DBconnect: Mining Research Community on DBLP Data

Osmar R. Zaïane Jiyang Chen Randy Goebel
Department of Computing Science
University of Alberta, Canada
{zaiane, jiyang, goebel}@cs.ualberta.ca

ABSTRACT

Extracting information from large collections of structured, semi-structured or even unstructured data can be a considerable challenge when much of the hidden information is implicit within relationships among entities within the data. Social networks are such data collections in which relationships play a vital role in the knowledge these networks can convey. A bibliographic database is an essential tool for the research community, yet finding and making use of relationships comprised within such a social network is difficult. In this paper we introduce **DBconnect**, a prototype that exploits the social network coded within the **DBLP** database by drawing on a new random walk approach to reveal interesting knowledge about the research community and even recommend collaborations.

1. INTRODUCTION

A social network is a structure made up of nodes, representing entities from different groups, that are linked with different types of relations. Viewing and understanding social relationships between individuals or other entities is known as *Social Network Analysis* (SNA). SNA methods [25] are used to study organizational relations, analyze citation or computer mediated communications, etc. There are many applications such as studying the spread of disease, understanding the flow of communication within and between organizations, and so on. As an important field in SNA, *Community Mining* [4, 15] has received considerable attention over the last few years. A community can be defined as a group of entities that share similar properties or connect to each other via certain relations. Identifying these connections and locating entities in different communities is an important goal of community mining and can also have various applications. We are interested in the application for finding potential collaborators for researchers by discovering communities in an author-conference social network, or recommending books (or other products) for users based on the borrowing records of other members of their communities in a library system. In this paper we are focusing on the social network implicit in the DBLP database which includes information about authors, their papers and the conferences they published in. DBLP [12, 2] is an on-line resource providing bibliographic information on major computer science conference proceedings and journals[1]. It is such an essential index for the community that it was included in the ACM SIGMOD Anthology[2].

In SNA, the closeness of two related concepts in the network is usually measured by a relevance score, which is based on selected relationships between entities. It can be computed with various techniques, e.g., *Euclidean distance* or *Pearson correlation* [25]. Here we use the random walk approach to determine the relevance score between two entities. A random walk is a sequence of nodes in a graph such that when moving from one node n to the subsequent one in the sequence, one of n's neighbours is selected at random but with an edge weight taken into account. The closeness of a node b with respect to a node a is the static steady-state probability that the sequence of the nodes would include b when the random walk starts in a. This probability is computed iteratively until convergence, and is used as an estimated relevance score. In this paper, we adapt a variation of this idea, which is the random walk with restart (RWR): given a graph and a starting node a, at each step, we move to a neighbour of the current node at random, proportionally to the available edge weights, or go back to the initial node a with a restart probability c. RWR has been applied to many fields, e.g. anomaly detection [22], automatic image captioning [17], etc.

In this paper, we use DBLP data to generate bipartite (author-conference) and tripartite (author-conference-topics) graph models, where topics are frequent n-grams extracted from paper titles. Moreover, we present an iterative random walk algorithm on these models to compute the relevance score between authors to discover the communities. We take into consideration the co-authorship while designing graphical models and the algorithm. We also present our ongoing work DBconnect, which provides an interactive interface for navigating the DBLP community structure online, as well as recommendations and explanations for these recommendations.

The rest of the paper is organized as follows. We discuss related work in Section 2. Graph models and Random walk algorithms for computing the relevance score are described in Section 3. The result and the ongoing DBconnect work

Joint 9th WEBKDD and 1st SNA-KDD Workshop '07 (WebKDD/SNA-KDD'07), August 12, 2007 , San Jose, California , USA.

[1]In April 2007, DBLP comprised more than 870,000 entries.
[2]http://acm.org/sigmod/dblp/db/anthology.html

is reported in Section 4 before the paper is concluded in Section 5.

