The current issue and full text archive of this journal is available at
www.emeraldinsight.com/0264-0473.htm

An approach to discover and recommend cross-domain bridge-keywords in document banks

Yu-Min Su
Department of Computer Science, National Chengchi University, Taipei, Taiwan
Ping-Yu Hsu
Department of Business Administration, National Central University, Taoyuan, Taiwan, and
Ning-Yao Pai
Institute of Information Management, National Chiao Tung University, Hsinchu, Taiwan

Received 21 May 2009
Revised 20 August 2009
Accepted 22 September 2009

Abstract

Purpose – The co-word analysis method is commonly used to cluster-related keywords into the same keyword domain. In other words, traditional co-word analysis cannot cluster the same keywords into more than one keyword domain, and disregards the multi-domain property of keywords. The purpose of this paper is to propose an innovative keyword co-citation approach called "Complete Keyword Pair (CKP) method", which groups complete keyword sets of reference papers into clusters, and thus finds keywords belonging to more than one keyword domain, namely bridge-keywords.

Design/methodology/approach – The approach regards complete author keywords of a paper as a complete keyword set to compute the relations among keywords. Any two complete keyword sets whose corresponding papers are co-referenced by the same paper are recorded as a CKP. A clustering method is performed with the correlation matrix computed from the frequency counts of the CKPs, for clustering the complete keyword sets. Since keywords may be involved in more than one complete keyword set, the same keywords may end up appearing in different clusters.

Findings – Results of this study show that the CKP method can discover bridge-keywords with average precision of 80 per cent in the *Journal of the Association for Computing Machinery* citation bank during 2000-2006 when compared against the benchmark of Association for Computing Machinery Computing Classification System.

Originality/value – Traditional co-word analysis focuses on co-occurrence of keywords, and therefore, cannot cluster the same keywords into more than one keyword domain. The CKP approach considers complete author keyword sets of reference papers to discover bridge-keywords. Therefore, the keyword recommendation system based on CKP can recommend keywords across multiple keyword domains via the bridge-keywords.

Keywords Databases, Data handling, Information retrieval, Cluster analysis

Paper type Research paper

1. Introduction

With the popularity of networking, a variety of documents and information are stored in online data banks. The documents can range from personal blogs to academic

The Electronic Library
Vol. 28 No. 5, 2010
pp. 669-687

0264-0473
DOI 10.1108/02640471011081951

papers and company profiles. Up-to-now, one of the most frequently used techniques to find information on the network including the internet and enterprise intranets is keyword search. Users search for information by placing predicted keywords and based on the results, iteratively tune the keywords entered. Therefore, tracking proper keywords in related fields are vital activities for users. However, the tasks tend to be daunting if not impossible. Therefore, a keyword recommendation system that can actively offer users proper related keywords after each querying is urgently needed. As a result, Google, Yahoo and Amazon all provide keyword recommendation mechanisms along with their search engines.

Most systems offer the service by recommending other keywords in the same domain to refine the search. A keyword domain is a set of related keywords that are grouped by a similarity-based methodology. Keywords located in the same keyword domain with the original query result are recommended to reduce the number of documents retrieved. The similarity-based approach is widely adopted by commercial search engines and web sites.

The approach assumes users know exactly what domains they are looking for and only need a guide to fine tune the search research. However, for users who are new to a field, the assumption is not entirely correct. Many users who are new to a field may only guess the possible keywords which may in fact lead to wrong domains. For instance, a rooky graduate student trying to forecast option premiums with "neural network" may use the two words as key words to place a search. The search result from Google recommends ten possible extension keywords for further search. These recommendations include "neural network java", "rbf neural network", "neural network wiki", etc. The result is shown in Figure 1. These extensions are designed to fine tune search on sub-areas of the original domain. On the other hand, "genetic algorithm" is a rising alternative research tool for this purpose and may also be worth the student's investigation. However, the student her/himself has no way of learning it with the traditional keyword recommendations. Therefore, a system that can refer users to keywords in other related domains and not just keywords in the same domain would be a useful contribution.

However, to the best of our knowledge, no related works have been proposed to recommend-users with related keyword domains. In this study, a technology needed to develop such a system is proposed. The technology is a keyword clustering method, which can compose the keyword domains and identify the bridge-keywords at the

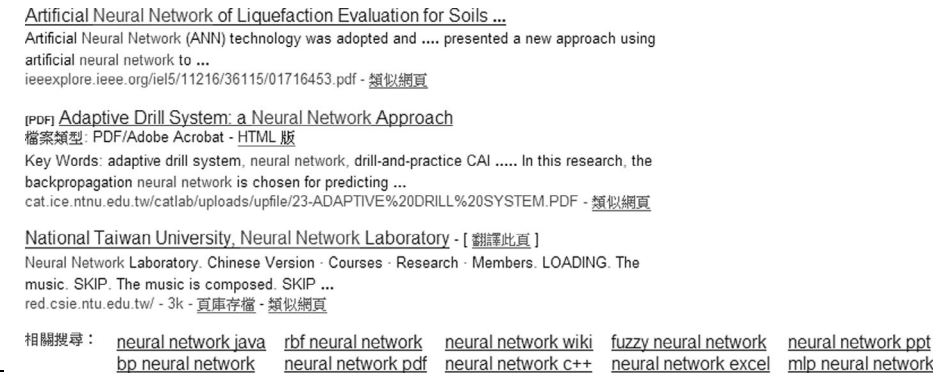

Figure 1.
The keywords recommended by Google

same time. Bridge-keywords are the keywords which are shared by more than one keyword domain. Traditional keyword co-occurrence analysis, namely co-word analysis, can only cluster a keyword into one keyword domain (Ding *et al.*, 2000; Lorence and Abraham, 2006; Whittaker *et al.*, 1989), and disregards the multi-domain property of keywords. The constraint therefore renders traditional co-occurrence analysis inappropriate for the purpose.

Figure 2 is designed to illustrate the idea of keyword domains and bridge-keywords. Domain 1 contains the keywords of neuron, neural network, self-organizing map, back propagation and soft computing, whereas domain 2 contains mutation, crossover, chromosome, gene, genetic algorithm and soft computing. Soft computing is a bridge-keyword of domain 1 and 2 since it is shared by both domains.

