Hierarchical Topic-based Communities Construction for Authors in a Literature Database

Chien-Liang Wu and Jia-Ling Koh*

Department of Information Science and Computer Engineering,
National Taiwan Normal University, Taipei, Taiwan, R.O.C.
wucl@ice.ntnu.edu.tw, jlkoh@csie.ntnu.edu.tw

Abstract. In this paper, given a set of research papers with only title and author information, a mining strategy is proposed to discover and organize the communities of authors according to both the co-author relationships and research topics of their published papers. The proposed method applies the CONGA algorithm to discover collaborative communities from the network constructed from the co-author relationship. To further group the collaborative communities of authors according to research interests, the CiteSeer^X is used as an external source to discover the hidden hierarchical relationships among the topics covered by the papers. In order to evaluate whether the constructed topic-based collaborative community is semantically meaningful, the first part of evaluation is to measure the consistency between the terms appearing in the published papers of a topic-based collaborative community and the terms in the documents related to the specific topic retrieved from other external source. The experimental results show that 81.61% of the topic-based collaborative communities satisfy the consistency requirement. On the other hand, the accuracy of the discovered sub-concept relationship is verified by checking the Wikipedia categories. It is shown that 75.96% of the sub-concept terms are properly assigned in the concept hierarchy.

Keywords: Social Network, Community Mining, Bibliographic database.

1. Introduction

In the area of social network analysis, one important issue is community discovery. A community in a social network is usually defined to be a densely connected sub-graph in the network. By detecting communities, it helps us to understand and exploit the networks more effectively. Especially in a bibliographic database, identifying communities from a co-authorship network can reveal academic activities as well as evolution of research areas; discovering communities in citation network can demonstrate the information diffusion within and between different research areas.

There have been several community mining algorithms proposed to identify meaningful communities from networks. These algorithms can be broadly classified into two main categories: graph partitioning based approaches [2, 12, 13] and modularity based approaches [3, 4, 11, 15]. Identifying the communities within a network has become one of the major concerns of social network analysis which has various applications. In this paper, we are interested to discover topic-based collaborative communities from co-authorship network.

This work was partially supported by the R.O.C. N.S.C. under Contract No. 98-2221-E-003-017 and NSC 98-2631-S-003-002.

Several studies have been proposed to analyze bibliographic databases. Zhang et al. [15] proposed the SSN-LDA model to discover flat communities from social networks by utilizing the topological information in social networks. Deng et al. [1] proposed a novel graph-based re-ranking model to improve the ranking of the retrieved documents with respect to a given query. Then the model was used to discover experts of a specific topic from the DBLP bibliographic data. Zaiane et al. [14] provided a new random walk approach to discover research communities with potentially collaborative relationship from the DBLP database [9]. An extended bipartite graph is built to model the relationships of authors and conferences. In order to include the topic information, the proposed model is further extended to be a tripartite graph. Then the random walk with restart algorithm was revised to calculate the relevance scores among researchers in the graph to group the highly-relevant researchers into the same community. Mei et al. [10] proposed the NetPLSA model which combined the statistical topic modeling and social network analysis to discover topical communities. The PLSA model proposed in [6] was exploited to get the weights of the predefined topics for each author. Thus, the topic similarity between each pair of researchers can be evaluated by comparing their topic weights. Moreover, the Harmonic function is used to evaluate the degree of collaborative relationship among each pair of researchers. Accordingly, an objective function is defined by integrating these two models. The topical communities are then discovered by minimizing the objective function.

The previous works [14] and [10] mentioned above are closely related to our work. In [14], for the members in a community discovered from the extended bipartite graph of author-conference relationship, they may have high weighted co-author relationship or often publish papers in the same set of conferences. However, the topics covered in each community are not explicitly specified. Moreover, although the members in a community discovered from the tripartite graph model have similar research topics, it is not necessary that they have strong co-work relationships. Likewise, although [10] regularize a statistical topic model with a harmonic regularization based on a graph structure, the concept hierarchy of topics covered in the communities is not shown explicitly.

