Knowl Inf Syst (2013) 36:59–82
DOI 10.1007/s10115-012-0523-8

REGULAR PAPER

Exploring heterogeneous information networks and random walk with restart for academic search

Meng-Fen Chiang · Jiun-Jiue Liou · Jen-Liang Wang · Wen-Chih Peng · Man-Kwan Shan

Received: 19 January 2011 / Revised: 14 March 2012 / Accepted: 14 July 2012 /
Published online: 26 August 2012

Abstract In this paper, we explore heterogenous information networks in which each vertex represents one entity and the edges reflect linkage relationships. Heterogenous information networks contain vertices of several entity types, such as papers, authors and terms, and hence can fully reflect multiple linkage relationships among different entities. Such a heterogeneous information network is similar to a mixed media graph (MMG). By representing a bibliographic dataset as an MMG, the performance obtained when searching relevant entities (e.g., papers) can be improved. Furthermore, our academic search enables multiple-entity search, where a variety of entity search results are provided, such as relevant papers, authors and conferences, via a one-time query. Explicitly, given a bibliographic dataset, we propose a Global-MMG, in which a global heterogeneous information network is built. When a user submits a query keyword, we perform a random walk with restart (RWR) to retrieve papers or other types of entity objects. To reduce the query response time, algorithm Net-MMG (standing for NetClus-based MMG) is developed. Algorithm Net-MMG first divides a heterogeneous information network into a collection of sub-networks. Afterward, the Net-MMG performs a RWR on a set of selected relevant sub-networks. We implemented our academic search and conducted extensive experiments using the ACM Digital Library. The experimental results show that by exploring heterogeneous information networks and RWR, both the Global-MMG and Net-MMG achieve better search quality compared with existing academic search services. In addition, the Net-MMG has a shorter query response time while still guaranteeing good quality in search results.

Keywords Academic search · Bibliographic information networks

M.-F. Chiang · J.-J. Liou · W.-C. Peng (✉)
National Chiao Tung University, Hsinchu, Taiwan
e-mail: wcpeng@cs.nctu.edu.tw

J.-L. Wang · M.-K. Shan
National Chengchi University, Taipei, Taiwan

1 Introduction

On-line bibliographic databases, such as DBLP, ACM Digital Library and CiteSeer, contain a large number of research papers. Furthermore, the literature search services (e.g., Google Scholar) are now widely used to retrieve research papers. The above systems utilize keyword-based approaches to retrieve papers that contain the query term issued. For example, if a user issues a query (e.g., "association rule") in Google Scholar, the top results are given, as shown in Table 1, which contain only those papers with titles that include the query term "association rule". However, some important papers that do not have the term "association rule" in their titles or abstracts cannot be retrieved. For example, one important paper about association rules related to FP-Growth [13], which is not included in the query result.

To determine the importance of papers in a search, research efforts have been dedicated to forming bibliographic information networks from bibliographic datasets [21,35,32]. In bibliographic datasets, there are several types of entities. Each paper is viewed as one type of entity, and each paper is associated with many authors (referred to as author entities). In addition, each paper is included in one conference proceeding or one journal (referred to as conference entities). To identify experts from bibliographic datasets, some techniques exploit co-author relationships from a co-author graph, where each vertex represents an author and the edges indicate co-authorships [35]. Without loss of generality, most techniques derive knowledge from a homogeneous graph in which vertices are of the same entity type. Clearly, the process of deriving homogeneous graphs from bibliographic datasets may result in a loss of certain information and thus cannot truly reflect all types of relationships. More recently, the authors in [27,29] modeled bibliographic databases as multi-type information networks because research papers usually have multiple types of entities. In light of these multiple types of entities in bibliographic information networks, not only the importance of papers but also the importance of other entities, such as authors and conferences, can be determined. However, the prior studies presented in [27,29] did not further explore group-based ranking information in multiple-entity academic searches.

In this paper, we aim to design an on-line academic search that can overcome the drawbacks of keyword-matching approaches and that can determine ranking scores of multiple entities. We first model a bibliographic dataset as a heterogeneous graph model, where each vertex represents one entity type (i.e., author, conference or paper) to fully reflect both identical-type

Table 1 Top-10 relevant papers by Google Scholar for the query "Association Rule"

Rank	Paper title
1	Mining association rules between sets of items in large databases
2	Fast algorithms for mining association rules
3	Integrating classification and association rule mining
4	Fast discovery of association rules
5	An efficient algorithm for mining association rules in large databases
6	Mining generalized association rules
7	Mining quantitative association rules in large relational tables
8	An effective hash-based algorithm for mining association rules
9	Sampling large databases for association rules
10	Discovery of multiple-level association rules from large databases

Table 2 Top-10 relevant papers by Global-MMG for the query "Association Rule"

Rank	Paper title
1	Fast algorithms for mining association rules in large databases
2	Mining association rules between sets of items in large databases
3	Mining generalized association rules
4	An efficient algorithm for mining association rules in large databases
5	An effective hash-based algorithm for mining association rules
6	Mining quantitative association rules in large relational tables
7	Mining sequential patterns
8	Mining frequent patterns without candidate generation
9	Discovery of multiple-level association rules from large databases
10	Dynamic itemset counting and implication rules for market basket data

entity and cross-type entity relationships. Specifically, we borrow the concept presented in [33], which explores mixed media graphs (MMGs) to model a given bibliographic dataset. In [33], the authors utilizes a heterogeneous information networks to model different types of media. In this paper, different entities are viewed as different media as well. Thus, the MMG is one example of a heterogeneous information network. With a heterogeneous information network, we further develop a Random Walk with Restart (RWR) to retrieve relevant papers by utilizing identical-type entity relationships. In this way, relevant papers and other attribute entities can be found using cross-type entity relationships. Finally, the relevant papers and attribute entities are ranked according to their probabilities for the query term. By adopting the RWR algorithm, relevant papers are determined by multiple relationships to the query term through heterogeneous information networks and the RWR so that relevant papers that do not contain exactly the same term may be discovered. In this paper, a Global-MMG is proposed with the input graph as a bibliographic dataset. For example, Table 2 shows the relevant papers retrieved by the Global-MMG for the query "Association Rule". As shown in Table 2, by considering multiple types of relationships to search for relevant documents, even though some papers that do not have the query term "Association Rule", Global-MMG can still retrieve relevant papers.

In applying RWR algorithm in an on-line academic search, a larger graph in the Global-MMG clearly results in poor performance. Therefore, we divide a heterogeneous information network into a collection of sub-networks such that each sub-network is related to a specific topic domain. As such, a set of sub-networks that are most relevant to the query terms is extracted for performing the RWR to determine the ranking scores of the entity objects. Accordingly, we propose the Net-MMG, which can significantly reduce the query response time for an on-line academic search. We conducted experiments on a real dataset, the ACM Digital Library. The experimental results show that both the Global-MMG and the Net-MMG achieve a better search quality as compared to existing academic search services. Moreover, the Net-MMG has a shorter execution time while still maintaining good quality in its query results. In addition, because a heterogeneous information network has different types of entities, our academic search can support multiple-entity searches. For example, when a user inputs the query term "association rules", our academic search returns four ranked lists, including the most relevant authors, conferences, terms and papers.

To summarize, the main contributions of this work are as follows:

- We explore heterogeneous information networks for academic search services. Relying on information regarding multiple relationships, related papers are ranked and retrieved. By exploring heterogeneous information networks, our academic search service can overcome the drawbacks of keyword-matching techniques.
- Using a heterogeneous graph model, the proposed on-line academic search service can provide a variety of entity types in the query results.
- To enhance the performance of on-line searches, we further propose the Net-MMG, in which topic-based sub-graphs are extracted for determining relevance scores regarding a query.
- We conducted extensive experiments on a real dataset, the ACM Digital Library, to demonstrate the efficiency and effectiveness of our proposed Global-MMG and Net-MMG.

The rest of this paper is organized as follows. Section 2 gives an overview of related work. In Sect. 3, both the Global-MMG and Net-MMG are described. Section 4 is devoted to the experimental results. Section 5 concludes the paper.

2 Related work

Existing academic search services, such as Google Scholar, Microsoft Academic Search and CiteSeer, are based on keyword-based matching approaches for retrieving related papers. Because of the commercial nature of these existing academic search services, the detailed algorithms used by these services have not been published. On the other hand, some academic research papers that mine academic social networks are presented. The authors in [30] developed a bibliographic search system, ArnetMiner, in which the author-conference-topic model (ACT) is proposed to simultaneously model topical aspects of papers, authors and publication proceedings. The ACT model enables many applications, including expertise searches and people association searches. The authors in [28] proposed BibNetMiner, which produces bibliographic information networks. In BibNetMiner, relationships among authors, papers and conferences are modeled as individual graphs. For each graph, the authors' proposed ranking and clustering approaches are used to determine the importance of nodes for ranking. However, the above two academic search services do not explore heterogeneous graph models for retrieving different entities at once when an arbitrary type of query is issued. Note that since a bibliographic dataset has temporal feature, entities could be modeled as social networks with temporal relationships. For temporal relation co-clustering on directional social network, the authors in [24] proposed a three-way alternating non-negative algorithm, TANPT. The work in [24] presents many interesting observations such as the author-topic correlations over years on DBLP datasets. Usually, bibliographic dataset is modeled as graph structures, and these graph structures are usually huge. To achieve good scale-up properties for peta-scale graph mining, the authors in [14] identified and developed a primary graph mining operation, Generalized Iterative Matrix-Vector multiplication (GIM-V), which is commonly used in PageRank, spectral clustering, diameter estimation and connected components. The authors in [14] propose several optimizations to speed up the GIM-V operation in a parallel computing environment.

