration do not appear and are not that important.

D. Weighted PageRank Algorithm

The last step of the proposed method applies weighted PageRank algorithm on both forward and reverse co-occurrence graphs of the extracted trigrams. Let P(t_i) denote the PageRank scores of the trigram t_i. Let w_{ji} denote the weight assigned to the edge connecting the two co-occurring pairs of nodes, t_i and t_j. w_{ji} is set to the corresponding Jaccard similarity. Then the PageRank score of the trigrams is calculated t_i by applying the Eq. (2) as below:

$$P(t_i) = \frac{1-d}{N} + d \times \sum_{t_j \in M(t_i)} \frac{w_{ji}}{\sum_{t_k \in M(t_j)} w_{jk}} P(t_j) \qquad (2)$$

where $M(t_i)$ denotes the set of nodes which point to t_i and N is the total number of extracted trigrams. The parameter d is the damping factor that is usually set to 0.85. $\sum_{t_k \in M(t_j)} w_{jk}$ is the sum of the weights assigned to each neighbor t_k in $M(t_j)$. Intuitively, if a node is pointed by many high-scored neighbors, the node may get a high score. However, the proposed method combines the Jaccard similarity and weighted PageRank algorithm. Therefore, if a node is pointed by many high-scored neighbors with large Jaccard similarities, then the node may obtain a high score.

IV. EXPERIMENTAL RESULTS

A. Evaluation in Text Classification

The trigrams extracted by the proposed method were evaluated as additional features in the paper title classification task. We used SVM (Support Vector Machine) for classification and checked whether the trigrams extracted by the proposed method improved the classification accuracy.

The proposed method was tested in the binary classification of the paper titles obtained from DBLP (Digital Bibliography & Library Project) [1]. Each DBLP record included a list of authors, title, conference name or journal name, year, page numbers, etc. Academic conferences were chosen in the four fields: Natural Language Processing (NLP), Data Management (DM), Algorithms and Theory (ALG), and Programming Languages (PRG). We only selected top conferences and used the research paper titles presented in the conferences shown in Table I. As a result, the total number of paper titles contained in NLP, DM, ALG, PRG data sets are 10,666, 27,573, 16,468, and 9,434, respectively. In the preprocessing, stop words were removed and Porter Stemmer [2] was used to stem words to their root forms.

Classification was conducted on the four data sets, i,e., NLP paper titles, DM paper titles, ALG paper titles and PRG paper titles. From these four data sets, six different pair of data sets were obtained as ALG_PRG, DM_ALG, DM_PRG, NLP_ALG, NLP_DM and NLP_PRG. For each pair, the data was randomly split into 90% of the paper titles for training and 10% for testing, and the classification was performed

[1] http://www.dblp.com/
[2] http://www.tartarus.org/martin/PorterStemmer/

with SVM. The accuracies were averaged over the ten results obtained from the 10-fold cross-validation.

TF-IDF term weighting was used to compose a feature vector for each paper title based on the formula: $\text{tf\_idf}(t, d) = \text{tf}(t, d) \times \log(N/\text{df}(t))$, where $\text{tf}(t, d)$ is the frequency of term t in document d, and $\text{df}(t)$ is the document frequency of t, i.e., the number of documents where t appears. N is the total number of documents in the corpus. In the experiment, the TF in TF-IDF was modified by using the trigrams obtained by the proposed method to improve the classification accuracy. First, we find the trigram having the largest PageRank score in each paper title. For each paper title d, the set of the three words of the trigram is denoted having the largest PageRank score by $W(d)$. Then, $\text{weight}(t, d)$ is used, defined by Eqs. (3) and (4) in place of $\text{tf}(t, d)$:

$$\text{weight}(t, d) \equiv \alpha \times \text{tf}(t, d) + 1 \text{ for } t \in W(d) \tag{3}$$

$$\text{weight}(t, d) \equiv \alpha \times \text{tf}(t, d) \text{ for } t \notin W(d) \tag{4}$$

In Eqs. (3) and (4), α is the term reweighting parameter and is chosen as an integer. All word counts are increased by a factor of α and then the word counts are increased by one only for the words of the trigram having the largest PageRank score (cf. Eq. (3)). For example, we assume that the trigram “probabilist inform flow” has the largest PageRank score among the trigrams appearing in the paper title “Decidability of Parameterized Probabilistic Information Flow” after stemming. Then only the counts of the three words “probabilist”, “inform”, and “flow” are increased by one after we increase the counts of all words by a factor of α. If the trigrams extracted by the proposed method are important in the sense that they are closely related to a particular research topic and thus help discriminating the research topic from other topics, the reweighting described above may improve the classification accuracy.

