DOI 10.2298/CSIS101124020C

Concept Extraction and Clustering for Search Result Organization and Virtual Community Construction

Shihn-Yuarn Chen[1], Chia-Ning Chang[2], Yi-Hsiang Nien[3], and Hao-Ren Ke[4]

[1] PhD Candidate, Dept. of Computer Science, National Chiao Tung University
Hsinchu, 300, Taiwan
sychen@cs.nctu.edu.tw
[2] Master, Dept. of Computer Science, National Chiao Tung University
Hsinchu, 300, Taiwan
chianing.chang@gmail.com
[3] Master, Institute of Information Management, National Chiao Tung University
Hsinchu, 300, Taiwan
hugo3318@gmail.com
[4] Corresponding Author, Professor and Deputy Library Director, Graduate Institute of Library and Information Studies, National Taiwan Normal University
Taipei, 106, Taiwan
clavenke@ntnu.edu.tw

Abstract. This study proposes a concept extraction and clustering method, which improves Topic Keyword Clustering by using Log Likelihood Ratio for semantic correlation and Bisection K-Means for document clustering. Two value-added services are proposed to show how this approach can benefit information retrieval (IR) systems. The first service focuses on the organization and visual presentation of search results by clustering and bibliographic coupling. The second one aims at constructing virtual research communities and recommending significant papers to researchers. In addition to the two services, this study conducts quantitative and qualitative evaluations to show the feasibility of the proposed method; moreover, comparison with the previous approach is also performed. The experimental results show that the accuracy of the proposed method for search result organization reaches 80%, outperforming Topic Keyword Clustering. Both the precision and recall of virtual community construction are higher than 70%, and the accuracy of paper recommendation is almost 90%.

Keywords: information retrieval, concept extraction, document clustering, virtual community, social network analysis, bibliographic coupling.

1. Introduction

In the digital library era, information retrieval (IR) systems become essential for researchers to discover literature on particular subjects. Two major concerns when a researcher uses IR systems are how to filter out irrelevant documents and how to discover latest or significant documents.

IR systems have been used in digital libraries for decades, and various approaches, such as query reformulation/expansion [1], have been applied to improve the search quality. However, it is not easy for IR systems to correctly identify information need of every individual within a few queries. Personalized search is one solution, but analyzing collected personal information (e.g. search log, click-through history) has to be done in advance. Moreover, one well-known problem of IR systems is the representation of search results, which are usually in an item-by-item list view. Even for a veteran, such kind of representation takes time to filter out irrelevant documents. This study focuses on refining the organization of results by clustering. Search results are clustered by concepts; in addition, clusters and relationships between clusters are represented in a visual form. Users can have an overview of the search result concepts, easily identifying which groups are closer to their interests, and then look into the group for literatures.

Besides better representation of search results, researchers also use IR systems to find out latest or important documents in their research areas. These documents are either highly cited or written by significant authors in particular areas, and provide the latest research trends or basic knowledge of a research domain. Citation databases [2] such as SCI/SSCI/A&HCI, Scopus, and Google Scholar can provide information on highly cited documents. This study tackles this issue by constructing a social network of authors from their co-authoring relationship so as to construct virtual research communities and find out representative authors in a research area. Then the system will recommend documents to the authors themselves or those who are interested in an author's documents.

This study extracts concepts from textual information of documents. A concept comprises a group of terms and is constructed on the basis of word co-occurrence statistics. On the basis of the extracted concepts, two value-added IR services that exploit clustering, citation relations, and social information among authors are developed for solving the two preceding issues. The rest of this paper is organized as follows. Section 2 describes related works. Section 3 provides an overview of the value-added IR services proposed by this study, and elaborates the core mechanism for concept extraction and clustering. Section 4 proposes the service for the organization and visual representation of search results. Section 5 presents the service for virtual research community construction and paper recommendation. Section 6 describes the evaluation, and Section 7 draws a brief conclusion and future work.

2. Related Works

This section reviews three topics related to this study, including document clustering, social network analysis, and different representations of search results.

2.1. Document Clustering

Among the various types of clustering techniques [3], hierarchical and partitional clustering are the most common. Hierarchical clustering often provides better results, but the time complexity is the vital issue. In contrast, partitional clustering methods, such as k-means [4], are more efficient than hierarchical methods, but the proper number of clusters is hard to define in advance.

Though lots of efforts had been devoted to document clustering research, there are still various issues discussed in these years, including clustering techniques [5], similarity metrics [6], online Web document clustering [7], local optimal problem [8], knowledge aided document clustering [9], etc. One of the most common issues of document clustering is that there are too many words and phrases in the corpus, and a high-dimensional clustering is usually time-consuming [5]. Many approaches are proposed to improve traditional document clustering. Some try to select representative features to represent a document, and perform document clustering based on these features. For example, Iliopoulos et al. proposed TEXTQUEST[1] for Medline documents clustering [11]. TEXTQUEST is based on statistical treatment of words, and TF-IDF is used to extract important words from abstracts. Each Medline abstract is represented as a vector, which is used as input for the document clustering.

Some approaches use word clusters to improve the quality of document clustering. For example, Slonim and Tishby [12] use word clusters to capture the mutual information [13] about a set of documents. The experiments on a corpus comprising 20 newsgroups show that word clusters are helpful in document clustering. Another example is Topic Keyword Clustering [14]. Firstly, Topic Keyword Clustering retrieves keywords from documents and clusters them to obtain concept clusters. Secondly, it computes the similarity of each keyword cluster and each document, and finally assigns each document to the most similar keyword cluster to achieve document clustering. This study modifies Topic Keyword Clustering to organize search results and find out virtual research communities.

[1] http://www.textquest.de/

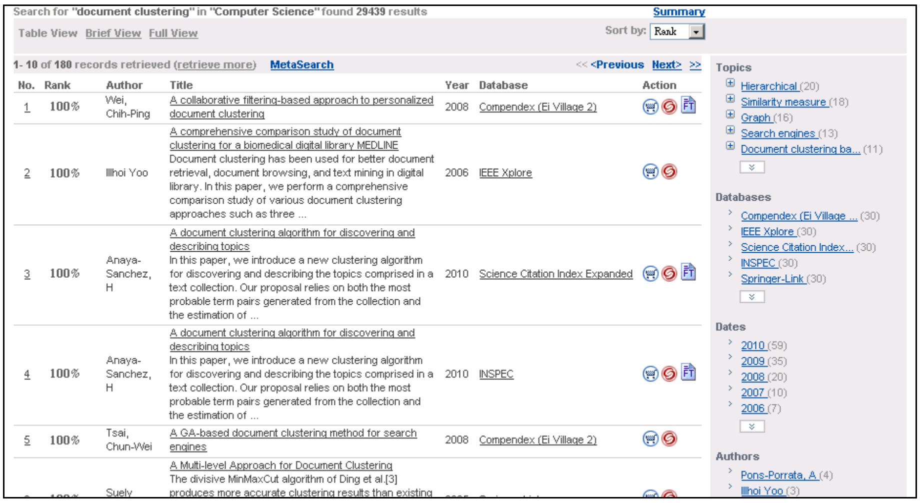
2.2. Social Network Analysis

Social Network Analysis (SNA) is a method based on sociometry for studying into social organizations, people relationships and interactions. Moreno quantizes people relationships and interactions, and represents the interactions and distances between people in a social network graph [15]. The vertices and edges in the social network graph represent social actors and the relationships of social actors, respectively. The properties of a social network graph can be described by global graph metrics and individual actor properties [16]. Global graph metrics describes the characteristics of a social network as a whole, e.g. average node distance, the number of components (fully connected subgraphs), clusters, etc. Individual actor properties cover actor distance, position in a cluster, etc. This graph representation allows applying graph theory [17] to analyze the tangled interaction of a society. Graph theory models objects and their relationships in mathematical structure for analyzing. In graph theory, objects are modeled as vertices, and the relationships of pairwise objects are modeled as directed or undirected edges, and the weight of edges represents the relationship degree. Graph theory is used in various applications, such as discovering implicit people relationships and exploring diseases propagation. Besides, with various metrics, researchers can know more about a social network and the significance of a member in the network.

Tyler et al. [18] use email to construct a social network in graph representation, and employ Betweenness Centrality to analyze communities in the network. Mika [19] proposes a system named Flink, which analyzes Web pages, emails, research papers and FOAF (Friend of a Friend) data to find out the social network of researchers, and visualizes the network and the ontology of a research area. Liu et al. [16] uses ACM, IEEE and JCDL (ACM/IEEE Joint Conference on Digital Libraries) publications as corpus for analyzing the co-author relationships; Liu et al. construct the social network of researchers and propose AuthorRank to compare with PageRank [20]. POLYPHONET uses participants of Japan Society of Artificial Intelligence conferences and their publications to explore their research interests and compute the context similarity for constructing social network relationships [21].

2.3. Different Representations of Search Results

The most common scenario that a user utilizes an IR system is – after submitting a query, the system returns a list of results to the user, and the user has to filter out unnecessary results in the list. Many systems pay a lot of efforts to provide better representation of results for users. Facet analysis is an example, which organizes results according to various types of criteria, such as publication date, topics, author, and publisher. Fig. 1 shows the interface of a federated search system called MetaLib[2], which not only returns a list of search results, but also displays the facet criteria (including clustering) at the right side. Users can look into an interesting group by clicking any facet hyperlink.