A significant amount of research effort has been focused on recommendation systems [1,25,4,16,19] and content-based filtering [3]. In recommendations, a list of items referring to products or papers is generated. The ranking of research papers in the query results is

closely related to the recommendation systems. Recommender systems are usually classified into two categories: content-based filtering and collaborative filtering. However, most existing content-based filtering and collaborative filtering recommender systems suffer from a number of drawbacks. For example, content-based techniques are limited to databases with rich textual content, where the textual information can lead to keyword-matching for retrieval.

A substantial amount of research effort has also been dedicated to exploring linkage relationships among data items during ranking [22,15,12]. With linkage relationships, the importance of data items can be determined. In general, a structured database is usually represented as a graph, and a link-based ranking algorithm is then utilized to determine the importance of the nodes. PageRank [22] is a link-based ranking algorithm based on a random walk model used to determine the PageRank score of Web pages. In [23,31], given a graph and one specific node, the Random Walk with Restart model is used to compute the similarities from this specific node to all other nodes. Several studies on the random walk model for similarity searching have been proposed to improve search quality in comparison with the standard collaborative filtering approach [5,17,35]. For example, the authors in [5] proposed QRank and QCRank algorithms for personalized recommendations on k-partite graphs, where QCRank is an approximation version of QRank for the purpose of efficiency. The improved recommendation quality achieved by their random walk model has been verified on the MovieLens and bibliographic datasets. The authors in [17] proposed a collaborative track recommender system by using multiple relationships among users, tracks and tags from music social networks. They compared standard collaborative filtering with the RWR model and showed that this model outperforms the standard collaborative filtering method. The authors in [35] proposed a co-ranking method by coupling two random walks on an authorship network and a document citation network. The authors of [20] introduced a link-based ranking indicator, AuthorRank, to find authoritative authors in a co-authorship network. They showed the advantages of PageRank and AuthorRank over several social network analysis metrics, including degree, closeness and betweenness centrality. Moreover, similarity measurements of nodes are thus determined by their linkage relationships. SimRank [12] is considered as a promising effective metric for measuring the similarity between two nodes. The underlying graph-theoretic model for SimRank is that two objects are similar if they are referenced by similar nodes. However, the main drawback of SimRank is its computational complexity because all of the similarity scores are computed, even if only a portion of them is required. However, most of them do not support multiple-entity searches. To enable multiple-entity searches, the authors in [34] proposed a search framework, SHINE, for retrieving multi-type entities in heterogeneous domains for a query that can be for any type(s) of entities. The key idea of SHINE is that it represents different types of entities in a unified vector space, the extended vector space model (E-VSM), so that similarities among different types of entities can be measured. In addition, the authors in [6] exploited the relationships among tags and resources (i.e., different entities) and proposed a novel clustering method, TagClus. As reported in [6], TagClus with a link-based relevance measure on del.ici.ous can achieve better clustering results for tags.

In this paper, we explore the heterogeneous graph models of given bibliographic datasets. Specifically, the authors in [23] presented an example of a heterogenous graph model, the MMG model, which is designed to capture both identical-type entity and cross-type entity relationships among various types of entities. In bibliographic datasets, an example of a cross-type entity relationship may be the publishing relationship between papers and conferences; an example of an identical-type entity relationship may be the similarity relationship among terms (or authors). Given a set of terms (or authors), a MMG can find identical-type entity relationships of terms via k nearest neighboring (kNN) approaches. Using these similar

terms, a set of relevant papers is determined. Moreover, with a Random Walk with Restart, the importance of relevant papers and other attributes can be determined. To reduce the query response time, we further perform Random Walk with Restart on some relevant sub-graphs instead of on a whole graph. It is worth mentioning that via the heterogeneous graph model, our on-line academic search can provide more search results with different entity types.

3 Algorithms for academic search

In this section, we aim at designing an on-line academic search which explores heterogeneous information networks and a Random Walk with Restart to retrieve and rank the entities required. In Sect. 3.1, given a bibliographic dataset, we propose a Global-MMG. On the other hand, to reduce the execution time for an on-line query, a Net-MMG is proposed in Sect. 3.2.

3.1 Global-MMG

The Global-MMG consists of two phases: *off-line Global-MMG (abbreviated as* G_{MMG}*) construction* and *on-line query in* G_{MMG}. In the off-line G_{MMG} construction, given a whole bibliographic dataset, G_{MMG} is constructed. Then, in the on-line query phase, the issued query navigates G_{MMG} to retrieve related entities and determine their scores via a Random Walk with Restart.

3.1.1 Off-line global-MMG construction

In a bibliographic dataset, each paper is associated with multiple attributes such as *authors*, the *conference* of publication, a set of *referenced papers* and a set of *terms* extracted from the paper title or paper abstracts. G_{MMG} is a weighted graph $G_{MMG} = \langle V, E, W \rangle$, where each distinct paper is viewed as one object and is represented as an object vertex. For each object vertex, the corresponding attributes are represented as attribute vertices. Note that in G_{MMG}, the referenced papers are viewed as attribute vertices. To generate G_{MMG}, two types of links should be added in the bibliographic dataset. One type of link consists of *object-attribute-value links* (abbreviated as OAV-links), and the other type of link consists of *nearest neighbor links* (abbreviated as NN-links). OAV-links are the edges between an object vertex and its attribute vertices. Namely, the edges between a paper object and each of its attribute vertices (i.e., conference, terms or authors) are constructed. NN-links are the edges between an attribute vertex and its k nearest neighbors. In G_{MMG}, NN-links could exist in attribute entities (i.e., the author, conference, terms and referenced papers). In this paper, we use the author attributes and the referenced paper attributes as example attributes for NN-links. The number of NN-links is given by a pre-defined parameter k.

An example of G_{MMG} with three papers and their associated attributes is illustrated in the left rectangle of Fig. 1. For example, the paper object d_1 appears in one conference $\{c_1\}$ and has three terms $\{t_1, t_2, t_3\}$, one author $\{a_1\}$ and two referenced paper vertices $\{r_1, r_2\}$. Suppose that the number of NN-links is 1. Then, each term vertex will link to its nearest neighbor term vertex, and each author vertex will link to its nearest neighbor author vertex.

Note that the edges between a paper object and its corresponding three attribute vertices (term, author and conference) are bi-directional. However, the edges between a paper object and reference paper vertices are unidirectional, namely the edges from a paper object to its reference paper vertices and the edges from a reference paper vertex to the identical paper object. Intuitively, papers are viewed as objects in MMG graph, and three attributes of papers

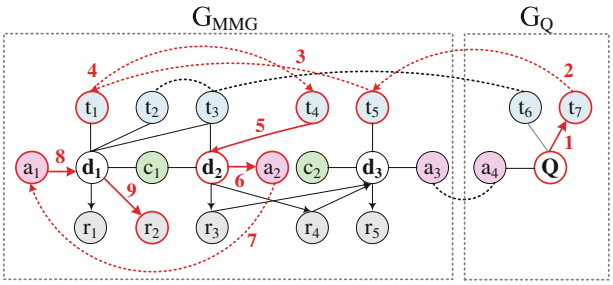


Fig. 1 An example of exploring MMG and RWR for academic search

are used to retrieve paper objects relevant to queries issued by users. Thus, bi-directional edges exist between objects and attributes. On the other hand, the unidirectional edges between a paper object and reference paper vertices are utilized to measure the importance degree of literatures (i.e., paper objects). From the importance degree of literatures, one could further evaluate the importance degree of attribute entities. Thus, in our paper, given a query, our proposed academic search is able to retrieve different attribute entities ranked by their corresponding ranking scores. The ranking scores will be presented later.

The G_{MMG} is built in an off-line manner and is then stored in memory for on-line queries. In summary, the G_{MMG} graph construction phase requires the following: (1) a single scan of distinct paper and attribute objects to construct vertices; (2) a single scan of cross-type entity relations to construct OAV-links and (3) a kNN search for term and author attribute vertices to construct NN-links.