SVM was trained with linear kernel by setting $C = 1.0$ and the classification accuracy was obtained in terms of Area Under the ROC curve (AUC). The term reweighting parameter α was varied from 3 to 27, and the mean and standard deviation of AUC in the 10-fold cross validation were recorded.

Tables II and III summarize the p-values obtained by comparing the standard TF-IDF (i.e., TF-IDF without modification of TF) and the TF-IDF based on the TF modified by Eqs. (3) and (4) in terms of AUC. The p-values are obtained in a paired two-sided t-test. If the classification accuracy of the proposed method is not as high as the frequency-based method, the p-value is assigned with a minus symbol. When the p-value is less than 0.05, we can say that the improvement is statistically significant and thus give the p-value in bold in Tables II and III. The results of Table II are given by using the trigrams obtained from the forward co-occurrence graph. On the other hand, the results of Table III are given by using trigrams obtained from the reverse co-occurrence graph. Tables II and III show the term reweighting factor α yielding the best p-values on each data set. Only for the two pairs, i.e., DM_PRG and NLP_PRG, in Table III, we could not get a statistically significant improvement. For all remaining cases in Tables II and III, we could get a significantly better accuracy than the standard TF-IDF. Based on these results, it can be said that the classification accuracy is improved by modifying the TF in TF-IDF by the trigrams the proposed method gives. So we claim that the proposed method can extract the features that are useful in discriminating different research topics as the trigrams having large PageRank scores.

TABLE II. p-VALUES FOR PAIRED T-TEST ON ROC CURVE AUC (FORWARD)

Data sets	***Standard TF-IDF***	***Modified TF-IDF***	***p-value***
ALG_PRG	0.942075	0.942493 (α = 3)	**0.009**
DM_ALG	0.978106	0.978225 (α = 8)	**0.021**
DM_PRG	0.971507	0.971669 (α = 10)	**0.029**
NLP_ALG	0.989345	0.989452 (α = 8)	**0.003**
NLP_DM	0.954356	0.954432 (α = 27)	**0.048**
NLP_PRG	0.985577	0.985633 (α = 19)	**0.047**

TABLE III. p-VALUES FOR PAIRED T-TEST ON ROC CURVE AUC (REVERSE)

Data sets	***Standard TF-IDF***	***Modified TF-IDF***	***p-value***
ALG_PRG	0.942075	0.943616 (α = 9)	**0.013**
DM_ALG	0.978106	0.978163 (α = 8)	**0.020**
DM_PRG	0.971507	0.971566 (α = 23)	0.052
NLP_ALG	0.989345	0.989422 (α = 12)	**0.016**
NLP_DM	0.954990	0.954883 (α = 20)	-0.053
NLP_PRG	0.985899	0.985959 (α = 22)	**0.041**

B. Comparing with Frequency-based Trigram Extraction

To discuss the special nature of the trigrams extracted by the proposed method, we compared the proposed method with a simple method for the extraction of trigrams, i.e., the frequency-based extraction. In the frequency-based method, the same data sets were used and the same preprocessing were applied as in the proposed method. Then, the number of occurrences, i.e., frequency, were counted for every trigram, and the higher-ranked trigrams based on their frequencies were obtained. The difference between two methods are clarified by displaying examples.