In this paper, a mining strategy is proposed for discovering topic-based collaborative community. First, CONGA algorithm [3] is applied to discover overlapping collaborative communities in the collaborative network. By applying the hidden information in the external source CiteSeer^X, the collaborative communities are further organized in a semantic level by automatically constructing a concept hierarchy of topic terms. Therefore, the collaborative communities corresponding to a research topic at arbitrary semantic level can be retrieved. In order to evaluate whether the constructed topic-based collaborative community is semantically meaningful, the first part of evaluation is to measure the consistency between the terms appearing in the published papers of a topic-based collaborative community and the terms in the documents related to the specific topic retrieved from other external source. The experimental results show that 81.61% of the topic-based collaborative communities satisfy the consistency requirement. On the other hand, the accuracy of the discovered sub-concept relationship is verified by checking the Wikipedia categories. It is shown that 75.96% of the sub-concept terms are properly assigned in the concept hierarchy.

The remaining sections of this paper are organized as follows. The related works are discussed in Section 2. Section 3 describes the proposed strategies for discovering the topic-based collaborative communities. The performance study is presented and discussed in Section 4. Finally, Section 5 provides the conclusion and future works.

2 Related Works

Community Discovery. The community mining algorithms can be broadly classified into two main categories: graph partitioning based and modularity based approaches. The partitioning approaches divide the vertices into different communities by minimizing the number of edges between the vertices in different communities, such as the min-max cut algorithm [2], normalized cuts algorithm [12], and spectral clustering algorithm [13]. On the other hand, the modularity based approaches provided a modularity measure to perform good data partitioning during the mining process. One of the representative methods is the Newman algorithm proposed by Newman et al. [11].

The limitation of the Newman algorithm is that it does not allow a node being assigned to more than one community. In reality, an individual may exist in more than one community to take on various roles, such as a blog user being a professional cook and an amateur photographer at the same time. For this reason, Gregory et al. [3] modified the Newman algorithm and proposed the CONGA algorithm, which introduced an operation for splitting a vertex. Suppose a node v is split into v_1 and v_2, a virtual edge is constructed to connect v_1 and v_2. The betweenness centrality of the virtual edge is called the split betweenness of node v. In each iterative step of the CONGA algorithm, either the edge with the maximum edge betweenness is removed or the node with the maximum split betweenness is split, depending on which one is greater. The CONGA algorithm is an extension of the Newman algorithm. Thus, it suffers from the huge computation cost for recalculating the betweenness of the nodes and edges repeatedly. For solving this problem, an improved version of the CONGA algorithm, which is named the CONGO algorithm, was proposed by the same author in [4]. In order to speed up the processing efficiency, instead of traversing every edge in the network globally, a parameter h is given to limit the search region locally when updating the betweenness after an edge was removed or a node was split.

Document Clustering. Most traditional document clustering methods adopted the "Bag of Words" (BOW) model to represent a document. A document is thus represented by a term vector; and the similarity of two documents is measured according to their term vectors. However, this approach ignored the relationships between important terms that do not co-occur in the documents, such as synonyms.

Recently, there is a growing amount of tasks on how to utilizing external background knowledge (e.g. WordNet and Wikipedia) to enhance document clustering [7, 8, 5]. Hotho et al. [7] used WordNet, a general ontology, to represent each document by a concept vector instead of a word vector. Furthermore, [8] and [5] considered Wikipedia is a more comprehensive resource to provide potential ontology which can be exploited for enriching text representation. Therefore, the mapping strategies were developed in [8] to match text documents to Wikipedia topics and further to Wikipedia categories. Then the text documents are clustered not only based

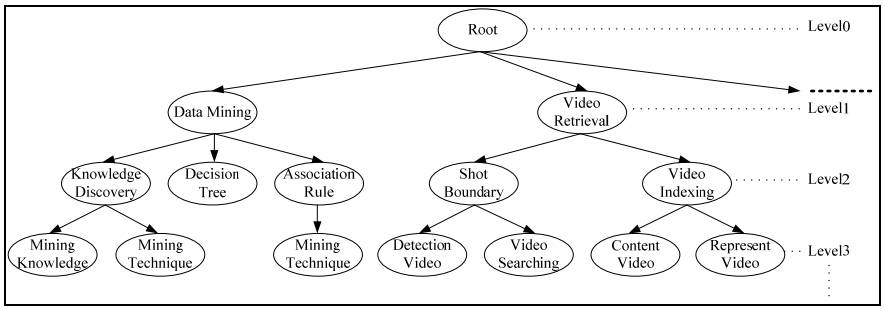

Fig. 1. An example of concept hierarchy of research topics**.**

on the similarity metric of document content but also the concept and category information. In [5], a document was modeled by a graph of terms with semantic links. By providing a semantic relatedness measure of terms according to Wikipedia, the Newman algorithm was performed to discover the communities of terms for extracting key terms in the document.