Term List Extraction: The term list and term weight for each paper can be determined as follows: after removing stop words and stemming the terms in the titles and abstracts, the importance of each remaining term is calculated by TF-IDF. Because the title uses words that are effective in capturing the idea of the entire paper, we consider the terms in the title to be more important than the term in the abstract. In practice, each occurrence of a word in the title was counted three times during the TF-IDF calculation. Finally, the top-10 terms with the highest TF-IDF scores were selected as the terms of the research paper.

NN-Link Construction: For both term and author attribute objects, kNN is performed to construct NN-links in the MMG graph. In this paper, we apply two widely used similarity functions, namely *cosine similarity* and *Jaccard similarity*. The cosine similarity function is often used in text mining; it measures the similarity between two vectors by finding the cosine of the angle between them. The Jaccard similarity function measures the overlap of two sets. When the cosine similarity is used, each term (author) attribute is represented as a document vector, in which each entry corresponding to one specific document (i.e., paper). Then, the cosine similarity is used to compare each pair of term (author) attributes. Eventually, for each term (author) attribute, the k nearest neighbors are determined and the NN-links for the k nearest neighbors are constructed. With Jaccard similarity, each term (author) attribute is represented as a set of documents that contain the term (author). Then, the Jaccard similarity measures the overlap of two terms (authors). Eventually, for each term (author) attribute, the k nearest neighbors are determined, and the NN-links for the k nearest neighbors are constructed. Note that we could also explore neighboring relationships for other attribute entities (i.e., referenced papers and conferences). Specifically, the similarity between reference paper entities can be measured in the same way. For conference entities, because two conferences do not share any document in common, instead of representing

each conference entity as a vector of documents, we can represent each conference entity as a vector of terms. Then, we can measure the similarity between conferences by the cosine or Jaccard measurement. For the sake of simplicity, we only show the NN-Links in the author and term attribute entities.

3.1.2 On-line query in G_{MMG}

In the on-line query phase, when a user issues a query that contains a set of entity values (e.g., author, conference, terms), this query is modeled as a star network, denoted as G_Q. Then, NN-links are built across G_Q and G_{MMG} as shown in Fig. 1. Note that NN-links across G_Q and G_{MMG} are constructed via nearest neighbor discovery techniques as in constructing NN-Links for G_{MMG}.

Once links are constructed between G_Q and G_{MMG}, a Random Walk with Restart is used to find the relevant entities. The principle idea of exploiting the Random Walk with Restart is as follows:

Assume that an agent traverses a heterogeneous information network. Both the information of cross-type entity relationships (i.e., authorship, publication, citation and content information) and identical-type entity relationships guide the agent when walking through the heterogeneous networks. Intuitively, given a query term, if a vertex has a higher probability of being visited by the agent, the vertex is then considered more relevant to the query term. The information about the relationship will implicitly guide the agent to visit vertices that are either directly or indirectly connected to the query. For example, Fig. 1 shows an example of a random walk from vertex t_7 in G_Q to another vertex r_2 in G_{MMG} via the sequence $t_7 \rightarrow t_5 \rightarrow t_1 \rightarrow t_4 \rightarrow d_2 \rightarrow a_2 \rightarrow a_1 \rightarrow d_1 \rightarrow r_2$. In this way, the vertex in G_{MMG} is considered more relevant to G_Q if the possibility of reaching it from G_Q is higher.

The procedure of applying a RWR in the constructed MMG graph is given in Algorithm 1. To determine the relevance of each vertex in the G_{MMG} regarding the query G_Q, an adjacency matrix, A_{MMG}, is used to represent the link relationships among vertices. For each column in A_{MMG}, we perform the column-normalized process (line 3). The vector $\vec{v}_q$ is the restart vector with all of its elements set to zero except for the entry that corresponds to the vertex in G_Q. In other words, $\vec{v}_q$ is an N-by-1 vector in which the ith element $\vec{v}_q(i)$ is $\frac{1}{|Q|}$, if vertex i is linked in G_Q. Otherwise, $\vec{v}_q(i) = 0$ (line 4). The vector $\vec{u}_q$ is the ranking vector, which records the ranking score of each vertex at each iteration (line 5). The RWR is performed by iteratively calculating the following equation until the ranking vector $\vec{u}_q$ converges (lines 6–8).

$$\vec{u}_q = (1 - c)A_{MMG}\vec{u}_q + c\vec{v}_q. \tag{1}$$

As a result, the ranking scores of all of the vertices, represented by $\vec{u}_q$, are determined. Finally, the top-L highly ranked vertices that include one object vertex type and other attribute types are generated. For example, when a user issues a query "association rules", our on-line search results include four query results (i.e., query results for papers, authors, conferences and terms), and each query result represents one entity type that is related to "association rules". Note that our academic search returns papers for which the entity type is a referenced paper, because papers with a higher number of citations are generally more important.

Although a RWR can effectively search for relevant entities of various types with respect to a query with arbitrary attribute types, an on-line query in Global-MMG requires navigation over the entire MMG graph. To improve on-line query efficiency, we introduce the Net-MMG, which performs a RWR over selected sub-graphs that are considered highly relevant to the query.

ALGORITHM 1: Discover and determine related entities

Input: ;
G_{MMG}: MMG graph $G_{MMG} = (V, E)$ with N nodes, ;
c: restart probability, ;
Q: a query object with a set of query attributes q,;
L: an integer.;
Output: ;
A ranked list containing top-L relevant entities for each type of attributes.

Construct G_Q for the query object and corresponding links between G_Q and G_{MMG};
Perform the column-normalization on the adjacency matrix A_{MMG} ;
Initialize the restart vector $\vec{v}_q$ with all zeros expect for the entries corresponds to the vertices in G_Q;
Initialize $\vec{u}_q = \vec{v}_q$;
while $\vec{u}_q$ *has not converged* **do**
  Update $\vec{u}_q$ by $\vec{u}_q = (1 - c)A_{MMG}\vec{u}_q + c\vec{v}_q$;
end
return top-L entities with high ranking scores for each attribute type ;

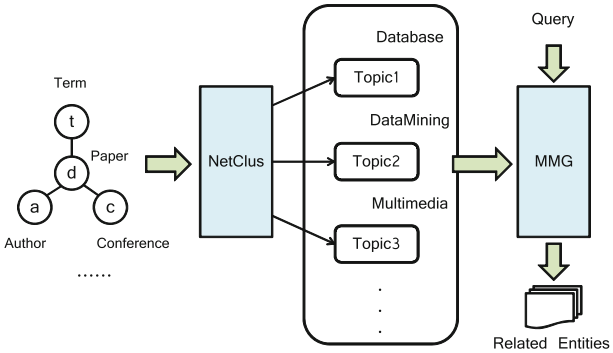


Fig. 2 Overview of Net-MMG

3.2 Net-MMG

In the Global-MMG, a RWR is performed in G_{MMG} which represents the whole bibliographic dataset. Clearly, the performance of the Global-MMG is affected by the size of G_{MMG}. With a greater G_{MMG}, Global-MMG will have a longer response time. Thus, to reduce the query response time, in Net-MMG, we utilize NetClus [29] to extract topic clusters, and each topic cluster includes those papers with similar topics. As such, we only perform a RWR in some clusters which are related to the query issued. Consequently, the query response time is shortened. Figure 2 illustrates the overview of Net-MMG.

Prior to ranking-based clustering, the bibliographic database is transformed to a graph structure that is used in NetClus, which is a heterogeneous information network with a star schema. The required graph structure resembles the MMG graph in that an object vertex (i.e., a paper) in the MMG graph is viewed as a target object in NetClus, and the attribute vertices (i.e., author, term and conference) in the MMG graph are equivalent to the attribute objects in NetClus. Note that the *referenced paper* attribute is not incorporated into the star schema because a referenced paper is usually covered by a broad range of latent topics, which may pose difficulties when identifying relevant papers. We use w_{x_i,x_j} to denote an element in the weight matrix W for the edge $e(x_i, x_j) \in E$, where x_j represents for a target object (i.e., a paper) and x_i is one of the attribute objects (i.e., author, conference or term). Formally, we

define the weight value for x_i and x_j as follows:

$$w_{x_i,x_j} = \begin{cases} 1, & \text{if the target object } x_j \text{ has attribute object } x_i \\ 0, & \text{otherwise.} \end{cases} \tag{2}$$

The NetClus algorithm is a probabilistic generative model that generates not only cluster information but also cluster-based ranking information. NetClus can be considered as an extension of traditional soft clustering, in that the memberships of more than one type of entity can be determined by NetClus. An example cluster generated by NetClus from a bibliographic information network is a cluster of author, conference, term and paper vertices relevant to "databases". To support efficient on-line queries, in the Net-MMG, we only extract some of the sub-graphs determined by NetClus for the Random Walk with Restart. The detailed steps in the Net-MMG are as follows:

- Step 0: Use NetClus to cluster the entire set of target objects in the heterogeneous network G into N clusters $\{C_1, C_2, \ldots, C_N\}$.
- Step 1: Construct an MMG graph G_k for each cluster C_k, where $1 \leq k \leq N$.
- Step 2: In the on-line query phase, when a user submits a query containing a list of query attributes, select the top-M relevant sub-graphs from $\{G_1, G_2, \ldots, G_N\}$ according to the relevance degree between the clusters and the query and exploit the RWR in these top-M sub-graphs.
- Step 3: Calculate the overall reaching probability for each attribute in the M sub-graphs.
- Step 4: Rank all attributes by their overall reaching probabilities in descending order, and select the top-L ranked entities of each type as query results.