A citation data bank of source papers published in the *Journal of the Association for Computing Machinery* (*JACM*) from 2000 to 2006 is built to show that there are many keywords crossing more than one domain and the proposed clustering method can catch them with reasonable accuracy. Association for Computing Machinery (ACM) Computing Classification System (CCM) trees are used as a benchmark in this experiment. Results of this experiment show that the method can discover bridge-keywords that cross two keyword domains with average precision of 80 per cent and average recall of 76 per cent computed against the benchmark of ACM Computing Classification System (CCS).

The rest of this paper is organized as follows. Section 2 illustrates the complete keyword pair (CKP) recommendation system proposed in this study. Section 3 introduces related works in keyword recommendation systems. Section 4 then defines data structure and presents the CKP method, along with demonstration of a synthetic example. Next, Section 5 performs the CKP method in the real-life *JACM* citation bank against the benchmark of the ACM CCS, and analyzes the results of the empirical experiment. Conclusions are finally drawn in Section 6, along with future work.

2. Prototyping of CKP keyword recommendation system

With the keyword recommendation system, when users submit a query with a keyword, they receive a number of search results along with recommended keywords for further search. A sound keyword recommendation system can shorten searching time, assist users to make a decision and motivate their next oriented search.

Figure 3 and 4 derived from the clustering result shown in Figure 2 present the prototype demonstration of the CKP cross-domain keyword recommendation system,

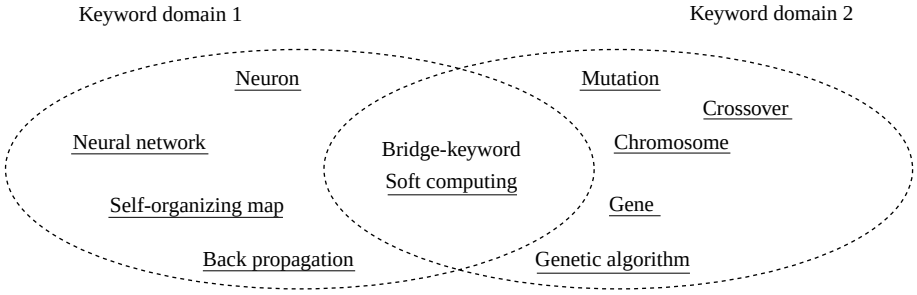

Figure 2.
An example of bridge-keyword shared by two domains

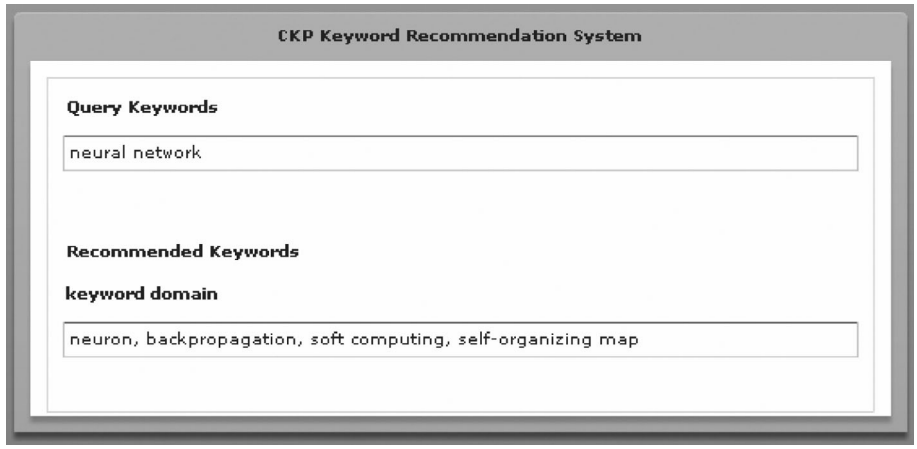

Figure 3.
Keywords suggested by CKP keyword recommendation system

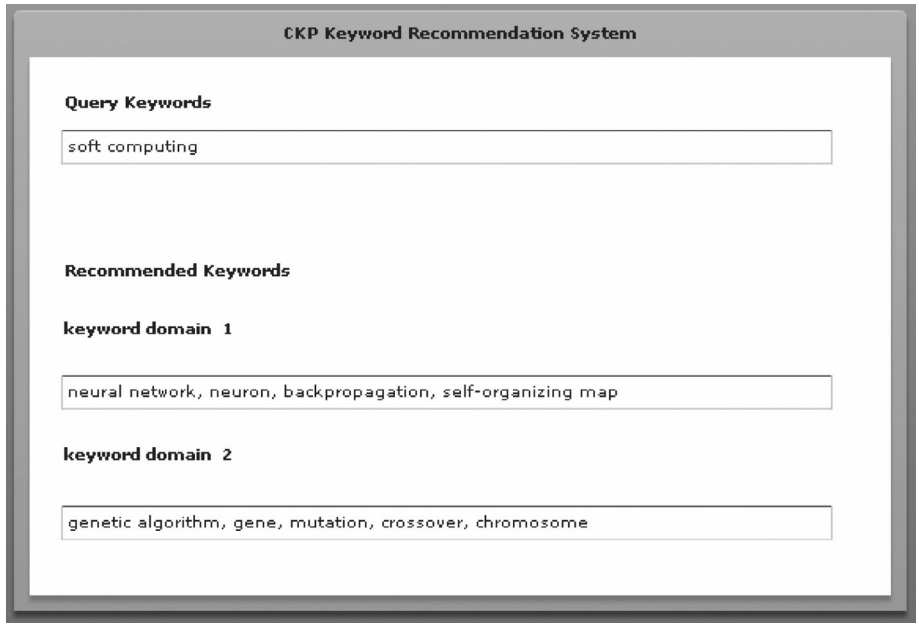

Figure 4.
CKP keyword recommendation system suggests keywords from two domains

which can automatically suggest query keywords out of the original query domain. Figure 3 shows that a user queries "neural network" and then the CKP keyword recommendation system returns the list of recommended keywords "neuron, back propagation, soft computing, self-organizing map".

If the user selects and clicks the recommended keyword "soft-computing", then the recommendation system returns two lists of recommended keywords shown in Figure 4, since the new query "soft-computing" is a bridge-keyword which is shared by domains 1 and 2. When the user clicks the recommended keyword "genetic algorithm"

in the list of keyword domain 2, only the keywords in the keyword domain 2 are returned. The CKP keyword recommendation system utilizes a keyword clustering method which can cluster a keyword into more than one cluster. Since the clustering method is the main novel technology deployed, the remaining of the paper will be devoted to discuss the technology.