3. Topic-based Collaborative Communities Discovery

3.1 Problem Definition

According to a given bibliographic dataset, the information of the co-authorship between researchers is modeled as a graph $G(V, E)$. Each node v_i in V represents a researcher. Besides, an edge $e=(v_i, v_j)$ in E connecting two nodes v_i and v_j if the two corresponding researchers have at least one co-publishing paper in the dataset. The constructed graph G is called a *collaborative network*.

A densely connected subgraph in graph $G(V, E)$ is called a *collaborative community*, whose nodes represent the researchers with strong co-work relationships. However, the collaborative communities only consider the co-author relationship as the basis of grouping researchers. In order to organize the researchers in a semantic level, a better way is to group the collaborative communities according to the research topics covered in the communities. Moreover, the research topics usually form a concept hierarchy as the example shown in Fig. 1. If the collaborative communities are further assigned to the concept hierarchy of research topics, a hierarchy of the *topic-based collaborative communities* can be constructed. As a result, the users can access the members in the same community not only by their co-author relationship but also the similar research interests at different concept level.

In a bibliographic dataset, suppose only the information of author, co-authors, and title is available for each publication, the challenge is how to extract the research topics covered in a collaborative community and construct the concept hierarchy of topics automatically.

3.2 Collaborative Community Discovery

We downloaded the bibliographic database from the DBLP website (http://dblp.uni-trier.de/xml/dblp.xml.gz). Each data in the DBLP database contains the names of researchers, published paper, journal/conference, year and other related information.

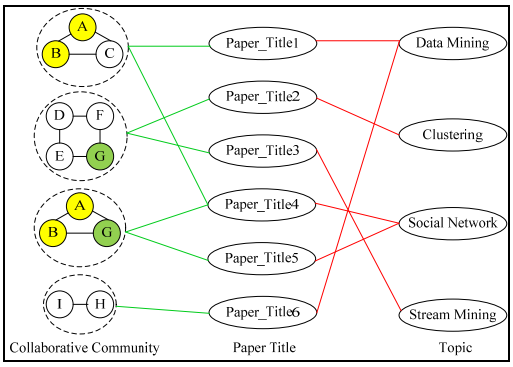

Fig. 2. Collaborative community with corresponding paper topics.

Among the related works of community mining, it is limited that each vertex is assigned to exact one community in most studies. However, in the real world, many researchers have more than one research interest. It is possible that a researcher has ever co-published papers with other researchers in different domains. It is improper to assign the researcher to a collaborative community with specific topic. Therefore, the CONGA algorithm proposed by [3] is used to discover the densely connected subgraphs with overlapping allowed in the graph. As a result, a researcher with multiple research interests will appear in many collaborative communities.

Fig. 2 shows an example of the discovered collaborative community from a collaborative network, where the nodes labeled by A, B, C, etc. represent the researchers. If two researchers have any co-published paper, a solid edge will connect them. Accordingly, the black dotted-line circles imply the collaborative communities discovered by the CONGA algorithm. It is shown that the researchers A, B, and G all belong to more than one collaborative community.

3.2 Concept Hierarchy Construction of Research Topics

A topic-based collaborative community is formed by the collaborative communities with a specific topic. The information of the topic covered by a collaborative community is implicit in the corresponding published papers. As Fig. 2 shows, the nodes in the middle are used to denote the titles of papers. A paper title is connected to the collaborative community which contains all its authors. Besides, the rightmost nodes in the Fig. 2 denote the implicit research topics of the papers, such as “Data Mining”, “Clustering”, and “Social Network” etc. If the research topics and the links between paper titles and topics can be extracted automatically, the discovered collaborative communities can be further grouped according the topics covered in their published papers. Using the collaborative communities {A, B, C} and {H, I} shown in Fig. 2 as an example, the authors in these two collaborative communities have never co-published any paper. However, both of the two collaborative communities have a paper with topic “Data Mining”. Therefore, these two collaborative communities should be grouped into a topic-based collaborative community with topic “Data Mining”.

