# BibNetMiner: Mining Bibliographic Information Networks*

Yizhou Sun† Tianyi Wu† Zhijun Yin† Hong Cheng† Jiawei Han† Xiaoxin Yin‡
Peixiang Zhao†
†University of Illinois, Urbana-Champaign ‡Microsoft Research
†{sun22,tw5,zyin3,hcheng3}@uiuc.edu, hanj@cs.uiuc.edu, pzhao4@uiuc.edu
‡xiaoxin@gmail.com

## ABSTRACT

Online bibliographic databases, such as DBLP in computer science and PubMed in medical sciences, contain abundant information about research publications in different fields. Each such database forms a gigantic information network (hence called *BibNet*), connecting in complex ways research papers, authors, conferences/journals, and possibly citation information as well, and provides a fertile land for information network analysis. Our BibNetMiner is designed for sophisticated information network mining on such bibliographic databases. In this demo, we will take the DBLP database as an example, demonstrate several attractive functions of BibNetMiner, including clustering, ranking and profiling of conferences and authors based on the research subfields. A user-friendly, visualization-enhanced interface will be provided to facilitate interactive exploration of a bibliographic database. This project will serve as an example to demonstrate the power of links in information network mining. Since the dataset is large and the network is heterogeneous, such a study will benefit the research on the analysis of massive heterogeneous information networks.

## Categories and Subject Descriptors

H.4.0 [**Information Systems**]: INFORMATION SYSTEMS APPLICATIONS—*General*

## General Terms

Algorithms, Management

## Keywords

Bibliographic Information Networks, Link Analysis, Clustering, Ranking

## 1. INTRODUCTION

There are numerous online bibliographic databases, such as DBLP (dblp.uni-trier.de) in computer science and PubMed (www.ncbi.nlm.nih.gov/entrez) in medical sciences. Such databases contain abundant information on papers, authors, conferences or journals, and possibly about citations as well. Currently, most bibliography search engines provide only keyword-based search or similarity search functions for retrieving information directly stored in the databases. However, each such database forms a gigantic information network (called *BibNet*), but rich information hidden in such bibliographic information networks are largely unexplored.

The exploration of massive link information in a bibliographic database may disclose much valuable, in-depth information about research, such as the clustering of conferences due to their sharing of many common authors, the reputation of a conference due to its serving as a common forum for many productive authors, research evolving with time, and the profile of a conference, an author, or a research area. This motivates us to study the information network mining on bibliographic databases and develop the BibNetMiner system, with the following distinct features:

1. It develops a ranking-based clustering mechanism that mutually enhances clustering and ranking by iterative link propagation and analysis, *e.g.*, conferences and authors can be clustered and ranked using this technique;
2. It performs clustering-based evolution analysis and discloses evolution regularities or irregularities for authors, conferences, and research themes;
3. It provides multidimensional profiling function, workout profiles for authors, conferences, research areas, and research evolutions; and
4. It also provides a user-friendly interface and a multi-resolution visualization tool for users to browse and comprehend the information derived from the above analyses.

## 2. GENERAL SYSTEM ARCHITECTURE

The BibNetMiner system has a three-layer architecture, as shown in Figure 1. The bottom layer contains the information extraction and analysis engine which performs link-based clustering and ranking analysis based on our recent and on-going research. The middle layer is the functional module layer, which implements the major function modules based on the clustering and ranking information derived from the information network analysis. The top layer contains a user-friendly and visualization-enhanced interface, which interacts with users and responds to their requests.

## 3. THE CORE LINK-ANALYSIS ENGINE

As information network data becomes ubiquitous, extracting knowledge from information networks has become an

*The work was supported in part by the U.S. National Science Foundation NSF IIS-05-13678 and NSF BDI-05-15813. Any opinions, findings, and conclusions or recommendations expressed here are those of the authors and do not necessarily reflect the views of the funding agencies.


*SIGMOD'08,* June 9–12, 2008, Vancouver, BC, Canada.
ACM 978-1-60558-102-6/08/06.