Tables IV and V summarize the trigrams obtained by the frequency-based method and by the proposed method for ALG and DM data sets, respectively. For example, “the complex of” and 381 in the top cell of the left column of Table IV mean that the frequency is 381 for the trigram “the complex of”. Moreover, “and relat problem” and 6.05 $\times 10^{-4}$ in the top cell of the middle column of Table IV mean that the PageRank is 6.05 $\times 10^{-4}$ for the trigram “and relat problem” in the forward co-occurrence graph. Furthermore, “on the complex” and 19.59 $\times 10^{-4}$ in the top cell of the right column of Table IV mean that the PageRank is 19.59 $\times 10^{-4}$ for the trigram “on the complex” in the reverse co-occurrence graph.

We can observe that many trigrams obtained from the forward co-occurrence graph can expand queries to the right. For example, the trigram “in web search” can expand the queries like “ranking” and “queries” to give more specific queries like “ranking in web search” and “queries in web search”. On the other hand, many trigrams obtained from the reverse co-occurrence graph can expand queries to the left. For example, the trigram “efficient algorithm for” can expand the queries like “computing” and “mining” to give more specific queries like “efficient algorithm for computing” and “efficient algorithm for mining”. This is a remarkable feature of the proposed method. In contrast, the frequency-based method

TABLE IV. TOP-10 (STEMMED) TRIGRAMS OF ALG

Frequency		***PageRank***			
		Forward ($\times$ ***10***$^{-4}$)		***Reverse*** ($\times$ ***10***$^{-4}$)	
the complex of	381	and relat problem	6.05	on the complex	19.59
lower bound for	259	in polynomi time	5.52	the complex of	19.23
approxim algorithm for	225	in linear time	5.46	lower bound on	8.89
algorithm for the	209	and it applic	4.10	approxim algorithm for	8.69
on the complex	162	term rewrit system	3.75	lower bound for	8.12
the power of	103	in the plane	3.72	a note on	7.20
with applic to	100	constraint satisfact problem	3.24	bound on the	7.09
bound on the	91	in planar graph	3.05	effici algorithm for	6.84
lower bound on	81	of complex class	2.37	the power of	6.13
effici algorithm for	73	and their applic	2.34	on the power	5.76

TABLE V. TOP-10 (STEMMED) TRIGRAMS OF DM

Frequency		***PageRank***			
		Forward ($\times$ ***10***$^{-4}$)		***Reverse*** ($\times$ ***10***$^{-4}$)	
a case studi	229	on the web	4.71	the impact of	8.06
the role of	219	for inform retriev	2.91	the effect of	6.56
the impact of	216	in inform retriev	2.68	the role of	4.41
a framework for	201	a case studi	2.62	a framework for	3.39
the effect of	184	in web search	2.50	a comparison of	3.04
for inform retriev	140	in social network	2.33	the influenc of	2.83
the case of	140	an empir studi	2.03	a studi of	2.71
in inform system	137	inform retriev system	2.00	the use of	2.06
on the web	134	an exploratori studi	1.94	effici process of	1.85
of inform system	123	in social media	1.90	an analysi of	1.84

TABLE VI. p-VALUES FOR PAIRED T-TEST ON ROC CURVE AUC (FORWARD)

Data sets	***Frequency-based***	***Proposed***	***p-value***
ALG_PRG	0.94904 (α = 5)	0.94928 (α = 3)	0.058
DM_ALG	0.98194 (α = 4)	0.98202 (α = 2)	0.546
DM_PRG	0.97613 (α = 2)	0.97622 (α = 2)	0.333
NLP_ALG	0.99140 (α = 5)	0.99148 (α = 4)	**0.031**
NLP_DM	0.96251 (α = 2)	0.96276 (α = 2)	0.079
NLP_PRG	0.98699 (α = 5)	0.98706 (α = 4)	0.336

TABLE VII. p-VALUES FOR PAIRED T-TEST ON ROC CURVE AUC (REVERSE)

Data sets	***Frequency-based***	***Proposed***	***p-value***
ALG_PRG	0.94245 (α = 4)	0.94256 (α = 3)	0.689
DM_ALG	0.97828 (α = 5)	0.97822 (α = 5)	−0.384
DM_PRG	0.97235 (α = 3)	0.97214 (α = 3)	−0.195
NLP_ALG	0.98942 (α = 5)	0.98941 (α = 5)	−0.774
NLP_DM	0.95450 (α = 6)	0.95422 (α = 6)	**−0.024**
NLP_PRG	0.98563 (α = 3)	0.98548 (α = 5)	−0.054

cannot give these two types of trigrams separately, because all trigrams are mixed in the same ranking, as shown in the left columns of Tables IV and V.