2. RELATED WORK

Community Mining. The ability to find communities within large social networks could be of important use, e.g., communities in a biochemical network might correspond to functional units of the same kind [7]. Since social networks can be easily modeled as graphs, finding communities in graphs, where groups of vertices within which connections are dense, but between which connections are sparser, has recently received considerable interests. Traditional algorithms, such as the spectral bisection method [18], which is based on the eigenvectors of the graph Laplacian, and the Kernighan-Lin algorithm [10], which greedily optimizes the number of within- and between-community edges, suffer from the fact that they only bisect graphs. While a larger number of communities can be identified by repeated bisection, there is no hint of when to stop the repeated partitioning process. Another approach to find communities is hierarchical clustering based on similarity measures between objects, but it cannot handle the case where some vertices are not close enough to any of the principal communities to be clustered. In the last few years, several methods have been developed based on iterative removal of between-community edges [4, 19, 24]. Important results on researcher community mining have been revealed by analysis of a co-relation (e.g., co-authorship in a paper or co-starring in a movie) graph. While Nascimento et al. [14] and Smeaton et al. [20] show some interesting properties of co-authorship graphs for several selected conferences, the Erdös Number Project[3] and the Oracle of Bacon[4] compute the minimum path length between one fixed person and all other people in the graph.

Community Information System. A related contribution in the context of recommending future collaborators based on their communities is W-RECMAS, which is an academic recommendation system developed by Cazella et al. [13]. The approach is based on collaborative filtering on the user profile data of the Brazilian e-government's database and can aid scientists by identifying people in the same research field or with similar interests in order to help exchange ideas and create academic communities. However, researchers need to post and update their research interests and personal information in the database before they can be recognized and recommended by the system, which makes the approach impractical. In order to efficiently browse the DBLP bibliographical database [12], Klink et al. [11] developed a specialized tool, DBL-Browser, which provides an interactive user interface and essential functionalities such as searching and filtering to help the user navigate through the complex data of DBLP. Another project to explore information for research communities is the DBLife system[5]. It extracts information from web resources, e.g., mailing list archives, newsletters, well-known community websites or research homepages, and provides various services to exploit the generated entity relationship graph [3]. While they do not disclose the process and the means used, they provide related researchers and related topics to a given researcher. In addition to the DBLife project supported by Yahoo, Microsoft Research Asia also developed a similar project called Libra[6], which discovers connected authors, conferences and journals etc. However, in our own experience of using the two systems, we found some incorrect instances of these related entities. Distinct from DBLife and Libra, our DBconnect focuses on finding related researchers more accurately based on a historical publication database and explicit existing relationships in the DBLP coded social network. Moreover, DBlife and Libra do not provide recommendations like DBconnect does.

Random Walk Algorithm. As a popular metric to measure the similarity between entities, the random walk algorithm has received increasing attention after the undeniable success of the Google search engine, which applies a random walk approach to rank web pages in its search result as well as the list of visited pages to re-index [1]. Specifically, Page-Rank [16] learns ranks of web pages, which are N-dimensional vectors, by using an iterated method on the adjacency matrix of the entire web graph. In order to yield more accurate search results, Topic-Sensitive PageRank [5] pre-computes a set of biased PageRank vectors, which emphasize the effect of particular representative topic keywords to increase the importance of certain web pages. Those are used to generate query-specific importance scores. Alternatively, SimRank [8] computes a purely structural score that is independent of domain-specific information. Similar random walk approaches have been used in other domains. For example, the Mixed Media Graph [17] applies a random walk on cross-modal correlation discovery. He et al. [6] propose a framework named MRBIR using a random walk for content-based image retrieval. Sun et al. [22] detect anomaly data in bipartite graphs using the random walk with restart algorithm. Recent work by Tong et al. [23] proposed a fast solution for applying random walk with restart on large graphs, to save pre-computation cost and reduce query time with some cost on accuracy.

In this paper, we extend the traditional random walk approach on tripartite graphs to include topic information, and increase the versatility of the random walk by expanding the original graph model with virtual nodes that take the co-authorship into consideration for the DBLP data. These extensions are explained in the following section.

3. PROPOSED METHOD

Searching for relevant conferences, similar authors, and interesting topics is more important than ever before, and is considered an essential tool by many in the research community such as finding reviewers for journal papers or inviting program committee members for conferences. However, finding relationships between authors and thematically similar publications is becoming more difficult because of the mass of information and the rapid growth of the number of scientific workers [11]. Moreover, except direct co-authorships which are explicit in the bibliographical data, relationships between nodes in this complex social network are difficult to detect by traditional methods. In order to understand relations between entities and find accurate researcher communities, we need to take into consideration not only the information of who people work with, i.e. co-authors, but also where they submit their work to, i.e., conferences, and

[3]http://www.oakland.edu/~grossman/erdoshp.html
[4]http://www.oracleofbacon.org/
[5]http://dblife.cs.wisc.edu/
[6]http://libra.msra.cn/