Under the limited information provided in a bibliographic database, only “paper title” best describes the content covered in a paper. Therefore, we perform the

```
Algorithm CCH
Input: all topic terms, Sub_concept(X) for each topic term X
Output: concept hierarchical paths
  For each topic term X
      Call M_DFS (<X>, Sub_concept(X));
Procedure M_DFS (P, Sub_concept)
  For each item t' in Sub_concept
      If (t' is a sub-concept of all the topic terms in P)
          P' = append item t' to P;
          Call M_DFS (P', Sub_concept(t'));
      else output(P);
```

Fig. 3. Pseudo codes for discovering all the concept hierarchical paths.

following processing to extract the potential topic terms. First, each paper title is processed by the basic text processing steps, including removing stop words and stemming. After that, the bigrams are extracted from the titles. The bigrams with frequency higher than a given threshold α is chosen to be the topic terms.

From the extracted topic terms, it is not easy to determine whether the topics of two papers are related. For example, suppose the title of paper A contains the topic term "Data Mining", while the title of paper B contains "Sequential Pattern". Although these two topic terms are different lexically, it is known that "Sequential Pattern" is an important research issue of "Data Mining" in computer science. For solving this problem, the external source CiteSeer^X is used in our approach to construct the hidden concept hierarchy of the extracted topic terms.

For each pair of topic terms X and Y, the confidences of the association rules X→Y and Y→X are measured to decide whether there exists a hidden sub-concept relationship between X and Y. For getting the confidences of the association rules, each topic term and each pair of topic terms are used as query keywords, respectively, to get the numbers of documents that the two topic terms separate occurrences and co-occurrences in CiteSeer^X. Thus, the *conf*(X→Y) of association rule X→Y is obtained from $|D(X \cap Y)|/|D(X)|$ where $|D(X)|$ denotes the number of documents contain topic term X and $|D(X \cap Y)|$ denotes the number of documents contain both X and Y. If *conf*(X→Y) is less than 1 and *conf*(Y→X) is 1, it is implied that if a document contains Y, it must also contain X, but the inverse is not true. In other words, topic term Y is a sub-concept of topic term X. By considering the noise in real data, when deciding whether a topic term Y is a sub-concept of topic term X, the *criterion* is relaxed to require that both *conf*(X→Y) and *conf*(Y→X) are greater than a given threshold value β, and *conf*(X→Y) is smaller than *conf*(Y→X).

The sub-concept relationship among topic terms is then used to construct a concept hierarchy of the topic terms as shown in Fig. 1. The algorithm for discovering all the concept hierarchical paths of the topic terms is described as the pseudo codes shown in Fig. 3. Let Sub_concept(X) denote the set of detected sub-concepts of a topic term X. If Y is a sub-concept of X, the discovered hierarchy path is denoted as <X, Y>. If Z is both a sub-concept of X and Y, the constructed hierarchy path is denoted as <X, Y, Z>. Initially, the Sub_concept(X) is discovered for each topic term X. The procedure M_DFS() is called to construct all the concept hierarchical paths existing

among the topic terms in a depth-first manner. Let $P=<t_1, t_2, \ldots, t_n>$ denote a discovered hierarchy path, where $t_i(i=1,\ldots, n)$ denote a topic term in the path. A topic term *t'* can be appended to the path only when *t'* is a sub-concept of all the topic terms in *P*.