Further, we can observe that the frequency-based ranking tends to provide trigrams having a general meaning like “lower bounds for”, “the power of”, “with applications to”, “bounds on the”, “lower bounds on”, “a case study”, “the case of”, etc., where the original form is recovered from the root form of each word. In contrast, the proposed method tends to provide trigrams having a specific meaning, e.g. like “in polynomial time ”, “in linear time”, “in planar graphs”, “term rewriting systems”, “constraint satisfaction problems”, “of complexity classes”, “information retrieval system”, “an empirical study”, etc., with respect to the forward co-occurrence graph. Also with respect to the reverse co-occurrence graph, many trigrams given by the proposed method have at least as specific a meaning as the trigrams given by the frequency-based method. Therefore, it can be said that, at least with respect to the forward co-occurrence graph, the top-ranked trigrams obtained by the proposed method have a more specific meaning than those obtained by the frequency-based method.

However, it is possible that the proposed method may degrade the quality of the extracted trigrams by providing them in two separate rankings. Therefore, we compared the proposed method with the frequency-based method also in text classification task described in Section IV-A. We also used SVM (Support Vector Machine) for classification and checked if the trigrams extracted by the proposed method were as useful as the trigrams extracted by the frequency-based method.

To obtain the best classification accuracy in terms of Area Under the ROC curve (AUC), SVM was trained with two different kernels, namely linear kernel by setting $C = 1.0$ and rbf (Radial Basis Function) kernel by setting $C = 2.0$ and $gamma = 2.0$. We selected the term reweighting parameter α yielding the best case from each kernel and recorded the mean and standard deviation of AUC in the 10-fold cross validation.

Tables VI and VII summarize the p-values obtained by comparing the frequency-based method and the proposed method based on the TF modified by Eqs. (3) and (4) in terms of AUC. The p-values are obtained in a paired two-sided t-test. The p-value is assigned with a minus symbol if the classification accuracy of the proposed method is not as high as the frequency-based method. When the p-value is less than 0.05, it can be said that the improvement is statistically significant. The results of Table VI are given by using the trigrams obtained from the forward co-occurrence graph, where SVM is trained by using the rbf kernel. On the other hand, the results of Table VII are given by using the trigrams obtained from the reverse co-occurrence graph, where SVM is trained by using the linear kernel. For all but one case in Tables VI and VII, we could get as good an accuracy as the frequency-based method. We could not get a comparable accuracy only for the NLP_DM data set pair in Table VII. Consequently, the result showed that the proposed method at least could extract as effective trigrams as the frequency-based method. It can be said that the bidirectional nature of the proposed method is an extra gain, which cannot be achieved by the frequency-based method.

C. A Possible Application: Query Expansion

Based on the experimental results, it can be said that the trigrams extracted by the proposed method represent technical research topics well. We here discuss how such trigrams can be used in query expansion for information retrieval.

For example, as presented in Fig. 3, the query word "clustering" can be expanded to the right by the trigrams "in sensor networks", "for web search", "for text categorization", etc., which are obtained by the proposed method from the forward co-occurrence graph. These trigrams can be used for the *right* expansion in this manner, because their first word (i.e., "for","in","of", etc.) is a function word that mainly follows a noun. As we discussed in Section IV-B, the trigrams obtained by the proposed method tend to represent a specific meaning, especially with respect to the forward co-occurrence graph. Therefore, we may expect that the search results obtained by the queries expanded in this manner will relate to specific research topics. Fig. 4 gives another example. The query word "clustering" is expanded to the left by the trigrams "a framework for", "automatic extraction of", "efficient algorithm for", etc., which are obtained from the reverse co-occurrence graph. These trigrams can be used for the *left* expansion, because their last word (i.e., "for", "of", etc.) is a function word that is mainly followed by a noun.