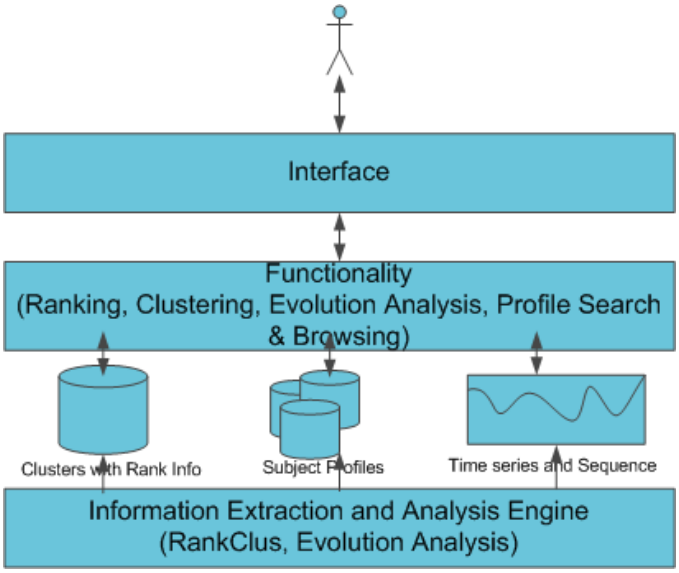


**Figure 1: System Architecture of** BibNetMiner

important task. Link analysis has been playing an essential role in the mining of massive information networks, which has also been shown in our recent studies, such as link-based clustering [4], object distinction analysis [5], and veracity analysis [6]. Both ranking and clustering can provide general views on information network data, and each of which has been a hot topic by itself. However, ranking objects globally without considering which clusters they belong to often leads to dumb results, *e.g.,* ranking database and computer architecture conferences together may not make much sense. Similarly, clustering a huge number of objects (*e.g.,* thousands of authors) in one huge cluster without distinction is very dull as well. In this demo, we propose a novel clustering framework called RankClus to integrate clustering with ranking, which applies conditional ranking relative to clusters to improve ranking quality, and uses accumulative ranking scores as the features to improve clusters. As a result, quality of clustering and ranking are mutually enhanced. Moreover, the clustering results with ranking can provide more informative views of data. For mining bibliographic databases, we have the following heuristics for effective ranking and clustering discovery.

1. A conference/journal is *reputed* if it attracts many papers from a good number of *prolific* or *highly-regarded authors*;
2. An author is *highly-regarded* if s/he publishes many papers in *reputed* conferences/journals;
3. Authors often publish in the *same* conferences/journals if they share the *same or similar research interests*;
4. A reputed conference/journal belongs to *one research field* if it collects papers of highly regarded authors mainly in that field; and
5. A group of conference/journal belongs to *the same field* if they mainly publish papers in that field/theme.

From these rules, one can see that clustering and ranking of authors and conferences/journals are intertwined, and it is difficult to perform separated, high-quality clustering or ranking. Therefore, for BibNet analysis, it is desirable to integrate the clustering and ranking processes together and consider them as one process with mutually enhanced measures. Moreover, except the information about prolific authors which can be extracted directly from a bibliographic database, reputed venues, highly regard authors, and other clustering and ranking information must be obtained by an iterative, progressive, and mutual enhancement process. This is similar to the PageRank and HITS algorithms popularly used in Internet search engines. Therefore, we propose a RankClus algorithm, with weights iteratively propagated and progressively enhanced among authors, conferences/journals, and research fields. Such a process terminates when the clusters and ranks do not change significantly.

In ranking part, to make ranking more meaningful, we propose *conditional ranking* and *within-cluster ranking*. Conditional ranking is the ranking of objects in an information sub-network determined by certain clusters of a different typed objects in the original network. For example, authors' conditional ranking is relative to each conference cluster. Within-cluster ranking is the ranking of objects in an information sub-network determined by a certain cluster of the same typed objects in the original network. For example, we can define conference's within-cluster ranking in the Database and Data Mining area or in the Hardware and Architecture area. Also, as ranking semantic may vary according to different users, we offer different ranking functions, which can integrate with users' rules as well. In clustering part, we fully use the ranking scores to define the distance measure between objects and clusters, which turns in an effective and efficient measure.

An initial implementation and testing of our method has demonstrated the high promise of this proposed approach. For example, Figure 2 illustrates a set of the conference clusters in DBLP network, in which, nodes represent conferences with high rank, and if similarity between two conferences is above a threshold, there is edge between them. Different colors stand for different clusters generated by our algorithm. Our experiment results also show that RankClus can gen-