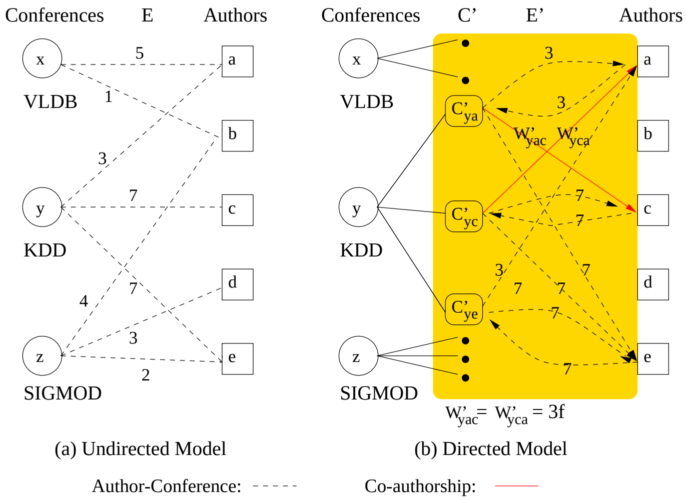

Figure 1: Bipartite Model for Conference-Author Social Network

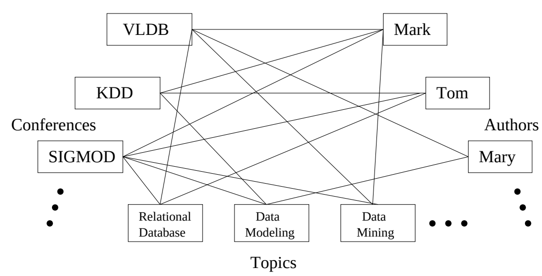

Figure 2: Tripartite Graph Model for Author-Conference-Topic

what they work on, i.e. topics. In this section, we first present the models that incorporate these concepts, then discuss the algorithms that compute the relevance scores for these models.

Given the DBLP database $D = (C \cup A)$, where conference set $C = \{c_i | 1 \leq i \leq n\}$ and author set $A = \{a_j | 1 \leq j \leq m\}$, we can model D as an undirected bipartite graph $G = (C, A, E)$: conference nodes and author nodes are connected if the corresponding author published in the conference and there are no edges in E within the same group of nodes. Figure 1 (a) shows an example of the bipartite graph, representing social relationships between conference and author entities. The weights of the edges are publishing frequency of different authors in a certain conference.

3.1 Adding Topic Information

As mentioned before, the research topic is an important component to differentiate any research community. Authors that attend the same conferences might work on various topics. Therefore, topic entities should be treated separately from conference and author entities. Figure 2 shows an example of linked author, conference, and topic entities. DBLP contains table of contents of some conference proceedings. These table of contents include session titles that could be considered as topics. Unfortunately, very few conference proceedings have their table of contents included in DBLP, and in the affirmative, session titles are often absent. To extract relevant topics from DBLP we resorted to the paper titles instead. We found that frequent co-locations in title text constitute reliable representation of topics. We concede that other methods are possible to get effective research topics.

We now consider a publication database $D = (C \cup A \cup T)$, where topic set $T = \{t_i | 1 \leq i \leq l\}$. We naturally use a tripartite graph to model such data: author/conference nodes are related to a topic node if they have a paper on that topic, the edge weight is the topic matching frequency. We apply the random walk algorithm on a tripartite graph by adjusting the walking sequence. For example, previously the random walker turns back to author nodes when it reaches a conference node; now it will go forward to topic nodes first, and then walk back to author nodes. By such modifications, the relevance score now contains both conference and topic influences, i.e., in a tripartite model, authors with high relevance score share similar conference experiences and paper topics with the given author.

	Publication Records
VLDB(x)	a(4), ab(1)
KDD(y)	ac(3), c(4), e(7)
SIGMOD(z)	b(4), d(1), de(2)

Table 1: Author Publication Records in Conferences. For example, a, b, c, d, e are authors, $ac(3)$ means that author a and c published three papers together in a certain conference.

3.2 Adding Co-author Relations

Table 1 shows the number of publications of five authors a, b, c, d, e in three conferences VLDB KDD and SIGMOD. Authors a and c have co-authored 3 papers in KDD, a and b co-authored 1 paper in VLDB and d, e co-authored 2 papers in SIGMOD. Unfortunately, the corresponding bipartite graph, which is shown in Figure 1 (a), fails to represent any co-authorships. For example, author a and c co-authored many papers at KDD, but there are no edges in the bipartite graph that can be used to represent this information: edge $e(y, a)$ and $e(y, c)$ are both used by relations between conference and author. On the other hand, author e seems more related to author c since the weights of edges connecting them to KDD are the heaviest ($w_{yc} = 7$, $w_{ye} = 7$). The influence of the important co-author a is neglected because the model only represents publication frequency.