Search for "document clustering" in "Computer Science" found 29439 results

Summary

Table View Brief View Full View

Sort by: Rank

1- 10 of 180 records retrieved (retrieve more) MetaSearch

<< <Previous Next> >>

No.	Rank	Author	Title	Year	Database	Action
1	100%	Wei, Chih-Ping	A collaborative filtering-based approach to personalized document clustering	2008	Compendex (Ei Village 2)	FT
2	100%	Illhoi Yoo	A comprehensive comparison study of document clustering for a biomedical digital library MEDLINE Document clustering has been used for better document retrieval, document browsing, and text mining in digital library. In this paper, we perform a comprehensive comparison study of various document clustering approaches such as three ...	2006	IEEE Xplore	
3	100%	Anaya-Sanchez, H	A document clustering algorithm for discovering and describing topics In this paper, we introduce a new clustering algorithm for discovering and describing the topics comprised in a text collection. Our proposal relies on both the most probable term pairs generated from the collection and the estimation of ...	2010	Science Citation Index Expanded	FT
4	100%	Anaya-Sanchez, H	A document clustering algorithm for discovering and describing topics In this paper, we introduce a new clustering algorithm for discovering and describing the topics comprised in a text collection. Our proposal relies on both the most probable term pairs generated from the collection and the estimation of ...	2010	INSPEC	FT
5	100%	Tsai, Chun-Wei	A GA-based document clustering method for search engines	2008	Compendex (Ei Village 2)	
		Suely	A Multi-level Approach for Document Clustering The divisive MinMaxCut algorithm of Ding et al.[3] produces more accurate clustering results than existing			

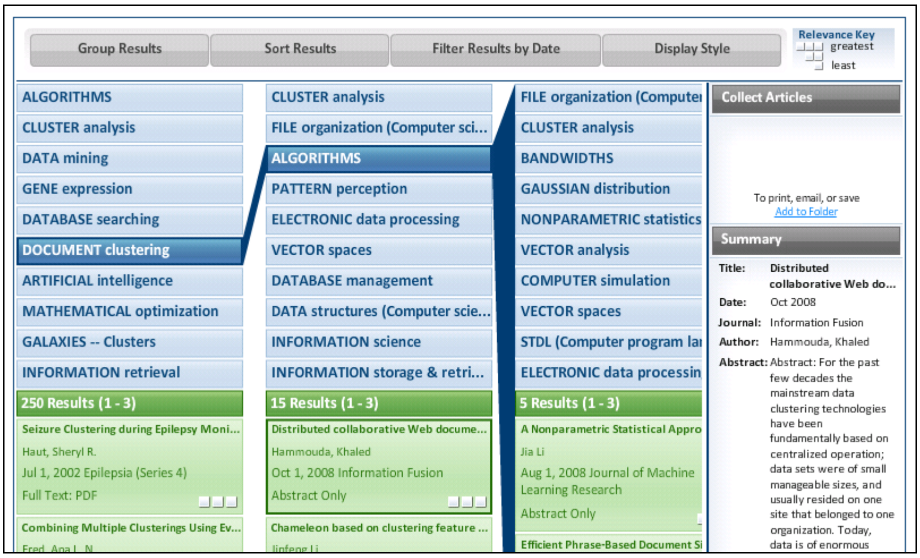
Topics
- ⊞ Hierarchical (20)
- ⊞ Similarity measure (18)
- ⊞ Graph (16)
- ⊞ Search engines (13)
- ⊞ Document clustering ba... (11)

Databases
- Compendex (Ei Village ... (30)
- IEEE Xplore (30)
- Science Citation Index... (30)
- INSPEC (30)
- Springer-Link (30)

Dates
- 2010 (59)
- 2009 (35)
- 2008 (20)
- 2007 (10)
- 2006 (7)

Authors
- Pons-Porrata, A (4)
- Illhoi Yoo (3)

Fig. 1. Facet analysis in MetaLib.



ALGORITHMS
CLUSTER analysis
DATA mining
GENE perception
DATABASE searching
DOCUMENT clustering
ARTIFICIAL intelligence
MATHEMATICAL optimization
GALAXIES -- Clusters
INFORMATION retrieval

250 Results (1 - 3)

Seizure Clustering during Epilepsy Moni...
Haut, Sheryl R.
Jul 1, 2002 Epilepsia (Series 4)
Full Text: PDF

Combining Multiple Clusterings Using Ev...
Fred, Ana L. N.

CLUSTER analysis
FILE organization (Computer sci...
ALGORITHMS
PATTERN perception
ELECTRONIC data processing
VECTOR spaces
DATABASE management
DATA structures (Computer scie...
INFORMATION science
INFORMATION storage & retri...

15 Results (1 - 3)

Distributed collaborative Web docume...
Hammouda, Khaled
Oct 1, 2008 Information Fusion
Abstract Only

Chameleon based on clustering feature ...
Jinfeng Li

FILE organization (Computer
CLUSTER analysis
BANDWIDTHS
GAUSSIAN distribution
NONPARAMETRIC statistics
VECTOR analysis
COMPUTER simulation
VECTOR spaces
STDL (Computer program la
ELECTRONIC data processin

5 Results (1 - 3)

A Nonparametric Statistical Appro
Jia Li
Aug 1, 2008 Journal of Machine Learning Research
Abstract Only

Efficient Phrase-Based Document Si

Collect Articles

To print, email, or save
Add to Folder

Summary

Title:	Distributed collaborative Web do...
Date:	Oct 2008
Journal:	Information Fusion
Author:	Hammouda, Khaled
Abstract:	Abstract: For the past few decades the mainstream data clustering technologies have been fundamentally based on centralized operation; data sets were of small manageable sizes, and usually resided on one site that belonged to one organization. Today, data is of enormous

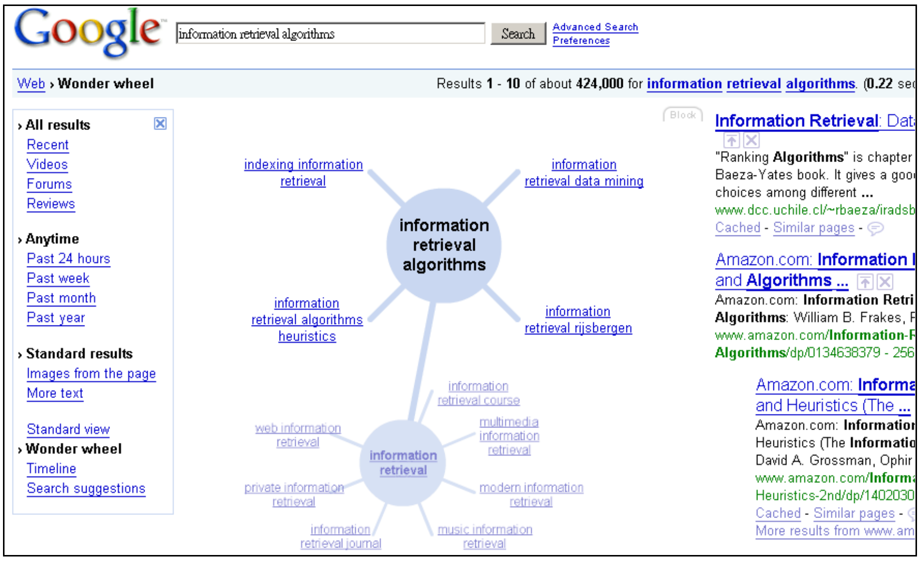
Fig. 2. Visual representation of search results in EBSCOhost.

[2] http://www.exlibrisgroup.com/category/MetaLibOverview

EBSCOhost[3] provides a visual representation of search results, as shown in Fig. 2. Users can click a result group to expand its sub-groups and result documents.

Google.com also provides "wonder wheel" view of search results (Fig. 3), which is similar to what EBSCOhost provides. Users can click a topic group on a wheel border to expand more detail results.

Grokker.com presents the results in a graphical view, as depicted in Fig. 4. Grokker.com organizes search results from Yahoo!, Wikipedia, and Amazon Books in a graphical interface. Each group is labeled with a keyword which helps users to figure out the concept of the group and to judge whether the group is relevant to their information need.


Google
Search
Advanced Search
Preferences
Web › Wonder wheel
› All results
Recent
Videos
Forums
Reviews
› Anytime
Past 24 hours
Past week
Past month
Past year
› Standard results
Images from the page
More text
Standard view
› Wonder wheel
Timeline
Search suggestions
indexing information retrieval
information retrieval data mining
information retrieval algorithms
information retrieval algorithms heuristics
information retrieval rijsbergen
information retrieval course
web information retrieval
multimedia information retrieval
information retrieval
private information retrieval
modern information retrieval
information retrieval journal
music information retrieval
Block

information retrieval algorithms

Results **1** - **10** of about **424,000** for **information retrieval algorithms**. (**0.22** sec

Information Retrieval: Dat
"Ranking **Algorithms**" is chapter
Baeza-Yates book. It gives a goo
choices among different **...**
www.dcc.uchile.cl/~rbaeza/iradsb
Cached - Similar pages -

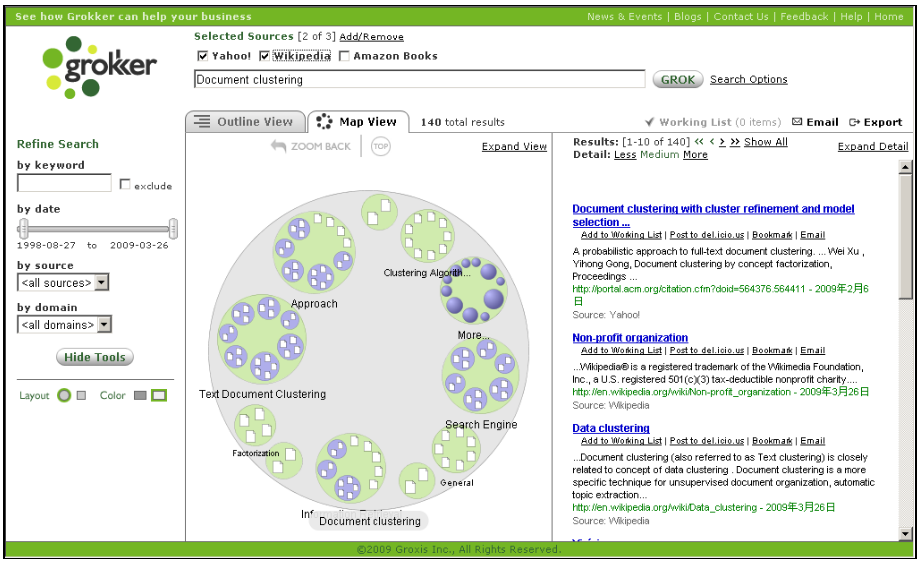
Amazon.com: **Information I**
and Algorithms ...
Amazon.com: **Information Retri**
Algorithms: William B. Frakes, F
www.amazon.com/**Information-F**
Algorithms/dp/0134638379 - 256

Amazon.com: **Informa**
and Heuristics (The **...**
Amazon.com: **Informatio**
Heuristics (The **Informatio**
David A. Grossman, Ophir
www.amazon.com/**Informa**
Heuristics-2nd/dp/1402030
Cached - Similar pages -
More results from www.am

Fig. 3. Google "wonder wheel" view of search results.