First, NetClus is performed to derive N clusters $C_1, C_2, \ldots, C_N$, where each cluster represents a latent topic. By simple ranking in NetClus [29], given a query Q with multiple attribute values, we determine the relevance degrees between query Q and the clusters derived by NetClus. Then, the top-M clusters with the largest relevance degree are selected. For each vertex within these M clusters, we derive its reaching probability with respect to Q. Finally, vertices with a higher reaching probability (i.e., the top-L vertices) are collected as the query results.

Let a query Q with multiple attribute values be $\{q_1, q_2, \ldots, q_n\}$, where the number of attribute values is n. To facilitate our presentation, the attribute type for q_i is expressed by T_{q_i}. Following two independence assumptions with respect to NetClus, given a network G and a query Q, the probabilities of visiting objects from different attribute types are independent. Another independence assumption is that for the same type of object, the probability of visiting two different objects jointly is also independent for the two different objects. Given the above assumptions, we then formulate the relevance degree between query Q and a cluster G_k as follows:

$$p(Q|G_k) = \prod_{q_i \in Q} p(q_i|T_{q_i}, G_k) \tag{3}$$

where $p(q_i|T_{q_i}, G_k)$ is the probability of occurrences of q_i with its attribute type T_{q_i} in G_k. Hence, the value of $p(q_i|T_{q_i}, G_k)$ is derived as follows:

$$p(q_i|T_{q_i}, G_k) = \frac{\sum_{y \in N_{G_k}(q_i)} w_{q_i,y}}{\sum_{q_i' \in T_{q_i}} \sum_{y \in N_{G_k}(q_i')} w_{q_i',y}} \tag{4}$$

where $N_{G_k}(q_i)$ represents the set of target objects (i.e., papers) that have attribute object q_i in G_k, and $w_{q_i,y}$ is the weight value that target object y contains attribute object q_i.

In this paper, if a cluster has a greater relevance degree with respect to query Q, this cluster is selected. Assume that cluster C_k is selected in the top-M clusters. Then, the ranking score for each vertex is determined by the Random Walk with Restart in G_k. Suppose that the ranking score for a vertex x in G_k is denoted as $u_k(x)$. Clearly, given a query Q, the probability of reaching one vertex x in G_k is proportional to the relevance degree between Q and G_k, and the ranking score of x in G_k. As such, the reaching probability of a vertex x in G_k is formulated as follows:

$$p(x|Q, G_k) \propto u_k(x|G_k) \times \prod_{q_i \in Q} p(q_i|T_{q_i}, G_k) \tag{5}$$

Note that it is possible that a vertex x could belong to multiple clusters because NetClus is a soft clustering method. Consequently, the overall reaching probability of vertex x with respect to query Q, denoted as $s(x, Q)$, is derived as follows:

$$s(x, Q) = \frac{\sum_{m=1}^{M} u_m(x|G_m) \cdot \prod_{q_i \in Q} p(q_i|T_{q_i}, G_m)}{\sum_{m=1}^{M} u_m(x|G_m)}. \tag{6}$$

In this paper, the overall reaching probability is viewed as the ranking scores in the Net-MMG.

3.3 Analysis of Global-MMG and Net-MMG

In this section, we analyze the time complexity of Global-MMG and Net-MMG in terms of the time complexity.

3.3.1 Time complexity of Global-MMG

The main off-line computational cost of Global-MMG is graph construction, which consists of the following computation costs. Given a MMG graph, E is the set of edges and V is the set of vertices. First, the time complexity for node creation is $O(\mid V \mid)$. Second, there are two types of links in MMG, NN-links and OVA-links for cross-type entity relations. An kNN search for term and author attribute vertices is used to construct NN-links. Thus, time complexity for NN-link creation is $O(\mid V_t^2 \mid)$ and $O(\mid V_a \mid)$, where V_t (respectively, V_a) is the number of term (respectively, author) entities. For cross-type entity relations, the time complexity for OVA-links is $O(\mid E' \mid)$, where $\mid E' \mid$ is the number of OVA-links in graph G_{MMG}. In all, the time complexity in the off-line Global-MMG construction phase can be derived as $O(\mid V \mid + \mid V_t^2 \mid + \mid V_a \mid + \mid E' \mid)$.

In the on-line phase, the dominant computation cost is the execution of RWR. As such, the time complexity of global ranking on the whole graph is $O(t_1 \mid E \mid)$, where t_1 is the number of iteration and $\mid E \mid$ is the number of edges in graph G_{MMG}.

3.3.2 Time complexity of Net-MMG

Time complexity of Net-MMG is analyzed as follows: First, the graph used in Net-MMG is the same as that in Global-MMG expect that NN-links are excluded. Hence, the time complexity of Net-MMG for graph construction is $O(\mid V \mid + \mid E'' \mid)$, where E'' is the total number of edges in M sub-graphs. Second, as shown in [29], the time complexity of executing algorithm Net-Clus is $O(c_1 \mid E \mid + c_2 N)$ [29], where c_1 and c_2 are constant and N is the number of objects (i.e., the paper objects in this paper).

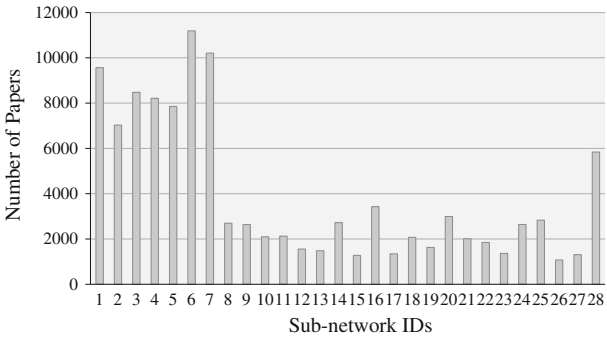


Fig. 3 Number of papers in each sub-network

In the on-line phase, the time complexity of Net-MMG includes the following parts: (1) the time complexity for selecting top-M most relevant sub-graphs is $O(c)$, where c is the number of sub-graphs derived by NetClus. (2) the time complexity for ranking attribute and target nodes by RWR is $O(t_i \mid E_i \mid)$ in each sub-graph G_i, where t_i is the number of iterations. Since we need to perform RWR in all sub-clusters, the time complexity for ranking attributes and target nodes is $O(t \mid E'' \mid)$, where E'' is total number of edges in M sub-graphs and t is the number of iterations. In all, the time complexity for Net-MMG in on-line phase is summarized as $O(t \mid E \mid)$.

Although Net-MMG requires an additional procedure for selecting relevant sub-graphs, the time complexity of the procedure only takes a constant computation time. Since each sub-graph derived by NetClus only involves a small number of edges, Net-MMG adopts the RWR on a set of extracted sub-graphs with smaller graphs, which can significantly improve the efficiency for on-line query.

4 Experiments

In this section, we first introduce a collected dataset and describe our experimental setting. Then, we compare our on-line academic search with existing on-line academic search services.

4.1 Dataset and experiment setting

We conducted experiments on a real dataset, the ACM Digital Library (abbreviated as ACM DL), to evaluate the proposed academic search. The ACM DL dataset contains rich citation information as compared to other academic datasets (e.g., DBLP). From the ACM DL dataset collection, we selected papers from 22 well-known proceedings in the domains of multimedia, user interface design, mobile/sensor networks, data mining, databases and information retrieval. The collected papers were all published before early 2009. In total, the collected ACM DL dataset contains 62,537 papers.

All of the experiments were conducted on an AMD 3.2-GHz computer with 8 GB of main memory, running on Ubuntu Linux 9.10. In our experiments, the sub-networks were recursively derived by NetClus. Eventually, 28 sub-networks were generated from the MMG graph, where each NetClus corresponds to a specific topic. The number of paper objects in each sub-network is plotted in Fig. 3. The average number of paper objects in the 28 sub-networks is approximately 3378 papers. The restart probability c was set c to 0.5. We used

the cosine similarity function to construct NN-links for both term and author attributes. The number of NN-links k was set to 4. The details of the parameter analysis are presented in Sect. 4.2.2.

4.1.1 Competitors

We compared our academic search frameworks, Global-MMG and Net-MMG, with four existing search frameworks: PageRank, Google Scholar, Microsoft Academic Search (abbreviated as MS Academic Search) and ArnetMiner.