Since the sub-concept relationship has the transitivity property, only the maximum hierarchical paths have to be maintained. For this reason, the hierarchical paths which are subsequences of any other discovered hierarchical path are removed. Finally, a concept hierarchy of topic terms is then constructed by constructing a prefix tree structure for the discovered hierarchical paths. As shown in Fig. 1, the topic terms located at level 1 represent the most general concepts in the concept hierarchy. The topic terms at level 2 are sub-concepts of their parent. For example, the nodes in the subtree rooted at the node "Data Mining" are all related topic terms in the field of "Data Mining". Besides, the children nodes of the node "Data Mining" represent the sub-concepts in the domain of "Data Mining", such as "Knowledge Discovery", "Decision Tree" and "Association Rule".

Next, according to the established concept hierarchy of topic terms, a collaborative community is assigned to the proper nodes in the concept hierarchy according to its published papers. Let $C_i.tt$ denote the set of topic terms in the titles of the papers whose authors are all in collaborative community C_i. For each topic term t in $C_i.tt$, if the term t exactly matches to the topic term of a node in the concept hierarchy, C_i is then assigned to this node. Otherwise, the following process is performed to look up the most related topic with t. First, among all the topic terms represented by the nodes at level 1 of the concept hierarchy, the topic t_i with the highest $conf(t \rightarrow t_i)$ is identified. If $conf(t \rightarrow t_i)$ is greater than or equal to the given threshold β, the above process will be performed recursively on the nodes in the subtree rooted at the node of topic t_i. The process will continue until the confidence $conf(t \rightarrow t_i)$ for each topic t_i at the level is smaller than the threshold β. Then the collaborative community C_i is assigned to the parent node of t_i. If the highest $conf(t \rightarrow t_i)$ obtained at level 1 is less than the given threshold β, the collaborative community C_i is topic-undetermined according to term t. The task of assigning the collaborative community to the concept hierarchy will repeat until all the terms in $C_i.tt$ have been examined.

For a topic t in the concept hierarchy, the corresponding topic-based collaborative community consists of the members in the collaborative communities which are assigned to the subtree rooted at the node of t. Therefore, if the publishing papers of a collaborative community cover multiple topic terms, the collaborative community will belong to multiple topic-based collaborative communities. For each topic-based collaborative community, the number of papers of a member containing the topic term is divided by the total number of papers assigned to the topic term to get the *participating degree* of the member in the community. Moreover, the *concentrate degree* of the member is obtained by dividing the number of papers of the member containing the topic term into his total number of published papers.

4. Experimental Evaluation

4.1 Testing Dataset

In the experimental evaluation, the testing dataset (named 23-CONF) was used, where

the papers published from 2006 to 2008 in the 23 data mining related conferences were extracted from the DBLP database.

The corresponding co-authorship network is constructed for the testing dataset, first. After performing the CONGA to discover the collaborative communities in the network, the topic terms are extracted and organized to a concept hierarchy. The related information of the testing dataset is as follows: the number of papers in the dataset is 10800, the number of nodes in the constructed network is 17216, and the number of edges is 34961. The threshold value α for filtering potential topic terms is set to be 0.05, and the threshold value β for detecting the hierarchical relationship between topic terms is set to be 0.13. Finally, the number of discovered collaborative communities is 2230, the number of nodes located at level 1 is 42, the average length of the discovered hierarchical paths is 4.7.

4.2 Evaluation Results

4.2.1 Consistency of a Topic-based Collaborative Community

In order to evaluate whether the constructed topic-based collaborative community is semantically meaningful, the first part of evaluation is to measure the consistency between the terms appearing in the published papers of a topic-based collaborative community and the terms in other documents related to the specific topic.

An abstract is a concise version of a paper that includes the paper's research purpose, methodology, experimental results, etc. We thus believe that the abstract of a paper contains more keywords that describe the topic of a paper than the title. Therefore, for each topic term t_i located at level 1 in the discovered concept hierarchy of topic terms, the abstracts of the papers in the corresponding topic-based collaborative community are extracted from CiteSeer^X. There are almost 22% of the papers whose abstracts can be obtained from CiteSeer^X. After performing the text processing steps on the abstracts, the unigrams are extracted from these abstracts to form the set of keywords: B_{Ti}. On the other hand, the ACM (http://www.acm.org) digital library is queried to retrieve the abstracts of the most related 200 papers for the topic term t_i. The unigrams extracted from this set of abstracts form another set of keywords: B_{ti}^{ACM}. Then we use the Jensen-Shannon Divergence (*JSD*) to measure the similarity between the probability distributions of terms in two sets of keywords.