These systems provide alternative organizations of search results to users, and save users' time on filtering out irrelevant results. However, these systems do not provide their users with sufficient relationship information among topic groups or virtual research communities.

[3] http://search.ebscohost.com/



☑ Yahoo! ☑ Wikipedia □ Amazon Books

Document clustering

140 total results

□ exclude

1998-08-27 to 2009-03-26

<all sources>

<all domains>

Results: [1-10 of 140] << < > >> Show All
Detail: Less Medium More

Document clustering with cluster refinement and model selection ...
Add to Working List | Post to del.icio.us | Bookmark | Email
A probabilistic approach to full-text document clustering. ... Wei Xu , Yihong Gong, Document clustering by concept factorization, Proceedings ...
http://portal.acm.org/citation.cfm?doid=564376.564411 - 2009年2月6日
Source: Yahoo!

Non-profit organization
Add to Working List | Post to del.icio.us | Bookmark | Email
...Wikipedia® is a registered trademark of the Wikimedia Foundation, Inc., a U.S. registered 501(c)(3) tax-deductible nonprofit charity....
http://en.wikipedia.org/wiki/Non-profit_organization - 2009年3月26日
Source: Wikipedia

Data clustering
Add to Working List | Post to del.icio.us | Bookmark | Email
...Document clustering (also referred to as Text clustering) is closely related to concept of data clustering . Document clustering is a more specific technique for unsupervised document organization, automatic topic extraction...
http://en.wikipedia.org/wiki/Data_clustering - 2009年3月26日
Source: Wikipedia

Fig. 4. Visual representation of search results of Grokker.com.

3. Value-Added IR Services Based on Concept Extraction and Clustering

First, this section describes the system architecture of the IR system with the two value-added services proposed in Section 4 and Section 5. Then, the core mechanism for the two value-added services, concept extraction and clustering, will be elaborated.

3.1. System Architecture

As shown in Fig. 5, an IR system stores document items. Each item has an internal representation and is indexed for efficient retrieval. For each new item, preprocessing tasks such as tokenization, lowercasing, stop words removing, part of speech (POS) tagging, and stemming [1] are conducted for extracting important features for the item; and these features, usually called index terms or keywords[4], and their weights (e.g. TF-IDF) associated with the item are inserted into the index. When a user submits a query, the system analyzes the user's information need and translates the information need into a formal query. The IR system then exploits the index to find relevant items, which are traditionally returned in an item-by-item list view.

[4] Hereafter, features, keywords, index terms, and terms are used interchangeably.

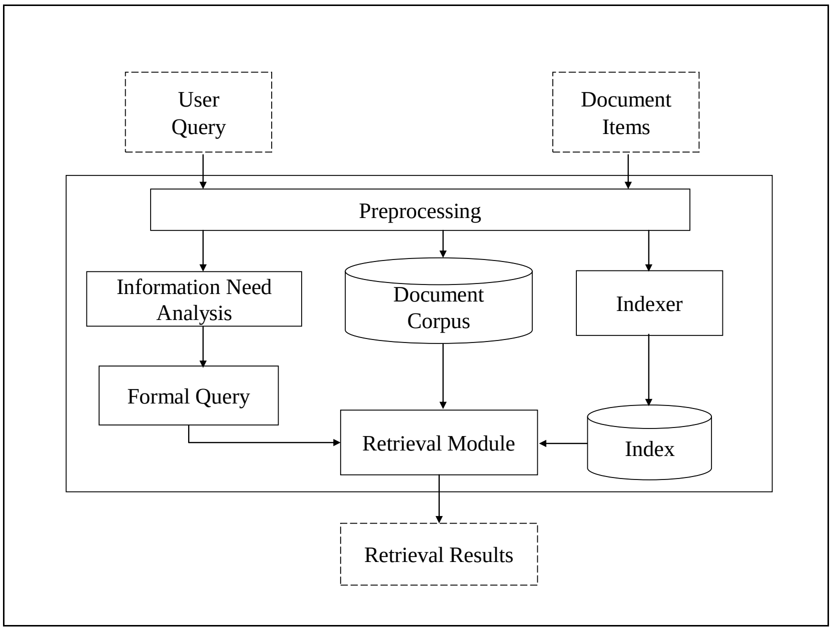


Fig. 5. System architecture of an IR system.

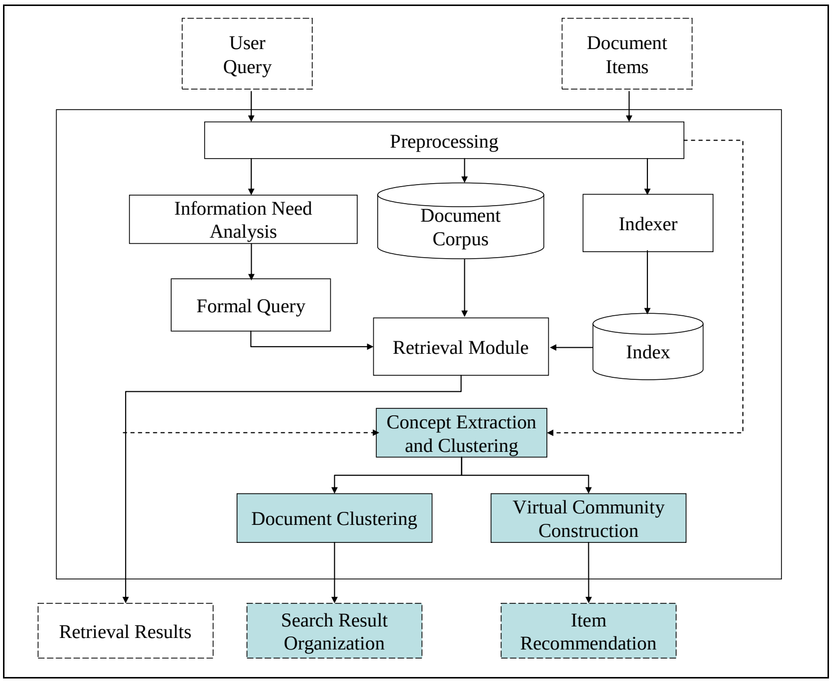


Fig. 6. System architecture of an IR system with proposed value added services.

This study aims at utilizing concept extraction and clustering to provide value-added services in an information retrieval system. These value-added services cover two aspects. The first reorganizes listed retrieval results into a graphical representation for intuitive browsing, and the second analyzes the relationships between authors of document items to generate virtual communities for item recommendation. The architecture of these proposed value-added services of an IR system is illustrated in Fig. 6. The following subsection will describe the core technique of these value-added services, concept-based extraction and clustering. The two value-added services will be explicated in Section 4 and Section 5, respectively.

3.2. Concept Extraction and Clustering

Concept extraction and clustering is the core technique of the proposed value-added IR services. A concept comprises a group of terms and is constructed on the basis of word co-occurrence statistics. The underlying assumption is that words co-occurring in a passage (such as document or paragraph) are likely to be in some sense similar or related in semantics [1]. The following subsections explicate how to conduct concept clustering, which is based on Topic Keyword Clustering [14].

3.2.1. Computing Semantic Relationship

The semantic relationship of two terms is based on their co-occurrences in a certain passage, for example, co-occurrences in the same paragraph. Different applications may choose different types of passages to judge word co-occurrences; furthermore, different applications may choose different methods to compute the semantic relationship of two terms. This study considers the co-occurrences in the same sentence and a few bibliographic fields like *title* and *keyword* fields. Instead of mutual information (MI), which is used in Topic Keyword Clustering, this study exploits Log Likelihood Ratio (LLR) [22][23] for computing the relationship between terms i and j, r_{ij} and keeps the confidence of r_{ij} larger than 99.9%. The main reason this study uses LLR is that MI is more proper for judging the independent degree rather than relative degree of two terms [34].

3.2.2. Concept Extraction and Clustering

After semantic relationships of terms (features) are computed, the concept extraction process is applied to construct concept clusters. The detailed processes are described below.