3. Related works

Thesauri are traditionally applied to support user searching, but have several significant drawbacks. First, thesauri are edited by various domain experts. Building and amending thesauri are extremely knowledge intensive and time consuming. Second, it quickly becomes out of date due to accelerated knowledge expansion through the internet popularity and the rise of Web 2.0. Finally, difficulties in searching by thesauri are largely owing to the lack of semantic clustering and linkages between relevant keywords (Chen and Lynch, 1992). To remedy the problem, other approaches have been developed (Ding *et al.*, 2000). The state-of-the-art keyword recommendation technology proposed by Ding *et al.* (2000) and Kitamura *et al.* (1999) uses the co-word analysis method to recommend keywords according to their similarities with query keyword. The co-word method is developed based on the similarity of the keywords to compute distances among keywords.

Ding *et al.* (2000) proposed the visual bibliometric information retrieval system (BIRS) based on the co-word approach with bibliometric data bank retrieved from Science Citation Index and Social Science Citation Index covering the period 1987-1997. For query expansion, BIRS can recommend related keywords to users according to keyword similarities with a current query. With the keyword co-word analysis of BIRS, the keyword co-occurrence frequency matrix is created by counting the number of times two keywords occur together in the bibliography of the same papers. The co-occurrence frequency matrix is then converted into the Pearson's correlation coefficient matrix. Finally, a classical clustering method is employed to group keywords into different clusters. BIRS creates a visual keyword map to structure knowledge and concepts into keyword domains by computing the relationships among keywords. Ding *et al.* (2000) concluded that the associations of keywords identified by co-word analysis are different from those obtained by thesauri or experts. Furthermore, co-word analysis can incorporate the changes in domains to provide a timely keyword map for users (Ding *et al.*, 2000). Chen *et al.* (1997) noted that co-word analysis can become a significant tool to support traditional thesauri in generating search varieties (Chen *et al.*, 1997; Ding *et al.*, 2000).

Kitamura *et al.* (1999) proposed a prototype of hybrid keyword recommendation system using the biological data and user log data for GenBank, which is a well-known DNA-sequence database with over five million entries. The keyword recommendation system proposed by Kitamura *et al.* (1999) can actively recommend proper keywords to narrow down the current search based on the classical co-word method. The method computes and groups-related keywords employing co-occurring keyword pairs, which is two keywords those occur in the same entries of GenBank. The keyword recommendation system proposed by Kitamura *et al.* (1999) considers the biological entries together with user log data to recommend-users-related keywords, for incorporating expert knowledge with user search experiences (Kitamura *et al.*, 1999).

The similarity-based recommendation systems proposed by Ding *et al.* (2000) and Kitamura *et al.* (1999) group each keyword into one cluster only. In other words, the recommendation systems based on the co-word methodology cannot identify bridge-keywords which can appear in more than one keyword domain. Furthermore, the recommendation system proposed by Kitamura *et al.* (1999) considers user log data to improve user experiences, thus unfortunately increases update frequency and lengthens response time for the keyword recommendation system.

This study will develop a keyword recommendation system to recommend keywords across different keyword domains via bridge-keywords, based on a co-citation methodology, called the "complete keyword pair (CKP) method". The CKP method can group the same keywords into multiple keyword domains by employing the concept of regarding all author keywords of an article as a complete keyword set, thus the bridge-keywords are identified and incorporated in the CKP keyword recommendation system. The users of the keyword recommendation system based on the CKP method can agilely search among connected keyword domains by clicking the bridge-keywords.

4. Methodology: CKP approach

Traditional co-word analysis method is unable to cluster the same keywords into more than one keyword domain, even if these keywords have multi-domain property. This study proposes an improved co-citation method, the "Complete Keyword Pair" method (CKP method for short), which regards all keywords of one paper as a complete keyword set. Two keyword sets have a relation if their originated papers are cited by the same paper. The sets with high similarities are clustered together to form keyword domains.

The CKP method is composed of four steps. The first step is the creation of a co-citation frequency matrix from the frequency counts of pairs of any two complete keyword sets whose papers are referenced by the same source paper in a citation data bank. The second step is the creation of a Pearson's correlation matrix by computing the Pearson's correlations between any two complete keyword sets from the co-citation frequency matrix. In the third step, clusters are generated from complete keyword sets based on the correlation matrix. The last step in CKP is joining the complete keyword sets in each cluster to form keyword domains. Figure 5 shows the procedure of CKP approach in this study.

The set of keyword information is assumed to be stored at a citation bank, C, which is a set of citation records. Each citation record, c, is composed of a source paper, s, and a complete keyword set, W. The complete keyword set W denotes the author keyword list of a paper referenced by the source paper s. Therefore, for a set of papers published in journal J, the citation bank C is:

$$\{ < s, W > | s \in \text{papers published in } J, \text{ and } W \text{ is the complete keyword set of a paper which is referenced by } s \}.$$

The data structure used in the CKP method is defined as follows.

Definition 1. 1. Given a citation bank, C, a set of keywords, W, denotes a complete keyword set if $\exists s, < s, W > \in C$.

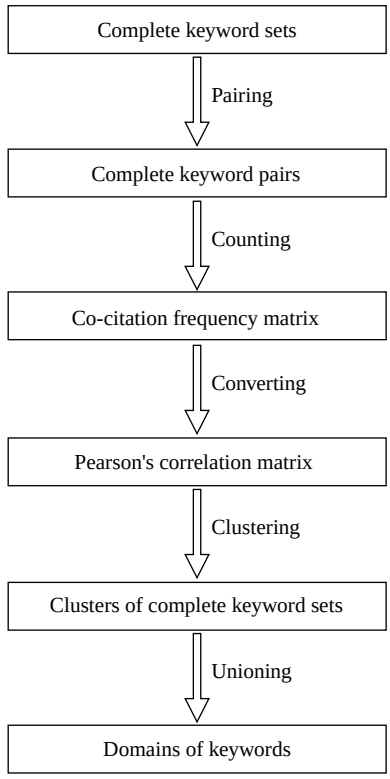

Figure 5.
Procedure of CKP approach

2. If there are N complete keyword sets in C, W_i and W_j are complete keyword sets, then $< W_i, W_j >$ is a CKP, where $0 \leqq i \leqq N-1$, $0 \leqq j \leqq N-1$.

Table I shows a sample citation bank, which contains three source papers with total of 12 citations annotated with four keyword sets. Table II lists all the complete keyword sets annotated with keyword set identifiers (KSIs).

4.1 Creation of co-citation frequency matrix

The frequencies of CKPs being co-cited are kept in the co-citation frequency matrix. The co-citation frequency matrix is defined as follows.