Let $\Pr(b \mid B_{ti})$ denote the probability of keyword b in B_{ti} and $\Pr(b \mid B^{ACM}{}_{tj})$ denote the probability of keyword b in $B^{ACM}{}_{tj}$ The *JSD* measure between B_{ti} and B_{tj}^{ACM} for each pair of topic terms t_i and t_j is shown below.

$$JSD(B_{t_i} \| B_{t_j}{}^{ACM}) = \frac{1}{2}(KLD(B_{ti} \| avg(B_{ti}, B_{t_j}{}^{ACM})) + KLD(B_{t_j}{}^{ACM} \| avg(B_{ti}, B_{t_j}{}^{ACM}))) \quad (1)$$

$$KLD(B_{t_i} \| avg(B_{ti}, B_{t_j}{}^{ACM})) = \sum_{b \in Bt_i} \Pr(b \mid B_{ti}) \log \frac{\Pr(b \mid B_{ti})}{\frac{1}{2}(\Pr(b \mid B_{ti}) + \Pr(b \mid B_{t_j}{}^{ACM}))} \quad (2)$$

When the probability distributions of terms in the two set of documents are more similar, the *JSD* measure will get lower value. It is indicated that the words appearing in the papers assigned to topic t_i is consistent with the papers searched by topic t_j from the ACM digital library.

In the constructed concept hierarchy of topic terms, there are 42 topic terms

located at level 1. For each topic term t_i at level 1, $JSD(B_{ti} \parallel B_{tj}^{ACM})$ is measured with all the terms at level 1 one by one. The measuring results are then sorted in ascending order. The corresponding topic t_j of the top 1 result represents the most consistent topic of t_i in the ACM digital library. The evaluation result shows that 81.61% of the topic terms have themselves as their most consistent topics in the ACM digital library. If the top 2 most consistent topics in the ACM digital library are identified, 84.38% of topic terms themselves are covered. By observing the situations that the most consistent topic of a topic term t_i is another topic term t_j, it usually occurs when the topic-terms are cross-domain such as “Neural Network” and “Data Mining”. The *JSD* between $B_{\{\text{Neural Network}\}}$ and $B_{\{\text{Data Mining}\}}^{ACM}$ is lower than that between $B_{\{\text{Neural Network}\}}$ and $B_{\{\text{Neural Network}\}}^{ACM}$. The reason is that many important keywords in $B_{\{\text{Neural Network}\}}$, such as "Supervised Learning" and "Markov Model", also appear in $B_{\{\text{Data Mining}\}}^{ACM}$. On the other sides, the papers published in the collaborative communities assigned to “Neural Network” do not contain the popular keywords related to "Neural Network", such as "Gaussian Process" and "Fuzzy Logic". Therefore, the probability distributions of keywords between $B_{\{\text{Neural Network}\}}$ and $B_{\{\text{Data Mining}\}}^{ACM}$ is more similar than that between $B_{\{\text{Neural Network}\}}$ and $B_{\{\text{Neural Network}\}}^{ACM}$.

4.2.2 Accuracy of the Sub-concept Relationship

In this part of evaluation, we would like to measure the correctness of the discovered hierarchical path of the topic terms. Since there is no ground truth to compare with the constructed concept hierarchy, we would like to use the external source Wikipedia to validate the accuracy of our discovered sub-concept relationship.

Let T_level2 denote the topic terms located at equal to or larger than level 2. For each topic term t in T_level2, it is used as a query term to search on Wikipedia. Let Category(t) denote the set of categories list in Wikipedia for t. If Category(t) contains any super-concept or which is the re-direction of any super-concept in the concept hierarchical path of t, the term t is considered to be properly assigned in the concept hierarchy. The evaluation result shows that 75.96% of the topic terms in T_level2 are properly assigned in the concept hierarchy.