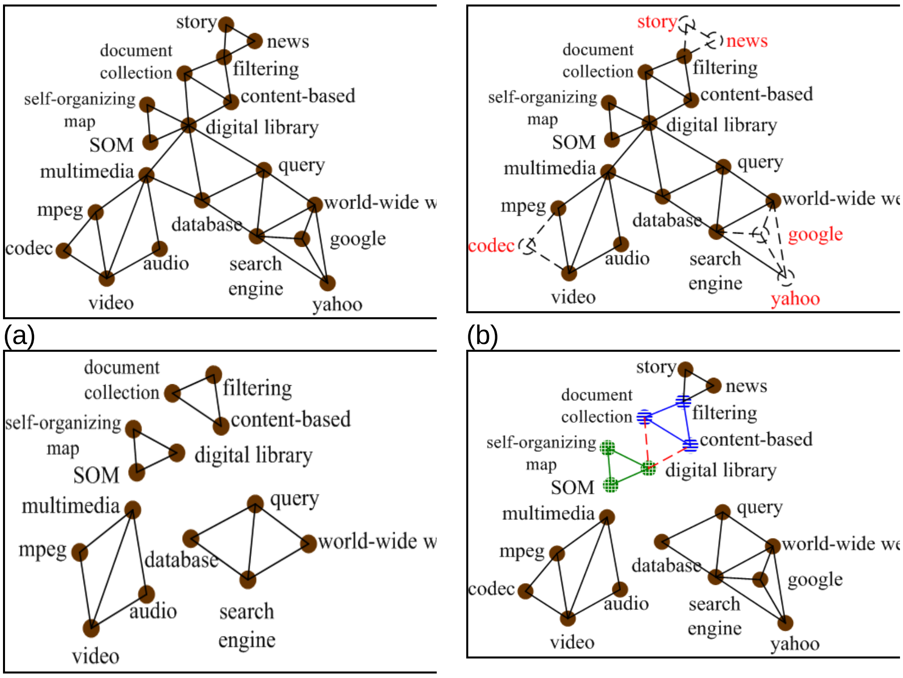
3.2.2.1. Feature Nethwork Construction

1. Constructing draft feature network

In the draft feature network, each vertex represents a feature, and the weighted edge represents the relationship of two features, which is computed in 0. Fig. 7 shows the idea of future network construction, and Fig. 7 (a) represents the draft feature network.

2. Removing insignificant edges and vertices

Insignificant edges and vertices may cause noises during concept extraction. To remove insignificant edges, edges with weight less than the threshold, which is defined as the average weight of edges in the network, are removed. For experimental comparison of Topic Keyword Clustering and our approach, this study follows the parameter and threshold settings of Topic Keyword Clustering.



(a)

(b)

(c)

(d)

Fig. 7. Example of concept extraction and clustering [14]

The weight of each vertex is CW_i, where r_{ij} is the relationship of two term vertices V_i and V_j computed in 3.2.1, and m is the number of vertices connected to V_i,. For example, in Fig. 7 (a), the vertex “story” is connected by “news” and “filtering”, so the weight of “story”, CW_{story}, is the average weight of the two connected edges (i.e. the average of relationships of story-news and story-filtering).

$$CW_i = \frac{\sum_{j=1}^{m} r_{ij}}{m} \tag{1}$$

To remove insignificant vertices, the threshold is the average weight of all vertices. Fig. 7 (b) shows the network after removing insignificant edges and vertices.

3.2.2.2. Concept Extraction and Concept Cluster Generation

1. *k*-Nearest neighbor (knn) grouping [24]

Each vertex and its *k* nearest neighbors can be grouped as a connected graph, named candidate feature unit. Different values of *k* affect the number of candidate feature units. A larger k results in fewer units, and each unit contains more features. If there are too many features in a candidate feature unit, its concept may not be well expressed, and the significance represented by each feature is also reduced. This study follows the setting of Topic Keyword Clustering and lets *k*=2. Fig. 7 (c) shows an example containing four candidate feature units.

2. Candidate Feature Subgroups Generation

Except the edges that cross two candidate feature units, out-linked edges from each candidate feature unit are recovered to form candidate feature subgroups. The weight of a subgroup is the sum of edge weights $\mathrm{W}_{\mathrm{G}_m}$, where G_m represents a candidate feature subgroup *m*, and r_{ij} is the weight of an edge.

$$W_{G_m} = \sum_{r_{ij} \in G_m} r_{ij} \tag{2}$$

3. Candidate Feature Subgroups Merging

The goal of subgroups merging is to reduce the number of subgroups by a greedy algorithm. The two most inter-connected subgroups are merged iteratively till the inter-connected degrees of all subgroup pairs are less than the threshold (empirically chooses 0.5 in this study). The degree of inter-connected subgroups, *RI(G_i, G_j)*,is computed as follows.

$$RI(G_i, G_j) = \frac{\left| W_{E(G_i, G_j)} \right|}{\left| W_{G_i} \right| + \left| W_{Gj} \right|} \tag{3}$$

$W_{E(Gi,\ Gj)}$ is the sum of the weights of the inter-connected edges of the two candidate feature subgroups, G_i and G_j. If *RI* is larger than the threshold, it indicates that the two subgroups are significantly related and should be merged. For example, in Fig. 7 (d), the two subgroups {digital library, SOM, self-organizing map} and {document collection, filtering, content-based} are merged.

$$AS(G) = \frac{\sum_{r_{ij} \in E(G)} r_{ij}}{|E(G)|}$$

$$CD(G) = \frac{(E(G))}{|V(G)| \times |V(G) - 1| \big/ 2} \tag{4}$$

$$CW = CD(G) \times AS(G)$$

4. Concept Clusters Generation

Subgroups with more features usually cover more documents, and a different amount of documents may affect the accuracy of clustering. To reduce the number of different documents in each subgroup, the concept weight of each subgroup (*CW*) should be considered.

CW is computed by multiplying the average edge weight (*AS*) and the connected density (*CD*). If the concept weight of a subgroup is less than the average or the number of features is larger than the average (as the threshold setting in Topic Keyword Clustering), the most non-related feature is removed. And, the final results are the concept clusters.

4. Search Result Organization

The first value-added IR service utilizing the concept extraction and clustering method in Section 3 lies in presenting search results in a graphical form. To demonstrate this service, a corpus containing about 500 "Information Retrieval" related documents from CiteSeer[5] is used. Three bibliographic fields, *title*, *abstract* and *citations*, are collected for each document in this corpus. The average length of abstracts is about 1000 words.

4.1. Concept Extraction

Before concept extraction, pre-processing is applied to *title* and *abstract* fields, and terms appearing in more than 8% of documents or less than three

[5] http://citeseer.ist.psu.edu/

times are filtered out. The concept extraction algorithm mentioned in Section 3.2 is then applied to the remaining terms for constructing concept groups.

This value-added service only considers the co-occurrences in the same sentence of the *title* and *abstract* fields. As mentioned previously, Log Likelihood Ratio (LLR) is used for computing the relationship between terms *i* and *j*, r_{ij}; in addition, to emphasize the importance of the *title* filed, terms co-occur in the *title* field are doubly weighted.

4.2. Document Clustering and Post-processing

After concept groups are generated, this study represents each document as a vector. Each element in the vector represents the semantic similarity of a document and a concept group. Documents can then be grouped by clustering these semantic similarity vectors. The detail of document clustering and the post processing are described in the following subsections.

4.2.1. Semantic Similarity Vector and Document Clustering

According to Section 3.2, *M* concept clusters can be generated and there are *n* features in total. A document D_j and a concept cluster C_m can be represented as vectors by these *n* features.

$D_j = (w_1, w_2, w_3, ..., w_n)$ where *j* = 1, 2, 3, …, *N*

w_i: the weights of the feature *i* (estimated as TF-IDF)

N: # of documents in corpus

$C_k = (t_{k1}, t_{k2}, t_{k3}, ..., t_{kn})$ where *k* = 1, 2, 3, …, *M* (5)

$$t_{ki} = \begin{cases} 1, & \text{If term } i \in C_k \\ 0 & \text{otherwise} \end{cases}$$

This study uses cosine similarity to compute the similarity between a document and a concept cluster. In this manner, each document D_j can also be represented as a vector of semantic similarities with all concepts, SD_j. SD_j also represents the degree of how well a document covers a concept.

$$SD_j = \left(sim(D_j, C_1), sim(D_j, C_2), sim(D_j, C_3), ..., sim(D_j, C_M)\right) \quad (6)$$

Bisection *k*-Means algorithm [25] is then applied to perform document clustering. After document clustering, post-processing, including labeling, cluster relationship construction, and visualization, are applied.

4.2.2. Labeling

Visualization of document clustering usually applies labeling to help users understand the concepts and meanings of each cluster. In this study, the following steps are applied to choose proper labels:

- Select the top 10 weighted features from each cluster;
- In the 10 features, nouns and noun phrases are selected;
- Choose the top 2 weighted candidates as the final labels.

4.2.3. Cluster Relation Construction

Besides labeling, the relation of documents is also helpful to understand relationships of clusters. As CiteSeer preserves the references made by each document, this study exploits the idea of bibliographic coupling [26] to establish the relation between clusters. The similarity of documents cited by two documents can be computed, and then the similarity can be used to construct the citation relation of the two documents. Furthermore, the citation relations of clusters can be computed and used in visualization. The similarity of cited documents for two documents is computed as follows:

1. Hyperlinks (URL) of cited documents

URLs of cited documents are considered in higher priority. In CiteSeer, most documents cited by a document have hyperlinks, and can be accessed through these links. A hyperlink URL can be treated as the unique id of the document. Then the citation similarity of two documents, *linksim*$_{i,j}$, can be computed, where d_i and d_j are two documents, and *link*(d_i) and *link*(d_j) are links to their cited documents, respectively.

$$linksim_{i,j} = \frac{\left|link\left(d_i\right) \cap link\left(d_j\right)\right|}{\min\left(\left|link\left(d_i\right)\right|, \left|link\left(d_j\right)\right|\right)} \tag{7}$$

2. Titles of cited documents

If the URL of a cited document is lost or there are multiple different hyperlinks of a cited document, the abovementioned citation similarity of two documents is affected. To fix the error, titles of cited documents should be considered. After pre-processing and vectorization, the similarity of titles, *textsim*, can be computed by cosine similarity.