Definition 2. Given a citation bank, C, with N different complete keyword sets, if $< W_i, W_j >$ is a CKP, then the corresponding entry in the co-citation frequency matrix is given by:

$$F[i,j] = \begin{cases} |\{s| < s, W_i > \in C \wedge < s, W_j > \in C\}|, & 0 \leqq i \leqq N-1, 0 \leqq j \leqq N-1, \quad \text{if i} \neq j \\ 0, & 0 \leqq i \leqq N-1, 0 \leqq j \leqq N-1, \quad \text{if i} = j \end{cases} \tag{1}$$

The co-citation frequency matrix of the running example is shown in Table III. According to Definition 2, the $N \times N$ co-citation matrix is clearly symmetrical along

Citation	Source	Complete keyword set (W)	KSI
c01	s1	Soft computing, neural network, neuron, back propagation, self-organizing map	0
c02	s1	Soft computing, neural network, neuron, back propagation, self-organizing map	0
c03	s1	Soft computing, genetic algorithm, gene, chromosome, mutation, crossover	1
c04	s1	Bayesian classifier, conditional probability, independency, Bayesian network	2
c05	s1	Bayesian classifier, conditional probability, independency, Bayesian network	2
c06	s1	Bayesian classifier, conditional probability, independency, Bayesian network	2
c07	s2	Soft computing, neural network, neuron, back propagation, self-organizing map	0
c08	s2	Soft computing, neural network, neuron, back propagation, self-organizing map	0
c09	s2	Decision tree, entropy, gain	3
c10	s2	Decision tree, entropy, gain	3
c11	s3	Soft computing, neural network, neuron, back propagation, self-organizing map	0
c12	s3	Soft computing, genetic algorithm, gene, chromosome, mutation, crossover	1

Table I.
A sample citation bank

KSI	Complete keyword set (W)
0	Soft computing, neural network, neuron, back propagation, self-organizing map
1	Soft computing, genetic algorithm, gene, chromosome, mutation, crossover
2	Bayesian classifier, conditional probability, independency, Bayesian network
3	Decision tree, entropy, gain

Table II.
Complete keyword sets used in the sample citation bank

		j		
KSI	0	1	2	3
0	0	3	6	4
1 i	3	0	3	0
2	6	3	0	0
3	4	0	0	0

Table III.
Co-citation frequency matrix

the diagonal, and all diagonal cell values are zero. That is, the cell value of a CKP composed of two identical complete keyword sets is treated as zero in the co-citation matrix (He and Hui, 2000).

4.2 Creation of Pearson's correlation matrix

A Pearson's correlation matrix is computed from the co-citation frequency matrix. Pearson's correlation coefficient (Pearson's r) is commonly used as a similarity

measure in co-citation analysis (McCain, 1990). The similarity between any two complete keyword sets in the citation bank is determined by Pearson's r formula, and kept in the corresponding cell of the Pearson's correlation matrix. Table IV shows the correlation matrix derived from the co-citation frequency matrix of Table III. Clearly, the $N \times N$ correlation matrix is symmetrical along the diagonal, and all diagonal cell values are positive one. In other words, the Pearson's r of a CKP composed of two identical complete keyword sets is positive one. The Pearson's r is a normalized similarity measure that varies between positive one and negative one. A CKP with a high positive correlation coefficient means that the distance of the two coordinate complete keyword sets is quite close, whereas that of the two coordinate complete keyword sets is very far. Pearson's r formula is presented as follows (Johnson, 1988):

$$r = \frac{\Sigma XY - ((\Sigma X \Sigma Y)/N)}{\sqrt{(\Sigma X^2 - ((\Sigma X)^2/N))(\Sigma Y^2 - ((\Sigma Y)^2/N))}} \tag{2}$$

where X and Y denote two input vectors, and N denotes the dimension of the co-citation matrix. Indeed, N is the number of different complete keyword sets in the citation bank. If W_i and W_j are any two complete keyword sets in the citation bank, then the row X is the cell values for a complete keyword set W_i with each individual complete keyword set, and the row Y is the cell values for a complete keyword set W_j with each individual complete keyword set in the co-citation frequency matrix. Then, the Pearson's r, which denotes the similarity between the two complete keyword sets W_i and W_j, can be determined.

4.3 Generating clusters of complete keyword sets

A clustering method is applied to group complete keyword sets into clusters after the correlation matrix has been created. The predominant clustering method, namely K-means, is adopted to group complete keyword sets into clusters, since it is commonly and easily used, and produces more outstanding clustering results than other popular clustering methods, such as agglomerative hierarchical clustering (AHC). K-means is a partitioning approach, which first considers K prearranged initial cluster centers, assigning each object to its respective nearest cluster center. K cluster centroids are then computed as new cluster centers, reassigning each object to its respective nearest new cluster center iteratively until the members of each cluster no longer change.

The critical parameter "K" is set to the number of clusters required by the users when running the CKP method with K-means. However, AHC approaches can be performed in no presetting the number of clusters. AHC is a bottom-up approach, it begins by considering each object to be a cluster and then merges the most similar

		j		
KSI	0	1	2	3
0	+1.00000	−0.11547	−0.94002	−0.86667
1 i	−0.11547	+1.00000	+0.30151	+0.57735
2	−0.94002	+0.30151	+1.00000	+0.87039
3	−0.86667	+0.57735	+0.87039	+1.00000

Table IV. Pearson's correlation matrix

(or closest) pair of clusters iteratively until all objects are merged into one cluster. AHC can be used to decide the appropriate number of clusters, i.e. the parameter "K", for K-means clustering in the CKP method.

Table V shows the clustering result of the complete keyword sets in the sample citation bank when K is set to 3. The complete keyword sets W_2 and W_3 are grouped into the same cluster after clustering, meaning that these keywords belong to the same keyword domain.

4.4 Deriving keyword domains

The final step of CKP converts the clusters to keyword domains. Given a cluster of complete keyword sets, t, the keywords in the cluster represented by the domain, d, and d is computed as $\cup_{W \in t} W$. Table VI presents the three keyword domains based on the clustering result as shown in Table V. Significantly, "soft computing" belongs to two keyword domains. The bridge-keyword which crosses multiple domains is thus identified in the above CKP demonstration.

4.5 Reducing number of CKPs with keyword support threshold

If there are n keywords to be examined, the number of possible complete keyword sets is 2^n. The complexity of the algorithm is $O(2^n)$, since the Person's Correlation has 2^n entries. Huge number of CKPs can potentially be generated when computing correlations between any two complete keyword sets. To lessen the issue, a keyword support threshold is added to screen keywords whose frequency counts are below the threshold. Table VII shows keywords with their frequency counts in the running example.