It should be noted that a similar expansion cannot be straightforwardly achieved by the trigrams obtained by the frequency-based method, because the trigrams that can be used for the right expansion and those that can be used for the left expansion are mixed in the same ranking as shown in the left columns of Tables IV and V. However, the proposed method provides two types of trigrams in two different rankings, as shown in the middle and right columns of Tables IV and V.

We here verify how the search results obtained by the expanded queries can focus on more specific research topics. Fig. 5 shows the three types of search results obtained from Google Scholar. Fig. 5(a) gives the search results for the query "clustering". Fig. 5(b) gives the search results obtained by the proposed method from the forward co-occurrence graph. Fig. 5(c) gives the search results obtained by the proposed method from the reverse co-occurrence graph. As presented in Fig. 5(a), we can get the search result having a general meaning when we only input a single query word "clustering". For example, the topics like "Algorithms for clustering data" and "A comparison of document clustering techniques" tend to provide a general meaning consisting of the words like "algorithms" and "techniques". These words tend to represent a wide range of topics. Consequently, the single query word has not exploited users' needs and interests, but users can't get relevant topics when each user has a specific need.

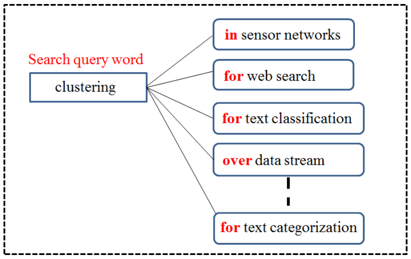


Fig. 3. Example of possible right expansions of the query word 'clustering' by using the trigrams obtained from the forward co-occurrence graph for DM data set

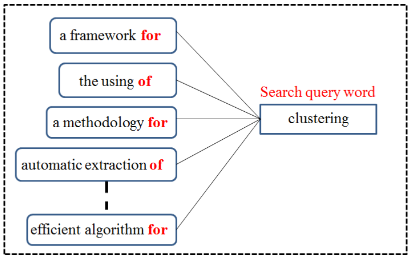


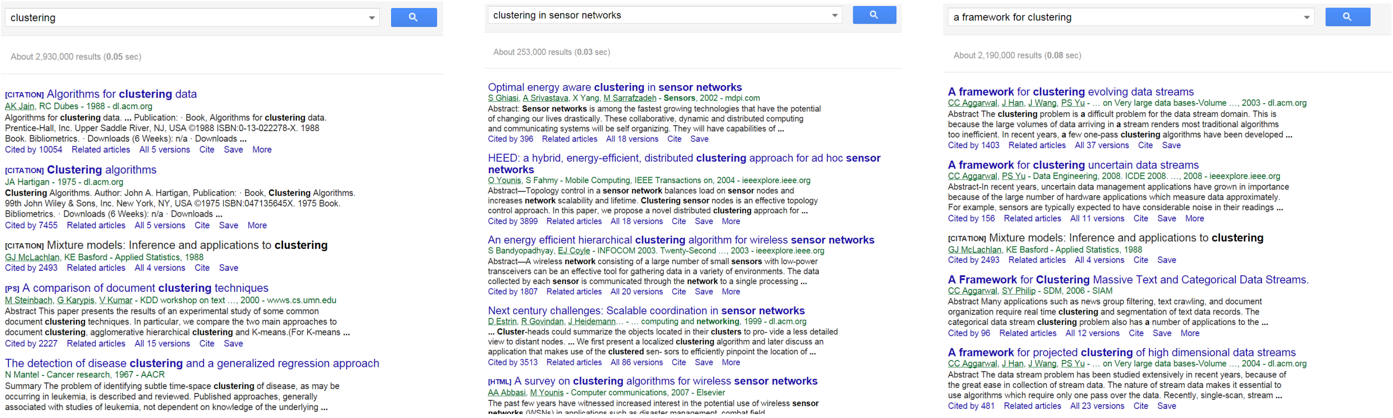
Fig. 4. Example of possible left expansions of the query word 'clustering' by using the trigrams obtained from the reverse co-occurrence graph for DM data set

In contrast, when a single query word "clustering" is expanded to the right by the phrase "in sensor networks", we can get the search results focusing on more specific topics as shown in Fig. 5(b). Most of the words or phrases appearing in the search results, e.g., "hybrid", "ad hoc", "hierarchical", and "wireless", have a specific meaning. On the other hand, when a single query word "clustering" is expanded to the left by the phrase "a framework for", we can also get the search results focusing on more specific topics as shown in Fig. 5(c). Some of the words or phrases occurring in the search results, e.g., "data streams", "high dimensional", and "Text and Categorical", represent a specific meaning.