The preceding two similarities, *linksim* and *textsim*, are linearly combined to obtain the citation similarity *DR(i,j)* of two documents d_i and d_j, and this study choose α = 0.5 empirically.

$$DR(i,j)=\alpha\cdot\left(linksim_{i,j}\right)+(1-\alpha)\cdot\left(textsim_{i,j}\right) \tag{8}$$

Citation relation of clusters is computed based on the citation similarity of documents. If C_m contains d_i, C_n contains d_i, and the citation similarity of d_i and d_i, *DR(i,j)*, is larger than threshold, θ (0.5, empirically chosen in this work), the citation count of C_m and C_n is increased by 1. Then, the cluster similarity of C_m and C_n, *CR(m,n)*, can be computed.

$$CR(m,n)=\frac{\text{\# of document pairs }\left(d_i,d_j\right)\text{ that }DR(i,j)>\theta}{\max\left(\left|C_m\right|,\left|C_n\right|\right)} \tag{9}$$

4.2.4. Cluster Visualization

Two visualization methods are exploited to represent the clustering result, and Table 1 shows the comparison of the proposed approach and other existed approaches mentioned in Section 2.3. The first method is used for inner document cluster representation, and Fig. 8 is an example. Each circle in Fig. 8 represents a cluster, and the more documents a cluster contains, the larger its radius is.

Table 1. Visual representation comparison of the proposed approach and others.

Concept label	Inner Cluster	Inter Cluster
Our Method	Radius of circle for cluster size. Different color for inner-cluster similarity	Radar graph to show cluster relations. Cluster relation degrees are labeled.
MetaLib	Numeric label for cluster size.	No relations are shown
EBSCOhost	Topic grouping only	Hierarchy-like structure for cluster relations
Google wonder wheel	Topic grouping only	Network-like graph for cluster relations
Grokker.com	Radius of circle for cluster size.	Hierarchy-like structure for cluster relations

Furthermore, the color of a pie represents the similarity / homogeneity of documents in a cluster, the darker the more similar / homogeneous they are.

Besides, the text above a circle is the generated labels of a cluster. In Fig. 8, the document similarity of the cluster “collaborative filtering” is the highest, followed by “speech Recognition”, “self-organizing map”, and “decision tree” is the lowest.

All the approach mentioned in Section 2.3 and the proposed approach provide document clustering/grouping according to the topics, but only MetaLib, Grokker.com and the proposed approach provides the amount information of each clusters/groups. MetaLib uses numeric label; Grokker.com and our proposed approach uses graphical view (the size of circles) to represent the cluster/group size. Besides, only our proposed approach uses different colors to represents the inner-cluster similarity.

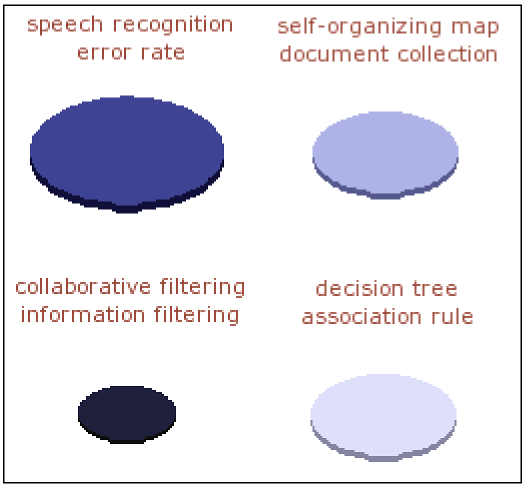


Fig. 8. Chart for cluster visualization.

The second method is a radar map for representing the relationships between clusters. In Fig. 9, the labels, “self-organizing map” and “document collection”, represent the concept of a cluster. Each apex represents a related cluster. The relation degree of two clusters can be realized easily by the radar map; in this example, the “text categorization and relevance feedback” cluster has the highest degree relation with the cluster labeled with “self-organizing map” and “document collection”.

As shown in Table 1, except MetaLib, all the approaches in Section 2.3 and the proposed approach provide the information of topic cluster/group relationship. EBSCOhost and Grokker.com model this information in a hierarchy-like representation, and Google wonder wheel models in a network-like representation. However, only our approach provides the relationship degree to users in a radar graph representation.

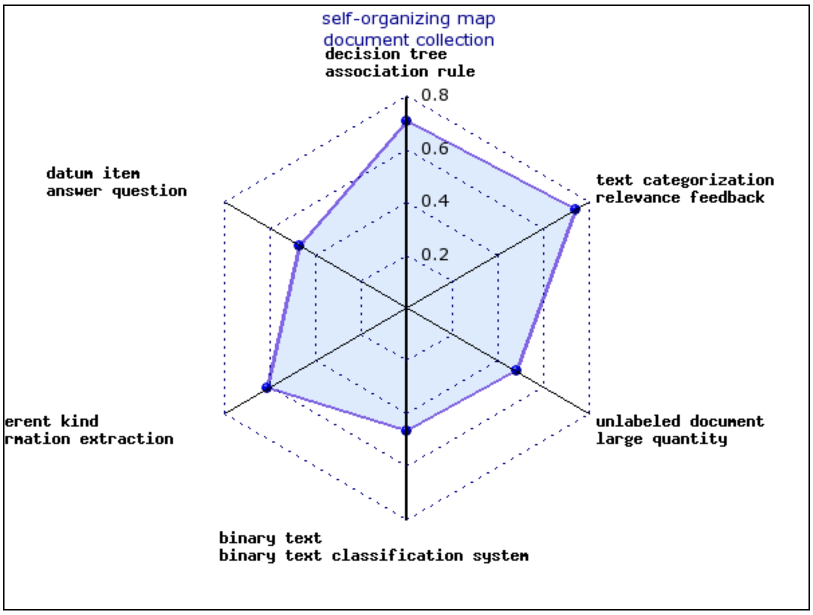


Fig. 9. Radar map for cluster relation representation.

5. Virtual Community Generation and Paper Recommendation

The second value-added IR service concerns the generation of virtual research communities and the recommendation of papers based on the virtual communities. To demonstrate this service, a corpus comprising 226 research papers from the institutional repository[6] of National Chiao Tung University (NCTUIR) is used, and four bibliographic fields, *title*, *abstract*, *keyword* and *author*, are collected for each paper in this corpus.

5.1. Author Model

Each term used in a paper is assumed to relate to the paper's concept, and also has a relationship to each author. The author model uses the relationship of terms and authors to model the research interests of authors. The relationship of terms and authors can be computed by TF-IAF (Term Frequency-Inverse Author Frequency). [27]

$$tf_{ij} = freq_i, \text{frequency of term } i \text{ associated with author} \quad (10)$$

[6] http://ir.lib.nctu.edu.tw/

j

$iaf_i = \log_2 \frac{N}{n_i}$, where *N*: # of total authors,

n_i: # of authors who use term *i*

$$w_{ij} = \begin{cases} tf_{ij} \times ia & \text{, if term } i \text{ is associated with author } j \\ 0 & \text{, otherwise} \end{cases}$$

$U_j = (w_{1j}, w_{2j}, ..., w_{mj})$, where *m* is the number of terms

After computing TF-IAF, the research interests of an author *j* can be represented as a vector, U_j. The author model is the collection of authors, and can be represented as a matrix, *U*.

$$U = \begin{bmatrix} U_1 \\ U_2 \\ U_3 \\ \vdots \\ U_N \end{bmatrix}, \text{where } N \text{ is the total number of authors} \quad (11)$$

5.2. Concept Clustering and Labeling

This value-added service only considers the co-occurrences in the same sentence of the *title* and *abstract* fields. As mentioned previously, Log Likelihood Ratio (LLR) is used for computing the relationship between terms *i* and *j*, r_{ij}; in addition, to emphasize the importance of the *title* filed, terms co-occur in the *title* field are doubly weighted.

The method introduced in Section 3.2 is used to gain the concept clusters. This value-added service considers the co-occurrences in the same sentence, and in the same *keyword* and/or *title* fields of a research paper. As mentioned previously, Log Likelihood Ratio (LLR) is used for computing the relationship between terms *i* and *j*, r_{ij}; in addition, to emphasize the importance of the *title* and *keyword* fields, the semantic relationship of terms co-occurring in the *title* and/or *keyword* fields equals to 1.0.

After concept clustering, each research cluster C_k can be represented as a vector.

$$C_k = (t_{k1}, t_{k2}, t_{k3}, ..., t_{kn}), \text{where } n \text{ is the number of terms,} \quad (12)$$

and

$$t_{ki} = \begin{cases} 1, & \text{if term } i \in C_k \\ 0, & \text{Otherwise} \end{cases}$$

Similar to Section 4.2.2, labeling is also applied for helping users to realize the concept of each research cluster. Table 2 shows the labels, keywords, and keyword weights of a few example research cluster.

Table 2. Concept clusters and their keywords and relative weights.

Concept label	Keywords	Keyword weight
Mobile Computing	MANET mobile ad hoc network route mobile computing wireless network	351.1895 152.5828 140.446 126.5852 95.5504
Genetic Algorithm	genetic algorithm dynamic linkage discovery particle swarm optimization GA economic dispatch	97.3025 66.415 53.332 41.3458 27.0827
Neural Network	neural network optimization problem constraint energy function	72.7728 49.7963 43.1446 38.9312

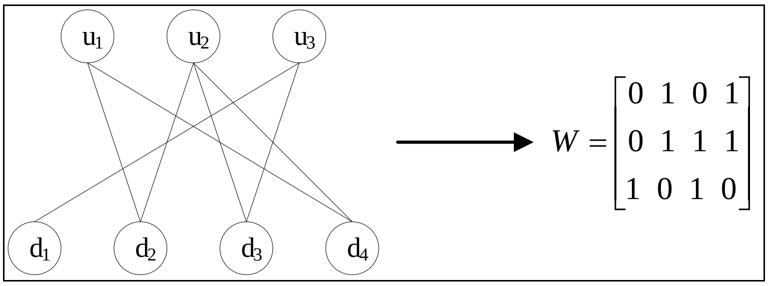
5.3. Virtual Community Construction

5.3.1. Author Social Network

Author social network is constructed from the co-author relationship, and the relative degree of two authors is calculated by Jaccard coefficient [28]. First, the author-paper relationship can be represented as a matrix, W_{ij}. As illustrated in Fig. 10, there are three authors, u_1 ~ u_3, four papers, d_1 ~ d_4, and each edge connecting an author and a paper represents the author-paper relationship.

$$W_{ij} = \begin{cases} 1, & \text{if user } i \text{ is one of the authors in paper } j \\ 0, & \text{Otherwise} \end{cases} \quad (13)$$

The co-author relationship, S, can be computed by $S = W \times W^T$, and each element S_{ij} represents the number of papers co-authored by i and j. Fig. 11 is the co-author relationship of Fig. 10.



$$\longrightarrow W = \begin{bmatrix} 0 & 1 & 0 & 1 \\ 0 & 1 & 1 & 1 \\ 1 & 0 & 1 & 0 \end{bmatrix}$$

Fig. 10. The relationship graph and matrix of authors and papers.

$$\boxed{S = W \times W^T = \begin{bmatrix} 0 & 1 & 0 & 1 \\ 0 & 1 & 1 & 1 \\ 1 & 0 & 1 & 0 \end{bmatrix} \times \begin{bmatrix} 0 & 0 & 1 \\ 1 & 1 & 0 \\ 0 & 1 & 1 \\ 1 & 1 & 0 \end{bmatrix} = \begin{bmatrix} 2 & 2 & 0 \\ 2 & 3 & 1 \\ 0 & 1 & 2 \end{bmatrix}}$$

Fig. 11. Co-author matrix of Fig. 10.