If keyword support threshold is set at 3, then keywords "genetic algorithm, gene, chromosome, mutation, crossover, decision tree, entropy and gain" are deleted from the complete keyword sets in the sample data bank. If all keywords of a citation record are

Table V.
Keyword clustered by CKP in the sample citation bank

KSI	Complete keyword set (W)	Cluster (T)
0	Soft computing, neural network, neuron, back propagation, self-organizing map	Cluster 1 (t_1)
1	Soft computing, genetic algorithm, gene, chromosome, mutation, crossover	Cluster 2 (t_2)
2	Bayesian classifier, conditional probability, independency, Bayesian network	Cluster 3 (t_3)
3	Decision tree, entropy, gain	Cluster 3 (t_3)

Table VI.
Keyword members of each keyword domain

Keyword domain (D)	Keyword members
Keyword domain 1 (d_1)	Soft computing, neural network, neuron, back propagation, self-organizing map
Keyword domain 2 (d_2)	Soft computing, genetic algorithm, gene, chromosome, mutation, crossover
Keyword domain 3 (d_3)	Bayesian classifier, conditional probability, independency, Bayesian network, decision tree, entropy, gain

deleted then the record is also purged. Table VIII shows the results for the cleansed citation bank. Significantly, some complete keyword sets are shortened, and the total numbers of the complete keyword sets and pairs are reduced. Therefore, the CKP method can be sped up due to reducing the dimension of the co-citation matrix.

5. Experiments with references in *JACM*

This study builds a real-life citation bank for the experiment, namely, the *JACM* citation bank of source papers published in *JACM* from 2000 to 2006. The CKP method is performed in the citation data bank to discover the bridge-keywords which cross multiple keyword domains against the benchmark of ACM CCS. The citation data are described in detail later. The measures of precision and recall are adopted to evaluate the effectiveness of the CKP method in discovering the bridge-keywords. Finally, the results of the empirical experiment are analyzed and discussed.

5.1 Benchmark of effectiveness evaluation: ACM CCS

The effectiveness of the CKP method in discovering bridge-keywords is compared with the benchmark of ACM CCS. The ACM has established an online digital library (DL), called the ACM Portal. The ACM DL comprises two major parts, the DL and the Guide. The DL is a full-text repository of papers published by ACM and other publishers that

Table VII.
List of all keywords with frequency counts in the sample citation bank

Frequency	Keyword
7	Soft computing
5	Neural network, neuron, back propagation, self-organizing map
3	Bayesian classifier, conditional probability, independency, Bayesian network

Table VIII.
Sample bank after deleting infrequent keywords

Citation record (*c*)	Source paper (*s*)	Complete keyword set (*W*)	KSI
c01	s1	Soft computing, neural network, neuron, back propagation, self-organizing map	0
c02	s1	Soft computing, neural network, neuron, back propagation, self-organizing map	0
c03	s1	Soft computing	4
c04	s1	Bayesian classifier, conditional probability, independency, Bayesian network	2
c05	s1	Bayesian classifier, conditional probability, independency, Bayesian network	2
c06	s1	Bayesian classifier, conditional probability, independency, Bayesian network	2
c07	s2	Soft computing, neural network, neuron, back propagation, self-organizing map	0
c08	s2	Soft computing, neural network, neuron, back propagation, self-organizing map	0
c11	s3	Soft computing, neural network, neuron, back propagation, self-organizing map	0
c12	s3	Soft computing	4

have co-publishing or co-marketing agreements with ACM. The Guide is a collection of bibliographic citations and abstracts of papers published by major computing publishers, but it does not yet include bibliographic citations of all papers in the computing field (Association for Computing Machinery, 2009b).

Query results from the DL of the ACM Portal often include the classification terms of the ACM Computing Classification System (CCS). The ACM CCS has served for 20 years as the primary and most generally used system for classifying and indexing published computing literature. The ACM CCS has become a *de facto* standard for identifying and categorizing computing literature, as well as areas of computing interest and/or expertise (Association for Computing Machinery, 2009a). Therefore, this study adopted the scheme of the ACM CCS to evaluate the CKP effectiveness in discovering bridge-keywords.

The ACM CCS scheme involves a four-level tree that applies to specific computing areas, and a set of general terms that applies to general areas. The four-level tree involves three or two coded category levels of nodes, and an uncoded subject level of leaves. The subjects of leaves generally appear at the fourth level, sometimes at the third level. The ACM CCS hierarchy comprises 11 first-level categories, 81 second-level categories, 400 third-level categories (or subjects) and 981 fourth-level subjects. Additionally, 16 separate concepts called general terms apply to general areas (Association for Computing Machinery, 2009a). In the ACM Portal, the bibliographies of papers normally include three kinds of ACM CCS index terms, namely primary classification terms, additional classification terms and general terms. The primary classification terms are more relevant than the additional classification terms and the general terms to a paper article.

Only the reference papers of the *JACM* which have both author keywords and fourth-level subjects of the primary classification terms of the ACM CCS in the ACM portal are collected in the citation data bank. Table IX shows that the 545 citation papers with 2,849 author keywords derived from 143 source papers are collected in the *JACM* citation data bank.

5.2 Discovering bridge-keywords by CKP method against ACM CCS

The effectiveness of the CKP method in discovering bridge-keywords is benchmarked against the ACM CCS. The number of fourth-level subjects in each yearly citation data set derived from the ACM CCS is employed as the value of K, i.e. the preset number of clusters for K-means clustering in CKP. Table X shows the numbers of author keywords which are keywords selected by paper authors and the fourth-level subjects of the citation papers recorded in the *JACM* citation bank during 2000-2006. This table

Table IX.
Statistics of the *JACM* citation bank used in this experiment

Journal of source papers	*JACM*
Publication years of source papers	2000-2006
Number of source papers of the citation bank	143
Number of citation papers collected in the citation bank	545
Number of different citation papers in the citation bank	503
Total number of author keywords of all citation papers	2,849
Total number of different author keywords of all citation papers	1,865
Average number of author keywords per citation paper	5.28

reveals that the number of subjects (K) increases as the number of author keywords grows in the *JACM* citation data bank during 2000-2006 except in 2001.