Therefore, it is found that the query expansion, i.e., the expansion to the left and the expansion to the right, can give the search results relating to specific topics. Further, we can get two different types of search results due to the bidirectional nature of the proposed method. These results are more specific when we expand query words than when we only use a single query word. We can observe that the proposed method works as a new query expansion scheme more oriented toward actual user needs and interests for informational retrieval.

V. Conclusion

In this paper, we proposed a new method for query expansion based on bidirectional extraction of phrases. The proposed



clustering

About 2,930,000 results (**0.05** sec)

[CITATION] Algorithms for **clustering** data
AK Jain, RC Dubes - 1988 - dl.acm.org
Algorithms for **clustering** data. **...** Publication: · Book, Algorithms for **clustering** data. Prentice-Hall, Inc. Upper Saddle River, NJ, USA ©1988 ISBN:0-13-022278-X. 1988 Book. Bibliometrics. · Downloads (6 Weeks): n/a · Downloads **...**
Cited by 10054 Related articles All 5 versions Cite Save More

[CITATION] **Clustering** algorithms
JA Hartigan - 1975 - dl.acm.org
Clustering Algorithms. Author: John A. Hartigan, Publication: · Book, **Clustering** Algorithms. 99th John Wiley & Sons, Inc. New York, NY, USA ©1975 ISBN:047135645X. 1975 Book. Bibliometrics. · Downloads (6 Weeks): n/a · Downloads **...**
Cited by 7455 Related articles All 5 versions Cite Save More

[CITATION] Mixture models: Inference and applications to **clustering**
GJ McLachlan, KE Basford - Applied Statistics, 1988
Cited by 2493 Related articles All 4 versions Cite Save

[PS] A comparison of document **clustering** techniques
M Steinbach, G Karypis, V Kumar - KDD workshop on text …, 2000 - wwws.cs.umn.edu
Abstract This paper presents the results of an experimental study of some common document **clustering** techniques. In particular, we compare the two main approaches to document **clustering**, agglomerative hierarchical **clustering** and K-means.(For K-means **...**
Cited by 2227 Related articles All 15 versions Cite Save

The detection of disease **clustering** and a generalized regression approach
N Mantel - Cancer research, 1967 - AACR
Summary The problem of identifying subtle time-space **clustering** of disease, as may be occurring in leukemia, is described and reviewed. Published approaches, generally associated with studies of leukemia, not dependent on knowledge of the underlying **...**

(a)

clustering in sensor networks

About 253,000 results (**0.03** sec)

Optimal energy aware **clustering** in **sensor networks**
S Ghiasi, A Srivastava, X Yang, M Sarrafzadeh - **Sensors**, 2002 - mdpi.com
Abstract: **Sensor networks** is among the fastest growing technologies that have the potential of changing our lives drastically. These collaborative, dynamic and distributed computing and communicating systems will be self organizing. They will have capabilities of **...**
Cited by 396 Related articles All 18 versions Cite Save

HEED: a hybrid, energy-efficient, distributed **clustering** approach for ad hoc **sensor networks**
O Younis, S Fahmy - Mobile Computing, IEEE Transactions on, 2004 - ieeexplore.ieee.org
Abstract—Topology control in a **sensor network** balances load on **sensor** nodes and increases **network** scalability and lifetime. **Clustering sensor** nodes is an effective topology control approach. In this paper, we propose a novel distributed **clustering** approach for **...**
Cited by 3899 Related articles All 18 versions Cite Save More