This study applies Jaccard coefficient to measure the similarity between two authors, and it is defined as the ratio of the number of co-author papers over the number of the union of each author's publication.

$$J(U_i, U_j) = \frac{S_{ij}}{|S_i| + |S_j| - S_{ij}} \tag{14}$$

Where $|S_i|$ and $|S_j|$ represents the amount of papers authored by author i and j, respectively. Applying Jaccard coefficient can also normalize the main diagonal of S to 1. Fig. 12 is the S after applying Jaccard coefficient, named S'.

As the co-authoring relationship may affect each author too much, a parameter α ($0 \leq \alpha \leq 1$) can be added to adjust the co-author relationship matrix, S'. If α is closer to zero, the co-authoring relationship exerts less effect to each author. The adjusted co-author relationship matrix is R, and each element R_{ij} can be represented as follows.

$$R_{ij} = \begin{cases} 1 & \text{, if } i = j \\ \alpha \times S_{ij} & \text{, otherwise} \end{cases} \tag{15}$$

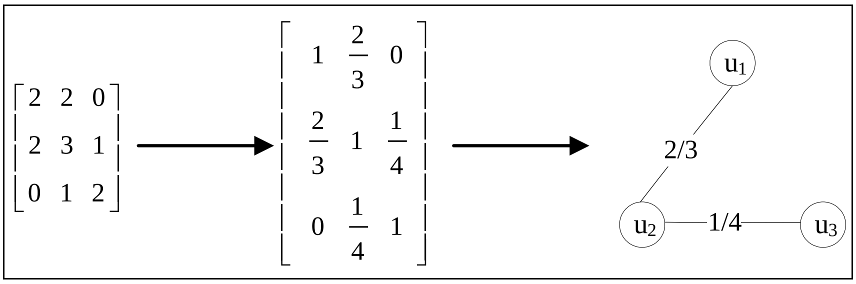

$\begin{bmatrix} 2 & 2 & 0 \\ 2 & 3 & 1 \\ 0 & 1 & 2 \end{bmatrix}$
$\begin{bmatrix} 1 & \frac{2}{3} & 0 \\ \frac{2}{3} & 1 & \frac{1}{4} \\ 0 & \frac{1}{4} & 1 \end{bmatrix}$
u1
2/3
u2
1/4
u3

Fig. 12. Co-author matrix normalized by Jaccard coefficient and relative social network graph.

5.3.2. Author Clustering and Paper Recommendation

This study models authors and terms using the author model. The author model *U* is represented as an *N*×*m* matrix, where *N* is the number of authors and *m* is the number of terms. Each row in *U* shows how an author uses terms in his/her research papers. For taking the co-author factor into account, the updated author model *U'* is the production of the adjusted co-author relationship matrix (*R*) and *U*. Each row in *U'* represents the relationship of an author and all terms when the co-author relationship is considered.

$$U = \begin{bmatrix} w_{11} & \cdots & w_{1m} \\ \vdots & \ddots & \vdots \\ w_{N1} & \cdots & w_{Nm} \end{bmatrix}$$

$$U' = R \times U = \begin{bmatrix} R_{11} & \cdots & R_{1N} \\ \vdots & \ddots & \vdots \\ R_{N1} & \cdots & R_{NN} \end{bmatrix} \times \begin{bmatrix} w_{11} & \cdots & w_{1m} \\ \vdots & \ddots & \vdots \\ w_{N1} & \cdots & w_{Nm} \end{bmatrix} = \begin{bmatrix} w'_{11} & \cdots & w'_{1m} \\ \vdots & \ddots & \vdots \\ w'_{N1} & \cdots & w'_{Nm} \end{bmatrix} \quad (16)$$

The similarity of author *j* and each concept, SU_{jk}, can be calculated by cosine similarity, and it also represents author *j*'s preference for concept *k*. The set of SU_{jk} is SU_j, the concept preference vector of author *j*.

$$SU_j = \{SU_{jk} \mid SU_{jk} = sim(U'_j, C_k), k = 1,2,\ldots,p\}, \quad (17)$$

where *j* = 1, 2, …, *N*; *p* is the total number of clusters

U'_j is a row of *U'*, representing the relationship of author *j* and all keywords. C_k, introduced in Section 4.3, is the vector representation of cluster *k*.

If SU_{jk} is larger than a pre-defined threshold, it means that author *j* has an apparent preference for concept *k*, and author *j* should be assigned to concept *k*. Each author cluster represents a virtual research community with a specific concept. The threshold used in this work is the average value of the similarity of an author and a cluster.

Virtual research communities and co-author relationship can be used to assist users in knowing the distribution of virtual research communities and identify significant researchers of each community. Furthermore, with the graph representation of virtual communities, SNA metrics mentioned in Section 2.2 can be computed[7].

While virtual research communities are constructed, authors in the same virtual research community have similar research interests. Researchers can be recommended according to the relationship of each community and their information need. In this work, two approaches are proposed for recommending papers to authors in the NCTUIR, and the two approaches can be easily extended to any users who are interested in the papers collected in the NCTUIR.

1. Recommendation by personal interests

If an author wants to find out some papers related to his/her research interests, the system can recommend him/her the top n (n=5 in this work) similar papers from the author's virtual communities, except the author's previous works. Cosine similarity of the author model and papers in the author’s communities is used to find out the top n papers for recommendation.

2. Recommendation by community

If an author wants to find out some representative papers of a specific virtual research community, the system can recommend him/her the top n (n=5 in this work) relevant papers. The relevant degree of a paper and a community is calculated by the cosine similarity of the paper and the community.

6. Evaluation

This section describes experiment setting and evaluation measurements for the two services presented in Section 4 and Section 5. In addition, some discussions about the proposed approach and legacy methods are also elaborated.

[7] Due to space limitation, the description of SNA metrics for the authors in the NCTUIR is omitted.

6.1. Evaluation for Graphical Representation of Search Results

The corpus containing 541 "Information Retrieval"-related documents is employed to compare our proposed approach with Topic Keyword Clustering. This study applies three evaluations. Firstly, domain experts classify documents into groups in advance, and then the clustering results of the proposed approach and Topic Keyword Clustering are evaluated with these paper classes by entropy, purity, recall and f-measure [29]. The second evaluation compares the compactness and separation of clustering results. In the last evaluation, the similarity values of papers calculated by the proposed approach and Topic Keyword Clustering are compared with the labelling decided by domain experts.

6.1.1. Entropy, Purity, Recall and F-measure

Domain experts classify all papers into 35 groups in advance, and four common evaluation methods – purity, recall, entropy and f-measure, are applied to evaluate the clustering results. The evaluation results are shown in Table 3, and our method is better than Topic Keyword Clustering (less is better in Entropy estimation).

Table 3. Evaluation of clustering results by proposed method and Topic Keyword Clustering.

# of clusters	Topic Keyword Clustering			Our Method		
Estimation	50	75	100	50	75	100
Purity	0.3008	0.4084	0.5259	0.5359	0.6096	0.6295
Recall	0.3239	0.2569	0.2138	0.4709	0.3811	0.3298
Entropy	2.6289	2.0581	1.4362	1.7427	1.348	1.1622
F-measure	0.2472	0.2971	0.3078	0.4339	0.4643	0.4586

6.1.2. Compactness and Separation Degrees

Table 4 shows the compactness (*Cmp*), separation (*Sep*), and overall cluster quality (*Ocq*) of clustering results [30].

$$Ocq(\beta) = \beta \cdot Cmp + (1-\beta) \cdot Sep \tag{18}$$

This study treats compactness and separation with the same weight, and chooses *β* = 0.5, and smaller *Ocq* value is better. Although the separation of 75 and 100 clusters using Topic Keyword Clustering is better, our method still leads in the overall cluster quality.

Table 4. Compactness, separation degree, and overall cluster quality of clustering results.

	Topic Keyword Clustering			Our Method		
# of clusters	50	75	100	50	75	100
Compactness	0.9217	0.8520	0.7841	0.4914	0.4350	0.3806
Separation	0.6086	0.4820	0.4531	0.5031	0.4991	0.4911
Overall cluster quality	0.7652	0.6670	0.6186	0.4973	0.4671	0.4358

6.1.3. Document Pair Similarity

In this evaluation, two domain experts were asked to label the similarity of 200 random document pairs, and the results is depicted in Table 5. Kappa statistics [31] is applied to evaluate the agreement, and the kappa value is 0.8584 (confidence level: 95%, confidence interval: 0.7893~0.9305), which means high agreement [32]. Consequently, the 186 documents (97+89) agreed by both experts are used for similarity evaluation. Topic Keyword Clustering and our method are evaluated and compared with domain experts' judgement.