Figure 6 shows both percentages of bridge-keywords discovered by the ACM CCS and the CKP method in the *JACM* citation data bank during 2000-2006. The ACM CCS-1 derived from the ACM CCS denotes the percentage of bridge-keywords which cross two or more subjects, whereas the ACM CCS-2 denotes the percentage of bridge-keywords which only cross two subjects. The CKP-1 derived from the CKP denotes the percentage of bridge-keywords which cross two or more keyword domains, whereas the CKP-2 denotes the percentage of bridge-keywords which only cross two keyword domains. Besides, both percentages rise as the value of K increases for the *JACM* citation bank. Obviously, most bridge-keywords only cross two specialized computing domains, as shown in Figure 6. Therefore, the evaluation of the CKP effectiveness in this experiment focuses on discovering the bridge-keywords across two keyword domains.

5.3 Measures and evaluation

Two measures, namely precision and recall, are used to evaluate the CKP effectiveness in discovering the bridge-keywords. The two measures are defined by the following formulae:

$$\text{Precision} = |A \cap B| \div |B| \tag{3}$$

$$\text{Recall} = |A \cap B| \div |A| \tag{4}$$

Year	2000	2001	2002	2003	2004	2005	2006
Number of author keywords	225	360	294	366	343	452	415
Number of subjects (K)	45	42	47	59	47	53	46

Table X.
Numbers of author keywords and subjects of the citation papers

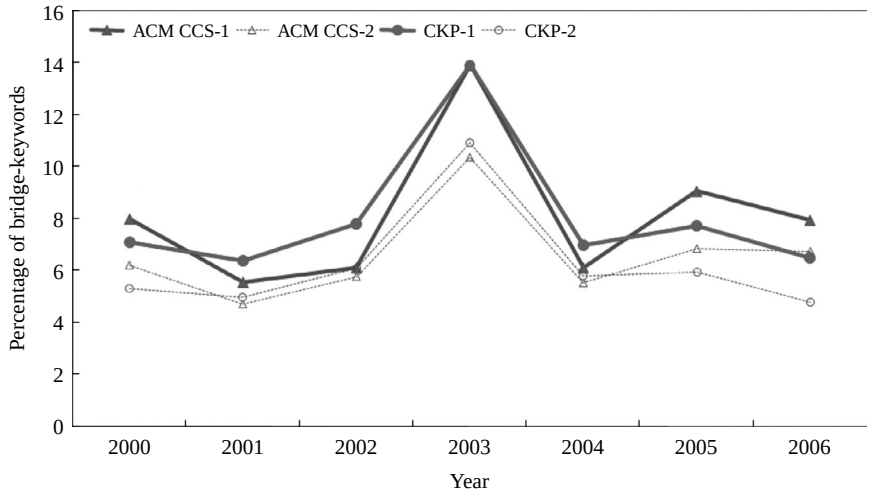

Figure 6.
Percentages of bridge-keywords discovered by ACM CCS and CKP

where A is a set of bridge-keywords across two specialized subjects identified by the ACM CCS, and B is a set of bridge-keywords across two keyword domains discovered by the CKP method.

Table XI shows the effectiveness of discovering bridge-keywords across two keyword domains by performing CKP in the *JACM* citation bank during 2000-2006. The precisions of CKP vary between 85 and 63 per cent, and the average precision is 80 per cent. The recalls of CKP vary between 89 and 61 per cent, and the average recall is 76 per cent.

The precisions and recalls have closely corresponding fluctuations in the *JACM* citation bank during 2000-2006 as shown in Figure 7. Furthermore, the precision and recall rise as the ratio of aggregated keyword counts to number of keywords increases.

5.4 Tuning parameter K in CKP

The step three of the CKP method, generating clusters of complete keyword sets, employs the clustering method to group the complete keyword sets. In the empirical experiment of this study, the well-known K-means method is elected as the clustering tool in the step three of the CKP. Relative to other clustering methods, K-means is easy to use, and excellent in clustering performance. The critical parameter "K" derived from classifying citation records by the ACM CCS subjects is used as the preset number of clusters for K-means. The K-means method is fitter than the statistical approaches for the experiment of the study since the real-life *JACM* citation data set after filtering is not plenty big enough to use statistical model. In other words, the

Table XI.
Effectiveness of CKP finding bridge-keywords based on K derived from ACM CCS

Year	2000	2001	2002	2003	2004	2005	2006	Average
CKP precision (%)	83	83	67	85	75	81	85	80
CKP recall (%)	71	88	71	89	79	71	61	76

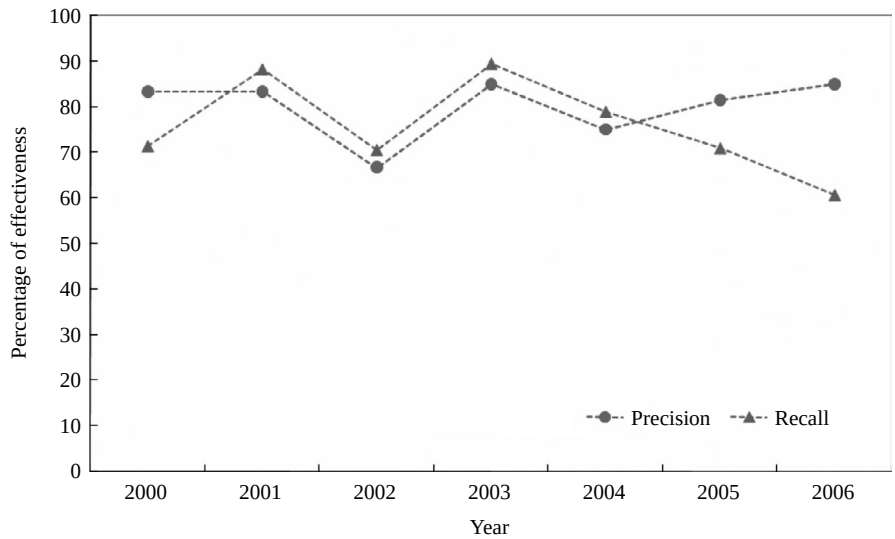

Figure 7.
Precisions and recalls have close relations

heuristic approach, e.g. K-means, is suitable for the empirical experiment in the data constraint. The measures of effectiveness evaluation in terms of the yearly precision and yearly recall in average during 2000-2006 to reduce the heuristic factor in the K-means employed in the CKP method. As the above-mentioned, the parameter "*K*" of the K-means clustering in the CKP method is subject to the citation records labeled by the *de facto* standard of ACM CCS. Therefore, the experiment results show that the CKP in discovering the bridge-keywords is provided with high effectiveness not owing to overfitting by tuning "*K*" or by accident. The hypothesis in this study is that the author keyword sets in the *JACM* citation bank is suitable to group into clusters with convex shape for K-means constraint in nature.