An energy efficient hierarchical **clustering** algorithm for wireless **sensor networks**
S Bandyopadhyay, EJ Coyle - INFOCOM 2003. Twenty-Second …, 2003 - ieeexplore.ieee.org
Abstract—A wireless **network** consisting of a large number of small **sensors** with low-power transceivers can be an effective tool for gathering data in a variety of environments. The data collected by each **sensor** is communicated through the **network** to a single processing **...**
Cited by 1807 Related articles All 20 versions Cite Save More

Next century challenges: Scalable coordination in **sensor networks**
D Estrin, R Govindan, J Heidemann… - … computing and **networking**, 1999 - dl.acm.org
... Cluster-heads could summarize the objects located in their **clusters** to pro- vide a less detailed view to distant nodes. **...** We first present a localized **clustering** algorithm and later discuss an application that makes use of the **clustered** sen- sors to efficiently pinpoint the location of **...**
Cited by 3513 Related articles All 86 versions Cite Save More

[HTML] A survey on **clustering** algorithms for wireless **sensor networks**
AA Abbasi, M Younis - Computer communications, 2007 - Elsevier
The past few years have witnessed increased interest in the potential use of wireless **sensor networks** (WSNs) in applications such as disaster management, combat field

(b)

a framework for clustering

About 2,190,000 results (**0.08** sec)

A framework for **clustering** evolving data streams
CC Aggarwal, J Han, J Wang, PS Yu - … on Very large data bases-Volume …, 2003 - dl.acm.org
Abstract The **clustering** problem is **a** difficult problem for the data stream domain. This is because the large volumes of data arriving in **a** stream renders most traditional algorithms too inefficient. In recent years, **a** few one-pass **clustering** algorithms have been developed **...**
Cited by 1403 Related articles All 37 versions Cite Save

A framework for **clustering** uncertain data streams
CC Aggarwal, PS Yu - Data Engineering, 2008. ICDE 2008. …, 2008 - ieeexplore.ieee.org
Abstract-In recent years, uncertain data management applications have grown in importance because of the large number of hardware applications which measure data approximately. For example, sensors are typically expected to have considerable noise in their readings **...**
Cited by 156 Related articles All 11 versions Cite Save More

[CITATION] Mixture models: Inference and applications to **clustering**
GJ McLachlan, KE Basford - Applied Statistics, 1988
Cited by 2493 Related articles All 4 versions Cite Save

A Framework for **Clustering** Massive Text and Categorical Data Streams.
CC Aggarwal, SY Philip - SDM, 2006 - SIAM
Abstract Many applications such as news group filtering, text crawling, and document organization require real time **clustering** and segmentation of text data records. The categorical data stream **clustering** problem also has **a** number of applications to the **...**
Cited by 96 Related articles All 12 versions Cite Save More

A framework for projected **clustering** of high dimensional data streams
CC Aggarwal, J Han, J Wang, PS Yu - … on Very large data bases-Volume …, 2004 - dl.acm.org
Abstract The data stream problem has been studied extensively in recent years, because of the great ease in collection of stream data. The nature of stream data makes it essential to use algorithms which require only one pass over the data. Recently, single-scan, stream **...**
Cited by 481 Related articles All 23 versions Cite Save

(c)

Fig. 5. Example of the search results for the query (a) ‘clustering’ (b) ‘clustering in sensor networks’ and (c) ‘a framework for clustering’

method extracted important phrases as trigrams based on a procedure consisting of four processing steps. The trigrams extracted by the proposed method were evaluated as additional features in the paper title classification task using SVM. The experimental results showed that the accuracy was improved. We also compared the trigrams given by the proposed method with those given by the frequency-based method. According to the experimental results, the proposed method could provide as good trigrams as the frequency-based method. However, the proposed method has an extra gain, i.e., the bidirectional nature of trigrams extraction, which cannot be achieved by the frequency-based method. Further, we discussed how we could use such trigrams for query expansion. A search system using this type of query expansion can give search results relating to specific topics.

We have a future plan to perform a quantitative evaluation of the search results obtained by the query expansion based on the proposed method in information retrieval task.

Acknowledgment

This work has been supported by the Grant-in-Aid for enhancement of engineering higher education of the Japan International Cooperation Agency (JICA) from 2014 to 2017. We are grateful for their support.