Table 6 shows the evaluation results, and according to this table, sensitivity, specificity and accuracy can be computed [33], as illustrated in Table 7. It is easy to find that our method is better than Topic Keyword Clustering.

Table 5. Similarity agreement results of domain experts.

		Expert A		
		No	**Yes**	**Total**
Expert B	**No**	97 (90.65%)	10 (9.35%)	107 (53.5%)
	Yes	4 (4.3%)	89 (95.7%)	93 (46.5%)
Total		101 (51.5%)	99 (49.5%)	200

Table 6. Similarity agreement evaluation.

Similarity Evaluation		Domain Experts	
		Yes	No
Topic Keyword Clustering	Yes	20	5
	No	69	92
Our Method	Yes	57	3
	No	32	94

Table 7. Accuracy of clustering results and domain experts' labeling.

	Sensitivity	Specificity	Accuracy
Topic Keyword Clustering	0.2247	0.9485	0.6022
Our Method	0.6405	0.9691	0.8118

6.1.4. Discussion

According to 6.1.1 – 6.1.3, our method is better than Topic Keyword Clustering. The differences of the two methods are as shown in Table 8, and the major differences are semantic correlation and the document clustering algorithm.

Table 8. Difference between proposed approach and Topic Keyword Clustering.

	Topic Keyword Clustering	**Proposed Approach**
Semantic Correlation	Mutual Information	Log likelihood ratio
Keyword Clustering	*k*-nearest neighbor graph approach	
Document Clustering	Cosine similarity measurement	Bisection *k*-Means

Topic Keyword Clustering uses mutual information (MI) to compute the semantic correlation, and our method uses LLR. The main reason is that MI is proper for judging the independent degree of two terms, but not proper for the relative (dependent) degree of two terms [34]. Table 9 shows an instance of concept subgroups computed by MI and LLR. The concept subgroup computed by MI contains more features, but the concepts of features are less identical, such as "part-of-speech" and "markov model". On the other hand, by LLR, these two feature terms, "part-of-speech" and "markov model", belongs to different concept subgroups, and the concept of each concept subgroup are more consistent.

Table 9. Concept subgroups computed by MI and LLR.

mutual information	**Log likelihood ratio**
part-of-speech tagging, tagger, institute, markov, text segmentation, markov model, estimation, disambiguation, information retrieval ir, workshop, resolution, tagging, rule-based, probability, series	part-of-speech tagger, part-of-speech tagging, text segmentation, story, tagger
	markov, markov model, threshold, relevance feedback

When a document contains multiple concepts, the clustering algorithm used in our method (Bisection k-Means) has better results than Topic Keyword Clustering (cosine similarity). The main reason is that Topic

Keyword Clustering only maps a document to the most similar cluster, but our method maps a document to a similarity vector of every concept subgroup. Thus, if two documents contain common or similar concepts, these two documents can be considered as similar documents. Table 10 is a real example, documents A and B are two search results using "machine learning" as query. In the experiment, our proposed method clusters documents A and B together, but Topic Keyword Clustering assigns them into different groups.

Table 10. Two searching result documents using “machine learning” as keyword.

A	Title: using reinforcement learning to spider the web efficiently
	consider the task of exploring the web in order to find pages of a particular kind or on a particular topic. this task arises in the construction of search engines and web knowledge bases. this paper argues that the creation of efficient web spiders is best framed and solved by reinforcement learning, a branch of machine learning that concerns itself with optimal sequential decision making. one strength of reinforcement learning is that it provides a formalism for measuring the utility of actions that give benefit only in the future. we present an algorithm for learning a value function that maps hyperlinks to future discounted reward by using naive bayes text classifiers. experiments on two real-world spidering tasks show a three-fold improvement in spidering efficiency over traditional breadth-first search, and up to a two-fold improvement over reinforcement learning with immediate reward only. keywords: reinforcement learning, text classification, world wide web, spidering,
B	Title: A machine learning approach to building domain-specific search engines
	domain-specific search engines are becoming increasingly popular because they offer increased accuracy and extra features not possible with general, web-wide search engines. unfortunately, they are also difficult and time-consuming to maintain. this paper proposes the use of machine learning techniques to greatly automate the creation and maintenance of domain-specific search engines. we describe new research in reinforcement learning, text classification and information extraction that enables efficient spidering, populates topic hierarchies, and identifies informative text segments. using these techniques, we have built a demonstration system: a publicly-available search engine for computer science research papers.

6.2. Evaluation for Virtual Community Generation and Paper Recommendation

The corpus used for virtual research community generation contains 235 authors and 226 papers stored in the NCTUIR, and all the authors and papers are computer science major. The evaluation contains two parts. The first one evaluates the result of author clustering, and the other one evaluates the accuracy of paper recommendation.

Table 11. Research concept clusters and relative labels and representative keywords.

Class label	Cluster label
Network Communication	Mobile Computing Routing Protocol PIM-SM Bandwidth Requests TCP Network Management
Artificial Intelligence	Genetic Algorithm Network Motif Brick Motif Content Analysis Neural Network SPDNN Divide-and-conquer Learning
Computer Graphics	Content-based Image Retrieval Watershed Segmentation Toboggan Approach
Information Retrieval	Semantic Query Content Management
Computer System	Memory Cache Parallel Algorithm
Information Security	End-to-end Security
Graph Theory	Interconnection Network
Software Engineering	Reliability Analysis

6.2.1. Evaluation of Author Clustering

The NCTUIR does not provide classifications of research areas in advance, so this study asks two domain experts to partition the formed concept clusters into classes. Domain experts create a total of 8 classes, and Table 11 lists the class labels and the labels of the contained clusters for all the 8 classes.

After concept classes are created, three domain experts then assign all the 235 authors into these classes, and the result is shown in Table 12. If there's an author not similar to any research area class, he/she would be assigned to "others" class.

Table 12. Author classification by domain experts.

Class label	# of authors
Network Communication	111
Artificial Intelligence	28
Information Retrieval	7
Computer System	6
Computer Graphics	23
Information Security	10
Graph Theory	29
Software Engineering	4
Others	17
Total	235

The correctness of author clustering is evaluated by precision and recall [35]. The relative degree of two authors is calculated by Jaccard coefficient, and is adjusted by a parameter, α, whose range is from 0 to 1. If α is 0, it means that the author clustering process does not take social relationship into consideration.

Table 13. Recall and precision of author clustering with different *α* value.

***α* value**	0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1
Precision	0.7071	0.6917	0.6981	0.7107	0.7172	0.7209	0.7209	0.7209	0.7209	0.7209	0.7209
Recall	0.6271	0.7606	0.7785	0.7839	0.7817	0.7828	0.7828	0.7828	0.7828	0.7828	0.7828

Table 13 and Fig. 13 show the recall and precision while adjusting *α* from 0 to 1 in steps of 0.1. According to the results, precision is about 0.7 and various *α* values do not significantly affect precision. Recall is about 0.78 while *α* is larger than 0.3. These show our approach is able to cluster authors to proper research communities. However, when *α* is small (<0.3), the recall is apparently worse, so 0.3 is chosen as the parameter of Jaccard coefficient for the following evaluation.

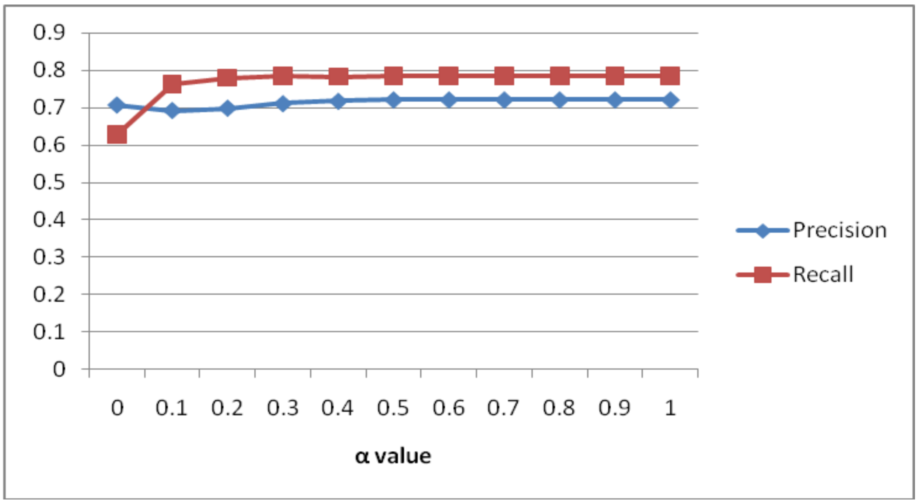


Fig. 13. Graph of recall and precision of author clustering with different *α* value.

6.2.2. Accuracy of Paper Recommendation

This evaluation uses different scenarios for paper recommendation. Each scenario represents a single user with identical information need. The recommended papers are evaluated by two domain experts, and Kappa Statistics is used for calculating the accuracy of paper recommendation. There are 219 papers recommended to two domain experts for evaluation, and the relevance judgement results of two domain experts is as Table 14.

Table 14. Relevance judgment results of domain experts.

		Expert A				**Total**	
		No		**Yes**			
Expert B	**No**	21	(9.6%)	9	(4.1%)	30	(13.7%)
	Yes	2	(0.9%)	187	(85.4%)	189	(86.3%)
Total		23	(10.5%)	196	(89.5%)	219	

The kappa value is 0.764, and the strength of agreement is substantial. Besides, for the 208 papers with the same relevance judgements of the two domain experts, the accuracy, 187/208 = 89.9%, is apparently good.

7. Conclusion and Future Work

This study shows the application of concept extraction and clustering in IR systems. The contribution is two fold. Firstly, the modified concept extraction and clustering method outperforms the original Topic Keyword Clustering. Secondly, from the application aspect, the proposed method facilitates the organization and visualization of search results, and the provision of virtual research communities and paper recommendation for researchers.