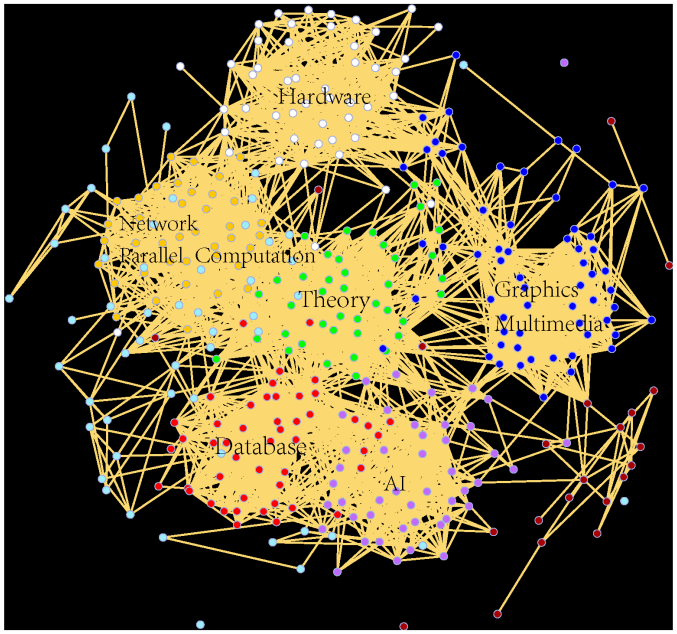


**Figure 2: Clusters of DBLP Conferences**

erate more accurate clusters than the state-of-art link-based clustering methods in a more efficient way.

# 4. MAJOR FUNCTIONAL MODULES

Ranking and clustering are two main functional modules in our system. Besides clustering and ranking, the system will provide two additional modules: evolution analysis and profiling. In this section, we introduce these functional modules in detail.

## 4.1 Ranking

Ranking aims at giving conferences and authors with higher authority higher rank. It's very helpful for users to quickly navigate to important objects. In order to give more accurate ranking results, we offer three mechanisms for users to generate and adjust ranking results.

First, considering ranking conferences and authors are only meaningful in a specific research area, we ask users to choose the area and conferences they are interested in, and give conditional ranking for authors and within-cluster ranking for conferences, relative to the conferences that users have chosen.

Second, ranking may have different semantics. Accordingly, we give different ranking functions for users to choose. For example, we can rank authors according to their productivity, *i.e.*, number of papers they published. Also, we can rank authors according to their authority.

Third, users may have their own rules when defining which conference or which author should rank higher. We allow users to adjust the rank order through the user interface, and our ranking algorithm will learn parameters through the process and give more accurate ranking according to users preferences by users' interaction.

## 4.2 Clustering

Group conferences or authors together will give a overall view for users. Unlike traditional clustering which only gives the cluster label to each object, our RankClus algorithm will also rank objects in each cluster as well. Obviously, clustering with ranking will give users a clearer view for each cluster.

In order to get the clusters, users need to specify the ranking function and the number of clusters, $k$. For conferences, each conference will be labeled one of the cluster label, and the order of their presenting is their rank in the cluster. Associating with each conference cluster, author clusters will be presented as well, and authors will also be ordered according to their ranks.

## 4.3 Evolution Analysis

The *evolution analysis* in BibNetMiner consists of two components: *trend analysis* and *structural evolution analysis*. The former refers to the change of some measures over time whereas the latter refers to the change of the network topological structures. There are interesting applications that require to accomplish the evolution analysis on the bibliographic data sets, *e.g.*, (1) how the research themes evolve in recent years; (2) how the productivity or visibility of a researcher evolves over time; (3) how a researcher's collaboration network evolves; and (4) how a research group evolves with members joining and departing. Among these tasks, the first two belong to the trend analysis and the remaining two belong to the structural evolution analysis. Such results will provide a temporal overview of historical events, and facilitate trend prediction as well as outlier detection.

To support the evolution analysis, we form a history-associated temporal bibliographic network with the time information, so that objects or events and their relationships are associated with the corresponding history. Different from a recent study in [1] which proposes a general statistics-based view on single graph/network evolution, we present an alternative view as *multiple graphs formed at different time durations*. For example, the author-paper graphs in DBLP can be viewed as multiple graphs separated by publication years so that comparison of such graphs across time can be performed to find evolution regularities.

For trend analysis, aggregated values with respect to some predefined measures can be extracted along with time. Sequential pattern mining and time series analysis methods will then be applied. For structural evolution analysis, structural pattern mining could be applied for comparison and contrast analysis on the networked data over time.

## 4.4 Profiling for Effective Browsing

A typical bibliography search engine provides the table of contents of conference proceedings or journals, or retrieves lists of publication for specific authors, besides returning concrete research papers based on search primitives. Profiling provides an alternative to summarize information buried in the bibliographic data by reorganizing data at a higher level of abstraction or from a different angle, which helps users gain insight into the data being examined. In BibNetMiner, we present profiling functions on three subjects: *author, conference/journal*, and *research area*.

### *4.4.1 Author Profiling*

By author profiling, bibliographic data concerning a specific author are re-structured in a more informative way. First, the author's research interests can be outlined and

the trend or evolution can be extracted as well. Second, the publications of the author can be organized and viewed from multiple perspectives, *e.g.*, by year, by area, or by conference/jounal, which may facilitate the access of literature published by the author. Third, information about co-authorship can be extended: instead of simply providing "*who cooperates whom*", questions like "*who shares similar research interests with me" or "who can form a firm research community with me*" can be effective answered. In the mean time, one may like to know "*how important*" and "*how productive*" an author is in a special field. This can be derived from the ranking in a particular field/topic.