5 Conclusion and Future Works

In this paper, a mining strategy is proposed for discovering collaborative community with similar research interests. The CONGA algorithm is applied to discover overlapping collaborative communities according to the collaborative relationship of authors. In order to organize the collaborative communities at semantic level, the topic terms are extracted from the paper titles, which are automatically constructed into a concept hierarchy by applying the hidden information in the external source CiteSeer^X. Therefore, the resultant topic-based collaborative community provided a semantic-meaningful and flexible view to explore the communities of authors in a bibliographic database. In the experiment, two evaluation methods are proposed to evaluate the topic consistency of a topic-based collaborative community and accuracy of the discovered sub-concept relationship. The experimental results show that 81.61% of the topic-based collaborative communities satisfy the consistency requirement. Besides, 75.96% of the sub-concept terms are properly assigned in the concept hierarchy.

The evolution analysis of communities and individuals is an interesting issue, which will discover the change of research interests of researchers, the change of the contribution of researchers to a collaborative community, the change of important topic terms. To take the information of publication time into account to detect the dynamic evolution of collaborative communities is under our investigation.

References

1. Deng, H., Lyu, M.R., King, I.: Effective Latent Space Graph-based Re-ranking Model with Global Consistency. In: Proceeding of the Second ACM International Conference on Web Search and Data Mining, pp.212-221 (2009).
2. Ding, C.H.Q., He, X., Zha, H., Gu, M., Simon, H.D.: A Min-max Cut Algorithm for Graph Partitioning and Data Clustering. In: Proceeding of the IEEE International Conference on Data Mining, pp. 107-114 (2001).
3. Gregory, S.: An Algorithm to Find Overlapping Community Structure in Networks. In: Proceeding of the 11th European Conference on Principles and Practice of Knowledge Discovery in Databases, pp. 91–102 (2007).
4. Gregory, S.: A Fast Algorithm to Find Overlapping Communities in Networks. In: Proceeding of the 12th European Conference on Principles and Practice of Knowledge Discovery in Databases, pp. 408–423 (2008).
5. Grineva, M.P., Grinev, M.N., Lizorkin, D.: Extracting Key Terms From Noisy and Multitheme Documents. In: Proceeding of the 18th ACM International Conference on World Wide Web, pp. 661-670 (2009).
6. Hofmann, T.: Probabilistic Latent Semantic Indexing. In: Proceeding of the 22th ACM SIGIR International Conference on Research and Development in Information Retrieval, pp. 50-57 (1999).
7. Hotho, A., Staab, S., Stumme, G.: Wordnet Improves Text Document Clustering. In: Proceeding of the 26th ACM SIGIR International Conference on Semantic Web Workshop, pp. 541–544 (2003).
8. Hu, X., Zhang, X., Lu, C., Park, E.K., Zhou, X.: Exploiting Wikipedia as External Knowledge for Document Clustering. In: Proceeding of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 389-396 (2009)
9. Ley, M.: The DBLP Computer Science Bibliography: Evolution, Research Issues, Perspectives. In: Proceeding of the 9th International Symposium on String Processing and Information Retrieval, pp. 1-10 (2002).
10. Mei, Q., Cai, D., Zhang, D., Zhai, C.: Topic Modeling with Network Regularization. In: Proceeding of the 17th ACM International Conference on World Wide Web, pp. 101-110 (2008).
11. Newman, M.E.J.: Modularity and Community Structure in Networks. In: Proceedings of the National Academy of Sciences of the United States of America, Vol. 103, No. 23., pp. 8577-8582 (2006).
12. Shi, J., Malik, J.: Normalized Cuts and Image Segmentation. In: IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 22, No. 8, pp. 888-905 (2000).
13. White, S., Smyth, P.: A Spectral Clustering Approach to Finding communities in Graphs. In: Proceeding of the SIAM International Data Mining Conference, pages 76-84 (2005).
14. Zaiane, O.R., Chen, J., Goebel, R.: DBConnect: Mining Research Community on DBLP Data. In: Proceeding of the First ACM Workshop on Social Network Mining and Analysis, pp. 74-81 (2007).
15. Zhang, H., Qiu, B., Giles, C.L., Foley, H.C., Yen, J.: An LDA-based Community Structure Discovery Approach for Large-Scale Social Networks. In: Proceeding of the IEEE International Conference on Intelligence and Security Informatics, pp. 200-207 (2007).