References

[1] K.S. Hasan and V. Ng, “Conundrums in unsupervised keyphrase extraction: making sense of the start-of-the-art,” in *Proc. of the 23rd International Conference on COLING 2010*, Beijing, pp. 365–373, August 2010.

[2] C. Caragea, F. Bulgarov, A. Godea, S.D. Gollapalli, “Citation-enhanced keyphrase extraction from research papers: a supervised approach,” in *Proc. of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 1435-1446, Doha, Qatar, October 2014.

[3] R. Mihalcea, P. Tarau and E. Figa, “PageRank on semantic networks with application to word sense disambiguation,” in *Proc. of the 20th International Conference on Computational Linguistics*, Stroudsburg, PA, USA, no. 1126, August 2004.

[4] S.D. Gollapalli and C. Caragea, “Extracting keyphrases from research papers using citation networks,” in *Proc. of the 28th AAAI Conference on Artificial Intelligence*, pp. 1629–1635, June 2014.

[5] S. Choi, S. Cha, and C.C. Tappert, “A survery of binary similarity and distance measures,” *J.Syst. Cybern, Inf.*, vol. 8, no. 1, pp. 43–48, 2010.

[6] D. X. Wang, X. Gao, and P. Andreae, “Automatic keyword extraction from single-sentence natural language queries,” in *Proc. of the 12th Pacific Rim International Conference on Artificial Intelligence*, Kuching, Malaysia, pp. 637–648, September 2012.

[7] K. Zhang, H. Xu, J. Tang, and J. Li, “Keyword extraction using support vector machine,” in *Proc. of the 7th International Conference on WAIM*, Hong Kong, China, pp. 85–96, June 2006.

[8] T. Nomoto and Y. Matsumoto, “A new approach to unsupervised text summarization,” in *Proc. of SIGIR2001*, pp. 26–34, 2001.

[9] M. R. Amini and P. Gallinari, “The use of unlabeled data to improve supervised learning for text summarization,” in *Proc. of SIGIR2002*, 105-112, 2002.

[10] S. N. Kim and M. Kan, “Re-examining automatic keyphrase extraction approaches in scientific articles,” in *Proc. of 2009 Workshop on Multiword Expressions, ACL-IJCNLP 2009*, pp. 9-16, Suntec, Singapore, August 2009.

[11] Z. Liu, P. Li, Y. Zheng and M. Sun, “Clustering to find exemplar terms for keyphrase extraction,” in *Proc. of 2009 Conference on Empirical Methods in Natural Language Processing*, pp. 257-266, Singapore, August 2009.

[12] T. Tomokiyo and M. Hurst, “A language model approach to keyphrase extraction,” in *Proc. of ACL 2003 Workshop on Multiword Expressions: Analysis, Acquisition and Treatment*, vol. 18, pp. 33–40, 2003.

[13] K. Barker and N. Cornacchia, “Using noun phrase heads to extract document keyphrases,” in *Proc. of 13th Biennial Conference of the Canadian Society for Computational Studies of Intelligence*, pp. 40–52, May 2000.

[14] S. N. Kim, T. Baldwin and M. Kan, “Evaluating n-gram based evaluation metrics for automatic keyphrase extraction,” in *Proc. of 23rd international conference on COLING 2010*, pp. 572-580, Beijing, August 2010.

[15] R. Mihalcea, “Graph-based ranking algorithms for sentence extraction, applied to text Summarization,” in *Proc. of ACL 2004 on Interactive Poster and Demonstration Sessions*, Article no. 20, Stroudsburg, PA, USA, July 2004.

[16] X. Wan and J. Xiao, “CollabRank: towards a collaborative approach to single-document keyphrase extraction,” in *Proc. of 22nd International Conference on COLING 2008*, pp. 969–976, Manchester, August 2008.

[17] M. Litvak and M. Last, “Graph-based keyword extraction for single-document summarization,” in *Proc. of 08 MMIES on COLING 2008*, pp. 17-24, Manchester, August 2008.

[18] Y. Win and T. Masada, “Exploring technical phrase frames from research paper titles,” in *Proc. of 29th IEEE International Conference on WAINA-2015*, pp. 558–563, Gwangju, South Korea, 2015.