Concept extraction and clustering is the core mechanism of this study. A concept is defined as a group of terms whose co-occurrence statistics reveal their similarity in semantics. The proposed approach modifies the original Topic Keyword Clustering. It treats terms and the relationships of term pairs as vertices and edges of a graph. Removing insignificant vertices and edges, and merging similar concept sub-groups generate the concept groups. The proposed method has better performance than Topic Keyword Clustering, and the evaluation results provide evidences. The main reason is, instead of MI used in Topic Keyword Clustering, LLR is applied to calculate semantic correlation. Furthermore, Topic Keyword Clustering assigns a document to a single concept group; on the other hand, the proposed method may assign a document to more than one concept groups with high similarity.

To well organize search results, concepts should be extracted from the corpus in advance. Then, documents are clustered according to the similarities of documents and concepts. Citation relations and labelling are also applied for assisting users to realize the concept of each cluster and the intra- and inter-relations among clusters.

For providing virtual research communities and paper recommendation, the author model is created to maintain author-term relationships. Second, the social information, i.e. co-author relationship, is applied to adjust the author model. Then, author clustering is done by assigning authors to concept groups according to the similarity of the updated author model and concept groups. Finally, each author cluster is a virtual research community, and each virtual community can provide users with information on representative researchers and papers. Besides, according to user's research interests, proper papers in a virtual community can be recommended.

The proposed concept extraction and clustering method generates concepts in a flat level and do not consider the hierarchical structure inside a concept group. With a hierarchical structure, users are able to know more details of a research concept when they track down deeply. For this purpose, the future study will track the process of concept group generation and the relationships of subgroups, and use ontology to model the hierarchical structure of groups. Besides, although at least two domain experts are asked to do qualitative evaluation, we will apply more experts for more rigorous evaluation in the future.

References

1. Manning, C. D., Raghavan, P. and Schütze, H. (2008), *Introduction to Information Retrieval*, Cambridge University Press, pp. 349-375.
2. Garfield, E. (1979). Citation indexing – Its theory and application in science, technology, and humanities. New York, NY: John Wiley & Sons.
3. Gan, G., Ma, C. and Wu, J. (2007), "Data Clustering: Theory, Algorithms, and Applications", SIAM, Society for Industrial and Applied Mathematics.
4. MacQueen, J. (1967), "Some methods for classification and analysis of multivariate observations", in *Proceeding of Fifth Berkeley Symposium on Mathematical Statistics and Probability*, pp. 281-297.
5. Killani, R., Rao, K. S., Satapathy, S. C., Pradhan, G. and Chandran, K. R. (2010), “Effective Document Clustering with Particle Swarm Optimization”, *Swarm, Evolutionary, and Memetic Computing, Lecture Notes in Computer Science*, Vol. 6466/2010, pp. 623-629
6. Taghva, J. and Veni, R. (2010), "Effects of Similarity Metrics on Document Clustering", in *Seventh International Conference on Information Technology*, pp. 222-226.
7. Oikonomakou, N. and Vazirgiannis, M. (2010), “A Review of Web Document Clustering Approaches”, *Data Mining and Knowledge Discovery Handbook*, Part 6, pp. 931-948
8. Mahdavi, M. and Abolhassani, H. (2009), “Harmony K-means algorithm for document clustering”, *Data Mining and Knowledge Discovery*, Vol. 18, No. 3, pp. 370-391
9. Zhang, X., Jing, L., Hu, X., Ng, M. and Zhou, X. (2007), “A Comparative Study of Ontology Based Term Similarity Measures on PubMed Document Clustering”, *Advances in Databases: Concepts, Systems and Applications, Lecture Notes in Computer Science*, Vol. 4443/2007, pp. 115-126.
10. Kogan, J. (2006), Introduction to Clustering Large and High-Dimensional Data, Cambridge University Press.
11. Iliopoulos, I., Enright, A. J. and Ouzounis, C. A. (2001), "TEXTQUEST: Document Clustering of MEDLINE Abstracts For Concept Discovery In Molecular Biology", in *Proceeding of Pacific Symposium on Biocomputing 2001*, pp. 384-395
12. Slonim, N., Tishby, N. (2000), "Document clustering using word clusters via the information bottleneck method", in Proceedings of the 23rd annual international ACM SIGIR conference on Research and development in information retrieval in Athens, Greece, ACM, New York, NY, USA, pp. 208-215.
13. Athanasios P. (1991), Probability, Random Variables and Stochastic Processes, McGraw-Hill, New York, NY.
14. Chang, H. C. and Hsu, C. C. (2005), "Using topic keyword clusters for automatic document clustering", *IEICE Transaction on Information System*, Vol. E88-D No. 8, pp. 1852-1860.
15. Moreno, J. L. (1934), *Who Shall Survive?*, Mental Health Resources.

16. Liu, X., Bollen, J., Nelson, M. L. and Van de Sompel, H. (2005), "Co-authorship networks in the digital library research community", *Information Processing and Management*, Vol. 41 No.6, pp. 1462-1480.
17. Wasserman, S. and Faust, K. (1994). “Social network analysis: Methods and applications”, Cambridge: Cambridge University Press.
18. Tyler, J. R., Wilkinson, D. M. and Huberman, B. A. (2003), "Email as spectroscopy: Automated discovery of community structure within organizations", in Huysman, M. H., Wenger, E. and Wulf, V. (Ed.), *Communities and technologies*, Springer, pp. 81-96.
19. Mika, P. (2005), "Flink: Semantic Web technology for the extraction and analysis of social networks", *Web Semantics: Science, Services and Agents on the World Wide Web*, Vol. 3 No. 2-3, pp. 211-223.
20. Page, L. and Brin, S. (1998), "The anatomy of a large-scale hypertextual Web search engine", *Computer Networks and ISDN Systems*, Vol. 30 No. 1-7, pp. 107-117.
21. Matsuo, Y., Mori, J. and Hamasaki, M. (2007), "POLYPHONET: An advanced social network extraction system from the Web", *Web Semantics: Science, Services and Agents on the World Wide Web*, Vol. 5 No. 4, pp. 262-278.
22. Lehmann, L. E. (1986), *Testing Statistical Hypotheses*, Wiley.
23. Neyman, J. and Pearson, E. (1967), *Joint statistical papers*, Hodder Arnold.
24. Gowda, K. C. and Krishna, G. (1978), "Agglomerative clustering using the concept of mutual nearest neighbourhood", *Pattern Recognition*, Vol. 10 No. 2, pp. 105-112.
25. Steinbach, M., Karypis, G. and Kumar, V. (2000), "A comparison of document clustering techniques", paper presented at the KDD Workshop on Text Mining, August 20-23, Boston, MA, USA, available at: http://www.cs.cmu.edu/~dunja/KDDpapers/Steinbach_IR.pdf (accessed 12 May 2009).
26. De Bellis, N. (2009), Bibliometrics and citation analysis: from the science citation index to cybermetrics, The Scarecrow Press, pp. 156-166.
27. Chan, S. E., Pon, R. K. and Cárdenas, A. F. (2006), "Visualization and Clustering of Author Social Networks", in International Conference on Distributed Multimedia Systems Workshop on Visual Languages and Computing in Grand Canyon, Arizona, USA, 2007, pp. 30-31.
28. Shah, M. A. (1997), ReferralWeb: A resource location system guided by personal relations, Master's thesis, M.I.T.
29. Rosell, M., Kann, V. and Litton, J. (2004), "Comparing comparisons: Document clustering evaluation using two manual classifications", in *Proceeding of International Conference on Natural Language Processing*, Allied Publishers Pvt. Ltd., pp. 207-216.
30. He, J., Tan, A., Tan, C. L. and Sung, S. Y. (2003), "On quantitative evaluation of clustering systems", in Wu, W., Xiong, H. and Shekhar S. (Ed.), *Clustering and Information Retrieval*, Springer, pp. 105-134.
31. Cohen, J. (1960), "A coefficient of agreement for nominal scales", *Educational and Psychological Measurement*, Vol.20 No.1, pp. 37-46.
32. Landis, J. R. and Koch, G. G. (1977), "The measurement of observer agreement for categorical data", *Biometrics*, Vol. 33, pp. 159-174.
33. Han, J. and Kamber, M. (2000), *Data Mining: Concepts and Techniques*, Morgan Kaufmann, San Fransisco, CA, USA.
34. Manning, C. D. and Schutze, H. (1999), *Foundations of Statistical Natural Language Processing*, MIT Press, Cambridge, MA, USA.

35. Makhoul, J., Kubala, F., Schwartz, R. and Weischedel, R. (1999), "Performance measures for information extraction", in *Broadcast News Workshop '99 Proceedings*, Morgan Kaufmann, San Fransisco, CA, USA, pp. 249-252.

Shihn-Yuarn Chen received his B.S. and M.S. degree in Computer Science and Information Engineering from National Chiao Tung University, Taiwan, R.O.C. Now he is a Ph.D. candidate in Computer Science, National Chiao Tung University. His research interests include digital library, information retrieval, social tagging and Web 2.0.

Chia-Ning Chang received her B.S. and M.S. degree in Computer and Information Science, from National Chiao Tung University, Taiwan, R.O.C. She is currently employed as a RD-DE-PCM engineer in Cyberlink Corp. in Taiwan.

Yi-Hsiang Nien received his M.S. degree in Institute of Information Management, from National Chiao Tung University, Taiwan, R.O.C. He is currently working as an engineer in Taiwan Semiconductor Manufacturing Company Ltd. in Taiwan

Hao-Ren Ke received his B.S. degree in 1989 and Ph.D. degree in 1993, both in Computer and Information Science, from National Chiao Tung University, Taiwan, R.O.C. Now he is a professor of the Graduate Institute of Library and Information Studies and the deputy library director in National Taiwan Normal University. His main research interests are in digital library/archives, library and information system, information retrieval, and Web mining.

Received: November 24, 2010; Accepted: May 17, 2011