### 4.4.2 Conference/Journal Profiling

Instead of presenting table of contents of a conference or journal in a year, the data can be organized in a multi-dimensional, multi-level way. Publications can be viewed by areas, sub-areas, authors, time, and evolution trends. Moreover, research papers in a conference or journal can also be grouped and ranked based on the author reputation to facilitate top-$k$ queries. Finally, conferences and journals can be clustered and ranked based on the same research themes.

### 4.4.3 Research Area Profiling

As discussed above, research areas can be identified based on the clustering of conferences and authors sharing common interests. By integration of the keyword information in the paper titles with some existing subject hierarchy or ontology information, these areas can be further partitioned into sub-areas, *etc.*. By profiling the research areas, one can distinguish "pure" areas from interdisciplinary ones, or conference dedicated to one area or interdisciplinary ones. Such information is important at predicting research trends. Moreover, given a research area, one can rank conferences/journals as well as authors based on the reputation, productivity, and some combined standards. Furthermore, one can observe the evolution of a research area over time, identifying an area or a conference as "*rising star*", "*peak time*", or "*declining*".

## 5. USER INTERFACE AND VISUALIZATION

User-friendly interface and visualization packages form the top-layer of the BibNetMiner system and will play an essential role in its usability. The design encourages user interaction and provides various facilities for explorative and multidimensional analysis of the bibliographic data, including clustering, ranking, profiling, and evolution analysis.

The system provides a set of typical, pull-down form-based user interface so that a user can interact with the system directly by posing a set of queries to ask clusters, ranks, profiling of authors, conferences, and research areas, given a set of query instantiations and constraints.

Alternatively, one can access the system using a DataScope-like [2] visualization-driven content browsing mode, as described below.

Initially, the system presents a high-level overall structure of the data, *e.g.*. the top-level clusters by areas in DBLP (*e.g.*, AI, Database, Theory, *etc.*.), and shows the most influential authors in each cluster. Given a limited screen space, the user is able to first get a high-level view of the data and then quickly zoom-in (or drill-down) to select the subset of data. The system allows users to progressively expand the clusters and explore the data at multi-level of resolution.

Similarly, the clustering-ranking analysis can be applied to different attributes of the data, such as clustering authors by fields, or group conferences based on their ranks. For ranking analysis, users are allowed to customize the system based on their own measure of interestingness. For example, one may rank the authors according to the total number of publications, or according to their prestige in a particular area.

In general, the browsing model will provide the following major functions: (1) explore the high-level structure of data (*e.g.*, show the clusters and link distribution); (2) browse interesting data first and then quickly narrow down the scope of data; and (3) customize parameters so that desired results can be generated accordingly.

To efficiently implement such a browsing mode, the system will take the results obtained in the core link analysis (the bottom layer), compute ranking cube as partial materialization based on the ranking-cube algorithm [3], and utilize the part of the DataScope system interface [2] to visualize the analysis results.

## 6. ABOUT THE DEMONSTRATION

The BibNetMiner is currently being implemented in C++ uder the Microsoft Windows XP system. Its user-friendly visual exploration component will take part of the DataScope system [2] that has been demonstrated at VLDB'07. The most updated DBLP data set will be used as the demonstration data set. The system will be made publicly web-accessible so that it can be popularly used for the exploration of the DBLP data.

## 7. REFERENCES


[1] J. Leskovec, J. Kleinberg, and C. Faloutsos. Graphs over time: Densification laws, shrinking diameters and possible explanations. In *KDD'05*, pp. 177–187, Chicago, IL, Aug. 2005.

[2] T. Wu, X. Li, D. Xin, J. Han, J. Lee, and R. Redder. Datascope: Viewing database contents in google maps' way. In *VLDB'07 (demo)*, Vienna, Austria, Sept. 2007.

[3] D. Xin, J. Han, H. Cheng, and X. Li. Answering top-k queries with multi-dimensional selections: The ranking cube approach. In *VLDB'06*, Seoul, Korea, Sept. 2006.

[4] X. Yin, J. Han, and P. S. Yu. Linkclus: Efficient clustering via heterogeneous semantic links. In *VLDB'06*, Seoul, Korea, Sept. 2006.

[5] X. Yin, J. Han, and P. S. Yu. Object distinction: Distinguishing objects with identical names by link analysis. In *ICDE'07*, Istanbul, Turkey, April 2007.

[6] X. Yin, J. Han, and P. S. Yu. Truth discovery with multiple conflicting information providers on the web. In *KDD'07*, San Jose, CA, Aug. 2007.