The alternative to decide the appropriate number of clusters, i.e. the parameter "*K*", is the AHC approach. The AHC allows of no presetting the number of clusters before clustering. The AHC merges the most similar (or closest) pair of objects iteratively until all objects are merged into one cluster. The AHC can be used to decide the appropriate number of clusters, namely parameter *K*, for K-means clustering in the CKP method. Table XII shows the appropriate numbers of clusters, i.e. *K*, in each *JACM* citation data set from 2000 to 2006, derived from the AHC approach. The AHC approach includes three major methods, namely, single linkage, complete linkage and average linkage. The appropriate values of *K*s derived from these AHC methods are the same in each citation set from 2000 to 2006, and very different from the values of *K*s derived from the ACM CCS, as shown in Table XII.

Table XIII shows the effectiveness of CKP in discovering bridge-keywords across two keyword domains based on *K* derived from AHC in each *JACM* citation data set from 2000 to 2006. The precisions of CKP based on K derived from AHC vary between 88 and 71 per cent, and the average precision is 77 per cent. The recalls of CKP based on K derived from AHC vary between 54 and 32 per cent, and the average recall is 44 per cent. The precisions of CKP based on *K* derived from AHC are close to the precisions of CKP based on *K* derived from ACM CCS. The results show that the precisions of CKP method retain stable in different values of *K*s. However, the recalls of CKP method based on *K* derived from AHC apparently decrease compared against the recalls of CKP method based on *K* derived from ACM CCS, since the numbers of clusters, i.e. *K*s, derived from AHC are far smaller than the numbers of clusters derived from ACM CCS, that leads to some keywords crossing multiple domains cannot be found by CKP.

Table XII.
Numbers of keyword domains (parameter *K*) derived from different approaches

Year	2000	2001	2002	2003	2004	2005	2006
K derived from ACM CCS	45	42	47	59	47	53	46
K derived from AHC	20	27	33	23	31	19	27

Table XIII.
Effectiveness of CKP finding bridge-keywords based on *K* derived from AHC

Year	2000	2001	2002	2003	2004	2005	2006	Average
CKP precision (%)	88	73	71	73	75	77	83	77
CKP recall (%)	50	47	39	39	47	32	54	44

6. Conclusions

Traditional co-word analysis focuses on co-occurrence of keywords, and therefore, cannot cluster the same keywords into more than one keyword domain. The CKP method considers complete author keyword sets of reference papers to discover bridge-keywords, i.e. keywords crossing multiple keyword domains. Hence, the keyword recommendation system based on CKP can recommend keywords across multiple keyword domains via the bridge-keywords.

In the final step of the CKP method, namely deriving keyword domains, the complete keyword sets have been grouped into clusters, each cluster corresponds to a keyword domain. Although each complete keyword set belongs to one cluster only, the same keywords may appear in more than one complete keyword set, and therefore, appear in more than one keyword domain. The bridge-keywords can then be discovered by the CKP method. However, traditional co-word analysis methods cannot discover any keywords across multiple keyword domains. Furthermore, the CKP method employs co-citation analysis to compute inter-relationships of reference paper keyword sets, whereas traditional co-word analysis employs co-occurrence analysis to compute intra-relationships of paper keyword sets. Therefore, the keyword map obtained by CKP will be different from that obtained by traditional co-word analysis method.

The *JACM* citation bank employed in this experiment is derived from source papers published in *JACM* during 2000-2006. The statistics reveal that the numbers of author keywords of most citation papers are between two to seven, and the average number of author keywords is 5.28 per citation paper. The CKP effectiveness of discovering bridge-keywords in the *JACM* citation bank is evaluated by the benchmark of ACM CCS. The CKP method can yield average precision of 80 per cent and average recall of 76 per cent in the *JACM* citation bank during 2000-2006. Briefly, the CKP method can discover the bridge-keywords that cross multiple keyword domains without relying on the ACM CCS classifier. In this study, the bridge-keywords existing in a real-life citation data bank are double identified by the *de facto* standard of ACM CCS and the proposed CKP method.

This study makes a good beginning for the research of bridge-keywords and cross-domain recommendation. A further study of applying the CKP method to large-scale of web documents should be in order to fully explore the effectiveness of the approach. Another avenue of extending the study is incorporating the CKP method with user profiles for recommendation.

References

Association for Computing Machinery (2009a), "ACM computing classification system toc", ACM, available at: www.acm.org/about/class

Association for Computing Machinery (2009b), "The ACM portal", ACM, available at: http://portal.acm.org

Chen, H. and Lynch, K.J. (1992), "Automatic construction of networks of concepts characterizing document databases", *IEEE Transactional on Systems, Man, and Cybermetics*, Vol. 22 No. 5, pp. 885-902.

Chen, H., Ng, T.D., Martinez, J. and Schatz, B.R. (1997), "A concept space approach to addressing the vocabulary problem in scientific information retrieval: an experiment on the

worm community system", *Journal of the American Society for Information Science*, Vol. 48 No. 1, pp. 17-31.

Ding, Y., Chowdhury, G. and Foo, S. (2000), "Organising keywords in a web search environment: a methodology based on co-word analysis", *Proceedings of the 6th International Society for Knowledge Organization (ISKO 6) Conference, Toronto, Canada*, pp. 28-34.

He, Y. and Hui, S.C. (2000), "Mining a web citation database for author co-citation analysis", *Information Processing & Management*, Vol. 38 No. 4, pp. 491-508.

Johnson, A.G. (1988), *Statistics*, Harcourt Brace Jovanovich, Orlando, FL.

Kitamura, Y., Nanbu, T. and Tatsumi, S. (1999), "A keyword recommendation system for GenBank", *Genome Informatics*, Vol. 10, pp. 206-7.

Lorence, D. and Abraham, J. (2006), "Analysis of semantic search within the domains of uncertainty: using keyword effectiveness indexing as an evaluation tool", *International Journal of Electronic Healthcare*, Vol. 2 No. 3, pp. 263-76.

McCain, K.W. (1990), "Mapping authors in intellectual space: a technical overview", *Journal of the American Society for Information Science*, Vol. 41 No. 6, pp. 433-43.