- **PageRank** [22]: An academic search engine running PageRank was implemented to investigate the effectiveness of homogeneous information networks. Specifically, the paper lists running PageRank were generated as follows: first, PageRank was used to compute the importance scores of each paper in the paper citation network. Afterward, given a query term, the set of papers containing that term was selected and ranked by their PageRank scores in descending order. The damping factor, which controls how often the random walk jumps to an arbitrary paper, was set to 0.85 in our experiments.
- **Google Scholar** [11]: For each query term, the top-500 papers returned from Google Scholar were collected as the academic search results.
- **Microsoft Academic Search** [9]: Similar to Google Scholar, for each query term, the top-500 papers returned by MS Academic Search were collected as the academic search results. Note that the MS Academic Search engine provides multiple-entity search functionality, where not only documents are returned but also relevant authors and conferences. Hence, the set of authors and conferences is also collected for performance comparison.
- **ArnetMiner** [10]: For each query, we collected the top-500 papers returned by ArnetMiner as the academic search results. Like MS Academic Search, ArnetMiner enables multiple-entity searches. Hence, the set of authors and conferences was also collected for performance comparison.

4.1.2 Performance measurements

We collected two query sets and their corresponding ground truth for performance study. First, we used a well-known data mining textbook [7] to evaluate the effectiveness of the query results. In total, four sample queries were used to evaluate the search quality, including "*Association Rule*", "*Frequent Pattern Mining*, "*Classification*", and "*Clustering*". The set of literature referenced in the corresponding chapters in the book [7] was taken as the ground truth for each query term. Second, ten sample queries were collected from five well-known text books in data mining [7,8], databases [26], web mining [18] and information retrieval [2]. The ten queries include "*Association Rule*", "*Frequent Pattern Mining*", "*Classification*", "*Clustering*", "*Database*", "*Opinion Mining*", "*Link Mining*", "*Sequential Pattern Mining*", "*Information Retrieval*" and "*Data Cube*". Similarly, the set of reference literatures in the corresponding book chapters is collected as the ground truth for each query term. The second dataset is used to derive the average precision for Net-MMG and Global-MMG.

Given a query term, suppose the set of literature in the ground truth is S_{GT} and the top-N papers returned by an academic search engine A is S_A. To evaluate the search quality, we compare the effectiveness of the different search engines in terms of precision and recall as follows:

$$P@N = \frac{|S_{GT} \cap S_A|}{|S_A|} \tag{7}$$

$$R@N = \frac{|S_{GT} \cap S_A|}{|S_{GT}|}. \tag{8}$$

4.2 Experimental results

4.2.1 Performance study of different academic searches

As previously mentioned, one of the advantages of the Global-MMG (or Net-MMG) is that it can consider multiple types of relationships to search for relevant documents. Table 2 lists the relevant papers retrieved by the Global-MMG for the query "Association Rule". Table 2 shows that, our approach can retrieve relevant papers for which the titles do not contain the query term. For example, for the query term "Association Rule", MS Academic Search only returns one paper included in the ground truth. Moreover, for the query term "Frequent Pattern Mining", all four other methods perform poorly. This performance results from the fact that most of the existing methods rely heavily on keyword-matching techniques. Under this constraint, users have to issue accurate terms in their queries to search for relevant literature. For the query terms "Classification" and "Clustering", the Global-MMG performs better as compared to the other methods.

The search quality of the different academic search engines for the top-500 search results is illustrated in Fig. 4. As Fig. 4 shows, the Global-MMG outperforms the academic search engine when running PageRank on citation networks and the other three academic search services. The academic search engine running PageRank has the worst performance in almost all of the cases due to the fact that only citation information is incorporated in searching, whereas the others consider more than one type of relationship within the academic datasets.

4.2.2 Sensitivity study of MMG and RWR in Global-MMG

Similarity Functions for NN-link Construction: We studied two widely used similarity functions, cosine similarity and Jaccard similarity, for NN-link construction. Figure 5 illustrates the search quality obtained by using the different similarity functions. In Fig. 5, the search qualities obtained using the cosine similarity and the Jaccard similarity are almost the same except that the search quality using the cosine similarity is slightly better than that of the Jaccard similarity in the cases of "Association Rule" and "Frequent Pattern Mining". Because the search quality is insensitive to the similarity function, we utilized the cosine similarity for NN-link construction in the remaining experiments.

Number of NN-links: The number of NN-links k clearly determines the scope of the term (author) expansion. Intuitively, if k is too small, then the search results may be limited by the expansion scope (i.e., some relevant papers cannot be found because they cannot be visited during the RWR). Figure 6 illustrates the search quality in terms of precision at rank 10 and 20, respectively, on the ten-query dataset with a varying number of NN-links k for Global-MMG. As can be seen in Fig. 6, the search quality with a larger K is almost similar to that with a smaller K. The reason is that both term and author types do not have a lot of similarity due to a short text information in these two types. In the rest of experiments, the k is set to three for the search efficiency purpose.

Restart Probability: The number of iterations required for the RWR to reach convergence is mainly influenced by the RWR restart probability c. If c is large, then the search scope will be limited to those vertices neighboring the start vertices (i.e., quick convergence). In contrast,

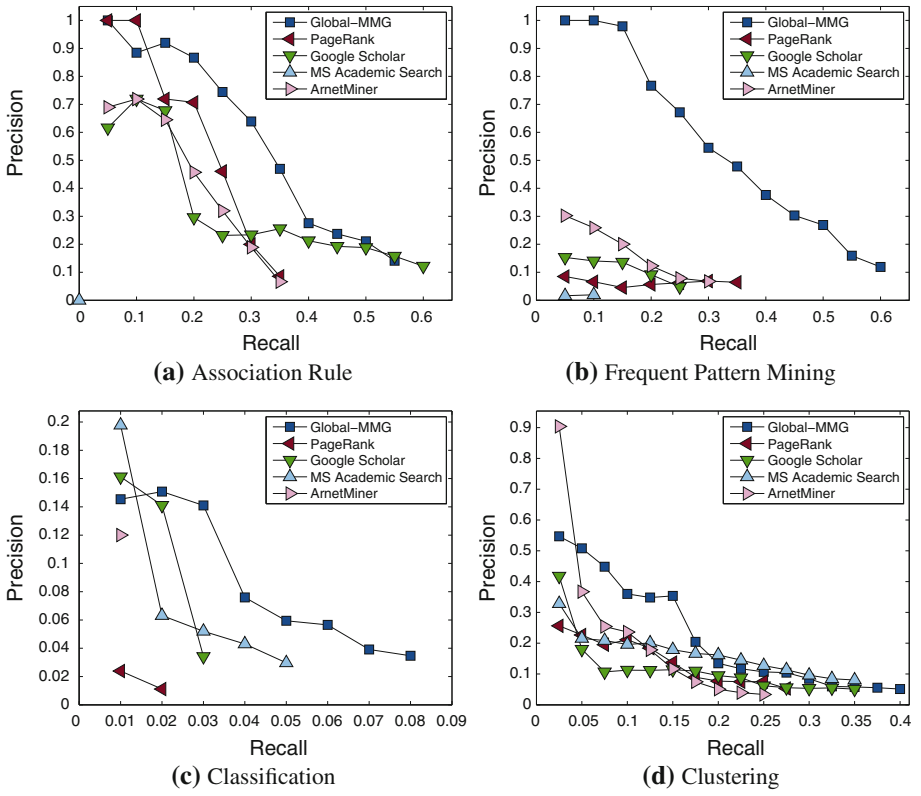


Fig. 4 Precision-recall plot for Global-MMG, PageRank, Google Scholar, MS Academic Search and ArnetMiner. **a** Association Rule, **b** Frequent Pattern Mining, **c** Classification, **d** Clustering

if c is small, then the RWR requires more iterations to reach convergence because new vertices distant from the start vertices are less likely to be visited as compared to cases with a large c in each iteration. Figure 7 shows the number of iterations required for convergence in the Global-MMG. As expected, the number of iterations required for convergence decreases as the restart probability increases.

Figure 8 shows the search quality in the precision-recall curve for varying c. In Fig. 8, we can see that a lower restart probability ($c = 0.1$) leads to a poorer search quality. This result is reasonable because, when c is small, the vertices distant from the start vertices have higher chances of being part of the search results. In that case, the search results may be inaccurate. In contrast, as c increases, the search quality can be improved because the vertices neighboring the start vertices have a higher chance of inclusion in the search results. Note that when $c \geq 0.5$, the search quality stabilizes. Thus, because the search quality is insensitive to c when $c \geq 0.5$, we set c to 0.5 in the remaining experiments.

4.2.3 Performance study of Global-MMG and Net-MMG

As previously mentioned, there are two advantages of Net-MMG over Global-MMG. First, Net-MMG requires less response time, and second, Net-MMG produces clustering results that can contain topic information.