Schatz, B.R., Johnson, E.H., Cochrane, P.A. and Chen, H. (1996), "Interactive term suggestion for users of digital libraries: using subject thesauri and co-occurrence lists for information retrieval", *Proceedings of the 1st ACM International Conference on Digital libraries (Bethesda, MD, March)*, ACM Press, New York, NY, pp. 126-33.

Whittaker, J., Courtial, J.P. and Law, J. (1989), "Creativity and conformity in science: titles, keywords and co-word analysis", *Social Studies of Science*, Vol. 19 No. 3, pp. 473-96.

Further reading

Ahlgren, P., Jarneving, B. and Rousseau, R. (2003), "Requirements for a cocitation similarity measure, with special reference to Pearson's correlation coefficient", *Journal of the American Society for Information Science and Technology*, Vol. 54 No. 6, pp. 550-60.

Avancini, H. and Straccia, U. (2004), "Personalization, collaboration, and recommendation in the digital library environment CYCLADES", *Proceedings of the IADIS Conference on Applied Computing, Lisbon, Portugal*, pp. 67-74.

Chang, C.C. and Chen, R.S. (2006), "Using data mining technology to solve classification problems: a case study of campus digital library", *The Electronic Library*, Vol. 24 No. 3, pp. 307-21.

Egghe, L. and Rousseau, R. (1990), *Introduction to Informetrics: Quantitative Methods in Library, Documentation and Information Science*, Elsevier Science, Amsterdam.

Eom, S.B. (1996), "Mapping the intellectual structure of research in decision support systems through author cocitation analysis (1971-1993)", *Decision Support Systems*, Vol. 16 No. 4, pp. 315-38.

Fuhr, N., Gövert, N. and Klas, C.P. (2001), "Recommendation in a collaborative digital library environment", technical report, University of Dortmund, Dortmund.

Gao, X., Murugesan, S. and Lo, B.W.N. (2006), "A simple method to extract key terms", *International Journal of Electronic Business*, Vol. 4 Nos 3/4, pp. 221-38.

Haruechaiyasak, C., Shyu, M.L. and Chen, S.C. (2005), "A web-page recommender system via a data mining framework and the semantic web concept", *International Journal of Computer Applications in Technology*, Vol. 27 No. 4, pp. 298-311.

Liang, T.P., Yang, Y.F., Chen, D.N. and Ku, Y.C. (2007), "A semantic-expansion approach to personalized knowledge recommendation", *Decision Support Systems*, Vol. 45 No. 3, pp. 401-12.

Liao, S.H. and Wen, C.H. (2007), "Artificial neural networks classification and clustering of methodologies and applications – literature analysis from 1995 to 2005", *Expert Systems with Applications*, Vol. 32 No. 1, pp. 1-11.

Matsuo, Y. and Ishizuka, M. (2004), "Keyword extraction from a single document using word co-occurrence statistical information", *International Journal on Artificial Intelligence Tools*, Vol. 13 No. 1, pp. 157-69.

Nichols, D.M., Twidale, M.B. and Paice, C.D. (1997), "Recommendation and usage in the digital library", Technical Report CSEG/2/97, Computing Department, Lancaster University, Lancashire.

Roussinov, D. and Zhao, J.L. (2003), "Automatic discovery of similarity relationships through web mining", *Decision Support Systems*, Vol. 35 No. 1, pp. 149-66.

Shiri, A.A., Revie, C. and Chowdhury, G. (2002), "Thesaurus-assisted search term selection and query expansion: a review of user-centred studies", *Knowledge organization*, Vol. 29 No. 1, pp. 1-19.

Tanaka, M., Nakazono, S., Matsuno, H., Tsujimoto, H., Kitamura, Y. and Miyano, S. (2000), "Intelligent system for topic survey in MEDLINE by keyword recommendation and learning text characteristics", *Genome Informatics*, Vol. 11, pp. 73-82.

Villarroel, M., Fuente, P., Pedrero, A., Vegas, J. and Adiego, J. (2002), "Obtaining feedback for indexing from highlighted text", *The Electronic Library*, Vol. 20 No. 4, pp. 306-13.

White, H.D. and Griffith, B.C. (1981), "Author cocitation: a literature measure of intellectual structure", *Journal of the American Society for Information Science*, Vol. 32 No. 3, pp. 163-71.

White, H.D. and McCain, K.W. (1998), "Visualizing a discipline: an author co-citation analysis of information science, 1972-1995", *Journal of the American Society for Information Science*, Vol. 49 No. 4, pp. 327-55.

Yang, Y. and Li, J.Z. (2005), "Interest-based recommendation in digital library", *Journal of Computer Science*, Vol. 1 No. 1, pp. 40-6.

Yang, C., Yang, K.C. and Yuan, H.C. (2007), "Improving the search process through ontology-based adaptive semantic search", *The Electronic Library*, Vol. 25 No. 2, pp. 234-48.

About the authors

Yu-Min Su received the BS degree from the Department of Biology, Fu Jen Catholic University, Taipei, Taiwan in 1991, and the MBA degree from the Graduate Institute of Business Administration, National Taiwan University, Taipei, Taiwan in 1993. He received the PhD degree from the Department of Business Administration of National Central University, Jhongli, Taiwan in 2009. He has taught in the areas of e-commerce, internet marketing and management information system from 1996. He teaches evolutionary computing and e-commerce in the Department of Computer Science of National Chengchi University in Taipei, Taiwan. His research interests include data mining, informetrics, information retrieval, social networks, e-commerce, bioinformation and system biology.

Ping-Yu Hsu graduated from the CSIE Department of National Taiwan University in 1987, received the master's degree from the Computer Science Department of New York University in 1991, and the PhD degree from the Computer Science Department of UCLA in 1995. He is a professor in the Business Administration Department of National Central University in Jhongli,

Taiwan. He also works as the Secretary-in-Chief of the Chinese ERP Association. His research interest focuses on business data applications, including data modeling, data warehousing, data mining and ERP applications in business domains. His papers have been published in *IEEE Transactions on Software Engineering*, *Information Systems* and various other journals. Ping-Yu Hsu is the corresponding author and can be contacted at: pyhsu@mgt.ncu.edu.tw

Ning-Yao Pai received the BA degree from the Department of Economics, National Chengchi University, Taipei, Taiwan in 1999, and the MBA degree from the Graduate Institution of Industrial Economics of National Central University, Jhongli, Taiwan in 2009. She is currently a doctoral student at the Institute of Information Management of National Chiao Tung University, Hsinchu, Taiwan. Her research interests include data mining, social networks, network economics and game theory.