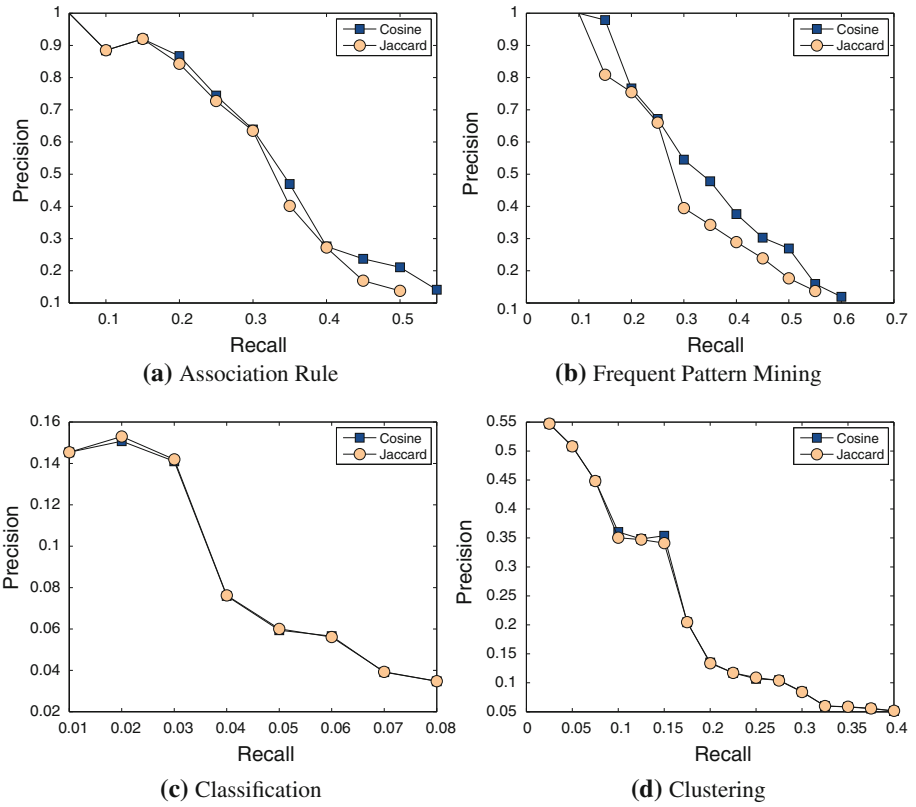


Fig. 5 Precision-recall plot for Global-MMG with different similarity functions. **a** Association Rule, **b** Frequent Pattern Mining, **c** Classification, **d** Clustering

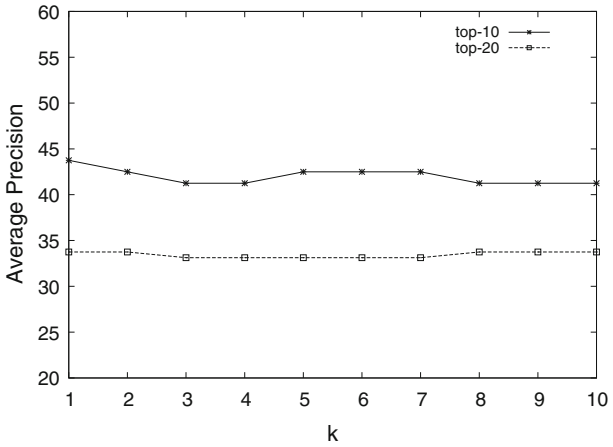


Fig. 6 Average precision for Global-MMG with the value k varied

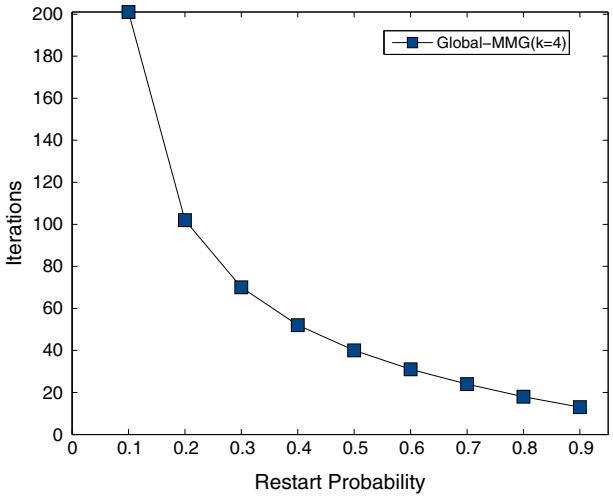


Fig. 7 Number of iteration required for RWR converge in Global-MMG with $k = 4$

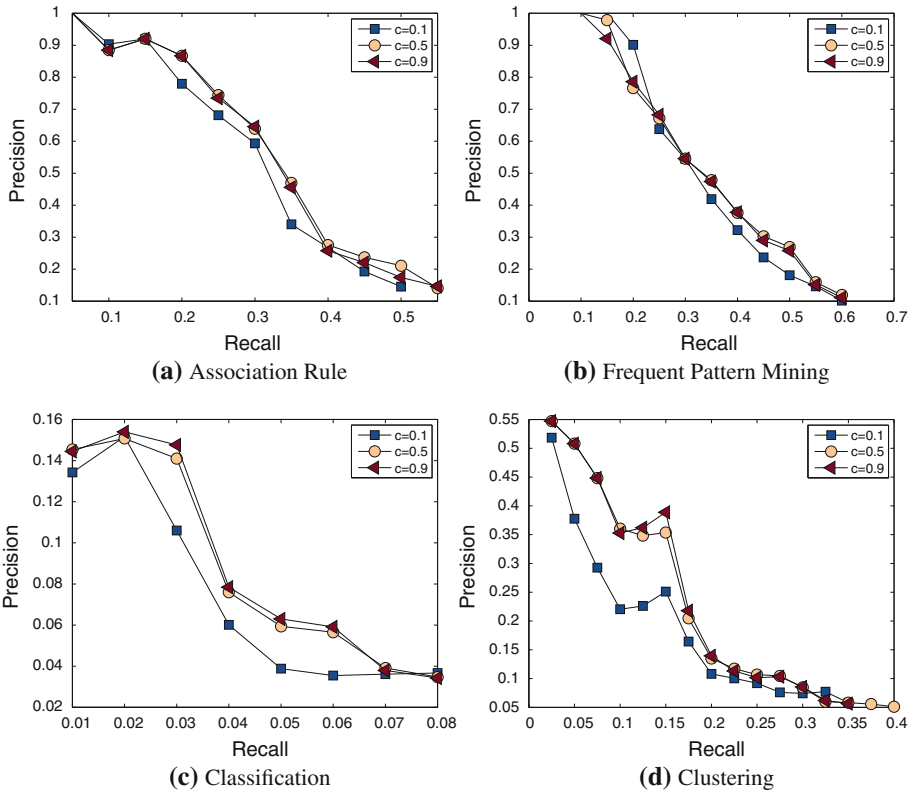


Fig. 8 Precision-recall plot for Global-MMG with varying restart probabilities. **a** Association Rule, **b** Frequent Pattern Mining, **c** Classification, **d** Clustering

Table 3 On-line query time (sec.) for Net-MMG with $M = 1, 3, 5, 7$ and Global-MMG

Method	Association Rule	Frequent Pattern Mining	Classification	Clustering
Net-MMG ($M = 1$)	**0.0188**	**0.0169**	**0.0137**	**0.0161**
Net-MMG ($M = 3$)	0.0448	0.0403	0.0397	0.039902
Net-MMG ($M = 5$)	0.0570	0.0550	0.0590	0.0544
Net-MMG ($M = 7$)	0.0696	0.0899	0.0697	0.0795
Global-MMG	**0.6039**	**0.5907**	**0.5912**	**0.5904**

The values in bold type emphasizes the potential difference in runtime that can be reduced by Net-MMG ($M = 1$) from Global-MMG

Table 4 P@10 and P@50 for Net-MMG with $M = 1, 3, 5, 7$ and Global-MMG, Google Scholar, ArnetMiner and MS Academic Search

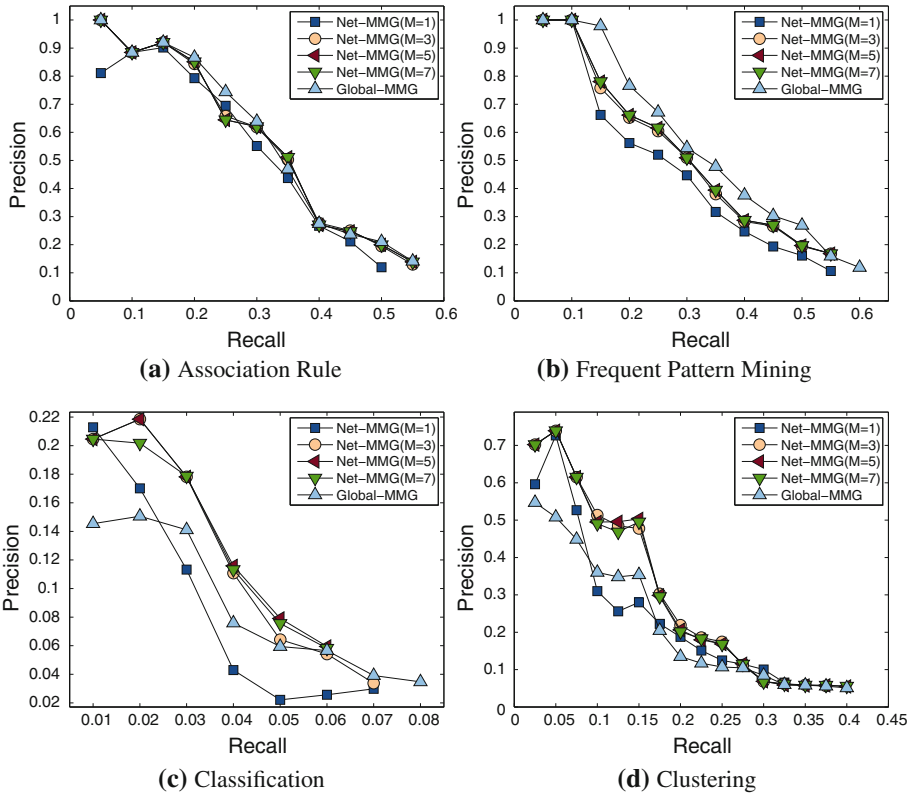
Method	Association Rule		Frequent Pattern Mining		Classification		Clustering	
	p@10	p@50	p@10	p@50	p@10	p@50	p@10	p@50
Net-MMG ($M = 1$)	0.8889	0.5102	1.0000	0.4489	0.2222	0.1224	0.4444	0.2245
Net-MMG ($M = 3$)	0.8889	0.5306	1.0000	0.4898	0.2222	0.1429	0.5556	0.2245
Net-MMG ($M = 5$)	0.8889	0.5306	1.0000	0.4898	0.2222	0.1429	0.5556	0.2245
Net-MMG ($M = 7$)	0.8889	0.5306	1.0000	0.4898	0.2222	0.1429	0.5556	0.2245
Global-MMG	0.8889	0.5306	1.0000	0.5102	0.2222	0.1224	0.5556	0.2245
Google Scholar	0.6667	0.3061	0.0000	0.1224	0.1111	0.1020	0.2222	0.1020
ArnetMiner	0.6667	0.3469	0.4444	0.2245	0.1111	0.0612	0.3333	0.1633
MS Academic Search	0.1111	0.0204	0.0000	0.0000	0.2222	0.0612	0.2222	0.1837

As shown in Table 3, the on-line query time of Net-MMG for $M = 1$ is approximately 30-40 times faster than that of Global-MMG. Because only relevant sub-networks are required in the search, the number of relevant vertices visited is significantly reduced in on-line queries. Overall, the search quality of Net-MMG is approximately equal to that of Global-MMG, while the on-line query response time of Net-MMG is much shorter.

Table 4 shows an overall comparison in terms of precision at rank 10 and 50 for the Net-MMG with different M values, for the Global-MMG and for the other methods. Table 4 shows that both Global-MMG and Net-MMG return better search results as compared to the other three academic search services. Note that the MMG used by Net-MMG is pruned by Net-Clus so that the reduced MMG can definitely improve the search efficiency. Nevertheless, the literature search results of the Net-MMG are still very close to those of Global-MMG. As M increases, the explored graph is more complete and thus could have better search results. For some cases such as "Classification", the Net-MMG even achieves better search quality as compared to the Global-MMG with when $M \geq 3$. Take a closer examination at the NetMMG and Global-MMG, we conduct experiments on the ten-query set. Table 5 shows an overall comparison in terms of precision at rank 10 and rank 20 for the Net-MMG with different M values and Global-MMG. As can be seen in Table 5, the search quality is slightly improved as M increases in terms of the precision. The reason is that Net-MMG always starts exploring relevant entities with the most relevant sub-graphs by given the number M. As M greatly increases, less relevant sub-graphs are included during the exploration process, which makes the search results less relevant. On the other hand, the design of selecting

Table 5 Average P@10 for Net-MMG with $M = 1,3,5,7$, Global-MMG and Google Scholar

	Google	Global-MMG	Net-MMG ($M = 1$) (%)	Net-MMG ($M = 3$) (%)	Net-MMG ($M = 5$) (%)	Net-MMG ($M = 7$) (%)
P@10	20.6	53.7	51.49	53.69	52.69	51.69
P@20	12.65	35.85	31.06	32.88	32.38	33.88



(a) Association Rule **(b)** Frequent Pattern Mining

(c) Classification **(d)** Clustering

Fig. 9 Precision-recall plot for Global-MMG and Net-MMG with varying M. **a** Association Rule, **b** Frequent Pattern Mining, **c** Classification, **d** Clustering

M sub-graphs for entity exploration is mainly for efficiency. With a smaller M, Net-MMG can spend much less time for on-line exploration while still guaranteeing the similar search quality of Global-MMG. The search quality of Google Scholar is also included to show a significant improvement in search quality of Global-MMG and Net-MMG.

An interpolated precision-recall curve for the top-500 ranked papers for the Net-MMG with different topic numbers M is illustrated in Fig. 9. Figure 9 shows that for some cases, both the Net-MMG and Global-MMG achieve the best search quality. For example, the Global-MMG presents slightly better literature search quality as compared to the Net-MMG when the query terms are domain-specific, such as "Association Rule" and "Frequent Pattern Mining", as shown in Fig. 9a, b. On the other hand, the Net-MMG has better search quality when the query terms are much more general and broadly used in different domains, such

Table 6 Top-5 relevant authors for the query "Association Rule"

Rank	Global-MMG	Net-MMG ($M = 3$)	ArnetMiner	MS Academic Search
1	Jiawei Han	Jiawei Han	Rakesh Agrawal	Mohammed Javeed Zaki
2	Rakesh Agrawal	Rakesh Agrawal	Jiawei Han	Masaru Kitsuregawa
3	Philip S. Yu	Philip S. Yu	Ramakrishnan Srikant	Osmar R. Zaiane
4	Bing Liu	Bing Liu	Philip S. Yu	Frans Coenen
5	Masaru Kitsuregawa	Ke Wang	David Wai-Lok Cheung	Sharma Chakravarthy

Table 7 Top-5 terms and authors relevant to "CIKM" and "MDM"

CIKM			MDM		
Rank	Author	Term	Rank	Author	Term
1	Philip S. Yu	Query	1	Shojiro Nishio	Network
2	James Allan	Information	2	Takahiro Hara	Mobile
3	W. Bruce Croft	Database	3	Katsumi Tanaka	Data
4	Ophir Frieder	Retrieval	4	Teruo Higashino	Sensor
5	Elke A. Rundensteiner	Data	5	Baihua Zheng	Service

as "Classification" or "Clustering" as shown in Fig. 9c, d. The main reason that the Global-MMG is not as effective as the Net-MMG in cases of general query terms is because the Net-MMG performs more accurate searches on focused and topic-relevant sub-networks with the help of topic classification by NetClus.

4.2.4 Multiple-type academic searches in Global-MMG and Net-MMG

As stated above, another advantage of the Global-MMG (or Net-MMG) is that it enables multi-type academic searches. We demonstrate the ability of a multi-type academic search from two perspectives, multiple-type search results and multiple-type queries. Table 6 lists the top-5 authors relevant to the query "Association Rule" for the Global-MMG, Net-MMG with $M = 3$, ArnetMiner and MS Academic Search. Additionally, our approach can search by issuing queries other than key terms. For example, we can use authors or conferences as a query and find all relevant entities for various types. Table 7 shows an example of a multiple-type search in which the top-5 terms and top-5 authors relevant to CIKM and MDM are listed. Table 8 shows the top-5 terms and top-5 conferences related to two authors, Philip S. Yu and Rakesh Agrawal. As expected, the term "Stream" is considered relevant to the researcher Philip S. Yu because we can find many papers related to stream mining from Philip S. Yu's publications. All of the above examples suggest that our approach provides more utility as compared to traditional academic search services (e.g., Google scholar). The multiple-type property introduced by our approach enables several applications in bibliographic databases such as query recommendation and expert finding.

To show the search capability with input as a set of different types of entities, we issue queries that have author and keyword object types. Table 9 illustrates the top-5 papers by issuing the query "Rakesh Agrawal" and "mining". Among the top-5 papers returned by Global-MMG, the top-3 papers are indeed written by Rakesh Agrawal. As for the papers ranked in 4 and 5, these papers are mining papers that improve the efficiency of the algorithms

Table 8 Top-5 terms and conferences related to two authors found by Global-MMG

Philip S. Yu			Rakesh Agrawal		
Rank	Term	Conf.	Rank	Term	Conf.
1	Data	KDD	1	Data	VLDB
2	Algorithm	CIKM	2	Mining	SIGMOD
3	Stream	SIGMOD	3	Database	WWW
4	Indexing	WWW	4	Query	ICDE
5	Query	VLDB	5	Relational	KDD

Table 9 Top-5 relevant papers by Global-MMG for the query "Rakesh Agrawal" and "mining"

Rank	Paper title
1	Fast algorithms for mining association rules in large databases
2	Mining association rules between sets of items in large databases
3	Mining sequential patterns
4	Mining frequent patterns without candidate generation
5	An efficient algorithm for mining association rules in large databases

proposed by Rakesh Agrawal. These algorithms are appeared in the top-3 papers in Table 9. Clearly, our academic search provides diverse search functionality with the capability of querying by different types of entities.

5 Conclusions

In this paper, we have developed an on-line academic search that explores MMGs to model a bibliographic dataset. Using an MMG, multiple types of entities relevant to a query can be determined by performing a RWR on the MMG. Specifically, we first modeled a bibliographic dataset as a heterogeneous graph model, where each vertex represents one object for each of the entity types (i.e., author, conference and paper) and the edges fully reflect cross-type entities and neighboring relationships from bibliographic databases. By adopting the Random Walk with Restart model, more relevant papers are retrieved even when the retrieved papers may not have the exact same issued keyword. To emphasize query efficiency, we further proposed the Net-MMG, in which a topic-based sub-graph is extracted for the RWR. We conducted experiments using the ACM Digital Library dataset. The experimental results show the effectiveness of the Global-MMG and Net-MMG as compared to existing academic search services. In addition, the Net-MMG has a shorter execution time but still maintains good search quality.

References

1. Adomavicius G, Tuzhilin A (2005) Toward the next generation of recommender systems: a survey of the state-of-the-art and possible extensions. IEEE Trans Knowl Data Eng 17(6):734–749
2. Baeza-Yates R, Ribeiro-Neto B et al (1999) Modern information retrieval. ACM press, New York

3. Bharat K, Kamba T, Albers M (1998) Personalized, interactive news on the web. Multimed Syst 6(5): 349–358
4. Breese JS, Heckerman D, Kadie C et al. (1998) Empirical analysis of predictive algorithms for collaborative filtering. In: Proceedings of uncertainty in artificial intelligence, pp 43–52
5. Cheng H, Tan PN, Sticklen J, Punch WF (2007) Recommendation via query centered random walk on K-partite graph. In: Proceedings of IEEE computer society international conference on data mining, pp 457–462
6. Cui J, Liu H, He J, Li P, Du X, Wang P (2011) Tagclus: a random walk-based method for tag clustering. Knowl Inform Syst 27(2):193–225
7. Han J, Kamber M (2006) Data mining: concepts and techniques. Morgan Kaufmann, Los Altos
8. Han J, Kamber M, Pei J (2011) Data mining: concepts and techniques. Morgan Kaufmann, Los Altos
9. http://academic.research.microsoft.com
10. http://arnetminer.org
11. http://scholar.google.com
12. Jeh G, Widom J (2002) Simrank: a measure of structural-context similarity. In: Proceedings of SIGKDD. ACM, New York, NY, pp 538–543
13. Jiawei H, Jian P, Yiwen Y (2000) Mining frequent patterns without candidate generation. In: Proceedings of SIGMOD, pp 1–12
14. Kang U, Tsourakakis CE, Faloutsos C (2011) Pegasus: mining peta-scale graphs. Knowl Inform Syst 27(2):303–325
15. Kleinberg JM (1999) Authoritative sources in a hyperlinked environment. J ACM (JACM) 46(5):604–632
16. Konstan JA, Miller BN, Maltz D, Herlocker JL, Gordon LR, Riedl J (1997) GroupLens: applying collaborative filtering to Usenet news. Commun ACM 40(3):87
17. Konstas I, Stathopoulos V, Jose Joemon M (2009) On social networks and collaborative recommendation. In: Procedings of SIGIR, pp 195–202
18. Liu B (2007) Web data mining: exploring hyperlinks, contents, and usage data. Springer, Berlin
19. Liu NN, Yang Q (2008) Eigenrank: a ranking-oriented approach to collaborative filtering. In: Proceedings of SIGIR. ACM, New York, pp 83–90
20. Liu X, Bollen J, Nelson ML, Van de Sompel H (2005) Co-authorship networks in the digital library research community. Inform Process Manag 41(6):1462–1480
21. Long B, Wu X, Zhang ZM, Yu PS (2006) Unsupervised learning on k-partite graphs. In: Proceedings of SIGKDD. ACM, New York, p 326
22. Page L, Brin S, Motwani R, Winograd T (1998) Bringing order to the web. The pagerank citation ranking.
23. Pan J-Y, Yang H-J, Faloutsos C, Duygulu P (2004) Automatic multimedia cross-modal correlation discovery. In: Proceedings of SIGKDD, pp 653–658
24. Peng W, Li T (2011) Temporal relation co-clustering on directional social network and author-topic evolution. Knowl Inform Syst 26(3):467–486
25. Sarwar B, Karypis G, Konstan J, Reidl J (2001) Item-based collaborative filtering recommendation algorithms. In: Proceedings of WWW. ACM, New York, p 295
26. Silberschatz A, Korth HF, Sudarshan S (2002) Database system concepts. McGraw-Hill, New York
27. Sun Y, Han J, Zhao P, Yin Z, Cheng H, Wu T (2009) Rankclus: integrating clustering with ranking for heterogeneous information network analysis. In: Proceedings of the 12th EDBT. ACM, New York, pp 565–576
28. Sun Y, Wu T, Yin Z, Cheng H, Han J, Yin X, Zhao P (2008) BibNetMiner: mining bibliographic information networks. In: Proceedings of SIGMOD. ACM, New York, pp 1341–1344
29. Sun Y, Yu Y, Han J (2009) Ranking-based clustering of heterogeneous information networks with star network schema. In: Proceedings of SIGKDD. ACM, New York, pp 797–806
30. Tang J, Zhang J, Yao L, Li J, Zhang L, Su Z (2008) Arnetminer: extraction and mining of academic social networks. In: Proceedings of SIGKDD. ACM, New York, pp 990–998
31. Tong H, Faloutsos C, Pan JY (2006) Fast random walk with restart and its applications. In: Proceedings of ICDM, pp 613–622
32. Tong H, Papadimitriou S, Yu PS, Faloutsos C (2008) Proximity tracking on time-evolving bipartite graphs. In Proceedings of SIAM. Citeseer, pp 704–715
33. Wang JL (2008) Academic literature search based on collaborative recommendation by authors. Master's thesis, National Chengchi University
34. Wang X, Sun J-T, Chen Z (2007) Shine: search heterogeneous interrelated entities. In: Proceedings of CIKM, pp 583–592
35. Zhou D, Orshanskiy SA, Zha H, Lee GC (2007) Co-ranking authors and documents in a heterogeneous network. In Proceedings of ICDM. IEEE Computer Society, pp 739–744

Author Biographies

Meng-Fen Chiang received a BS degree and an MS degree from Chengchi University, Taipei, Taiwan, in 2004 and 2006, respectively. She is currently a Ph.D. student at the Department of Computer Science, Chiao Tung University, Hsinchu, Taiwan. Her research focus is on data mining, especially, querying of large-scale, semi-structured and heterogeneous network data such as Web browsing streams and social networks.

Jiun-Jiue Liou was born in TaiChung, Taiwan, R.O.C in 1985. He received the BS and MS degrees from the National Chiao Tung University, Taiwan, in 2008 and 2010, respectively. He is interested in data mining and programming completion. Currently, he is a software engineer in TrendMicro.

Jen-Liang Wang was born in Taipei, Taiwan, R.O.C in 1975. He received the MS degrees from the National Chengchi University, Taiwan, in 2008. He is interested in data mining, system analysis and system design. Currently, he is a project manager in SYSTEX Corporation.

Wen-Chih Peng was born in Hsinchu, Taiwan, R.O.C in 1973. He received the BS and MS degrees from the National Chiao Tung University, Taiwan, in 1995 and 1997, respectively, and the Ph.D. degree in Electrical Engineering from the National Taiwan University, Taiwan, R.O.C in 2001. Currently, he is an associate professor at the department of Computer Science, National Chiao Tung University, Taiwan. Prior to joining the department of Computer Science and Information Engineering, National Chiao Tung University, he was mainly involved in the projects related to mobile computing, data broadcasting and network data management. Dr. Peng serves as PC members in several prestigious conferences, such as IEEE International Conference on Data Engineering (ICDE), IEEE International Conference on Data Mining (ICDM) and ACM International Conference on Information and Knowledge Management (ACM CIKM). His research interests include mobile data management, data mining and sensor networks. He is a member of IEEE.

Man-Kwan Shan is a professor of both Department of Computer Science and Program in Digital Content and Technologies at National Chengchi University. He received the Ph.D. degree in Computer Science and Information Engineering from National Chiao Tung University in 1998. His current research interests include social network mining, cloud databases, multimedia systems, and computer music.