To capture the co-author relations, just adding a link between a and c does not suffice, since it misses the role of KDD, where the co-authorship happens. Making the link connecting a and c to KDD directional does not work either, as from KDD there are edges to many other authors, which would make the random walk infeasible (i.e., yielding undesirable results). Our approach is to re-structure the bipartite model by adding surrogate nodes of KDD and having them link to a and c so that the random walk calculation can be applied while the connection between related nodes remains the same. In more detail, we add a virtual level of nodes between the two sides, and add direction to the edges. Figure 1 (b) shows details of node KDD as an example. We first split y into 3 nodes to represent relations between y and authors who published there (a, c and e). These author

Algorithm 1 The Random Walk with Restart Algorithm

Input: node $\alpha \in A$, bipartite graph model G, restarting probability c, converge threshold ϵ.
Output: relevance score vector $\vec{A}$ for author nodes.
1. Construct graph model G' for co-authorship based on G. Compute the adjacency matrix J of G'.
2. Initialize $\vec{v}_\alpha = 0$.
set value for α to 1: $\vec{v}_\alpha(\alpha) = 1$.
3. While $(\Delta\vec{u}_\alpha > \epsilon$)
$\vec{u}_\alpha = (1-c)(\frac{(Norm(N))^T \vec{u}_{\alpha(n*m+1:n*m+m)}}{Norm(M)^T \vec{u}_{\alpha(1:n*m)}}) + c\vec{v}_\alpha$
4. Set vector $\vec{A} = \vec{u}_{\alpha(n*m+1:n*m+m)}$
5. Return $\vec{A}$.

nodes then connect to their own splitted relation nodes with the original weight $(e'(a, C'_{ya}), e'(c, C'_{yc}), e'(e, C'_{ye}))$. Then we connect from C' nodes to all author nodes that have published at KDD. If the author node has a co-author relation with the author included in the C' node, the edge is weighted by co-author frequency multiplied by a parameter f (which is explained in the following), otherwise, the edge is weighted as original. We can see that the co-authorship, which is missed in the simple bipartite graph, is now represented by extra weight of edge $e'(C'_{yc}, a)$ and $e'(C'_{ya}, c)$, which shows author a is more related to c then author e through KDD due to their collaborations. The parameter f is used to control the co-author influence, usually we set $f = k$ (k is the total author number of a conference).

3.3 Random Walk on DBLP Social Network

Before presenting the random walk algorithms, we define the problems we are solving: given an author node $a \in A$, we compute a relevance score for each author $b \in A$. The result is a one-column vector containing all author scores with respect to a. We measure closeness of researchers so we can discover implicit communities in the DBLP data.

Recall that we extend the traditional model into a directed bipartite graph $G' = (C', A, E')$, where A has m author nodes, C' is generated base on C and has $n * m$ nodes (we assume every node in C is split into m nodes). The basic intuition of our approach is to apply random walks on the adjacency matrix of graph G' starting from a given author node. To form the adjacency matrix, we first generate a matrix for directional edges from C' to A, which is $M_{(n*m)\times m}$, then form a matrix for edges from A to C', which is $N_{m\times(n*m)}$. In these two matrices, $M(\alpha, \beta)$ or $N(\alpha, \beta)$ indicates the weight of the directed edge from node α to node β in G' (0 means no such edge). A random walk starting from a node represented by row α in M (the same applies to N) takes the edge (α, β) based on the probability which is proportional to the edge weight over the weigher of all outgoing edges of α. Therefore, we normalize M and N such that every row sums up to 1. We can then construct the adjacency matrix J of G':

$$J_{(n*m+m)\times(m+n*m)} = \begin{pmatrix} 0 & (Norm(N))^T \\ (Norm(M))^T & 0 \end{pmatrix}$$

We then transform the given author node α into a one-column vector $\vec{v}_\alpha$ consisting of $(n * m + m)$ elements. The value of the element corresponding to author α is set to 1. We now need to compute a steady-state vector $\vec{u}_\alpha$, which contains relevance scores of all nodes in the graph model.

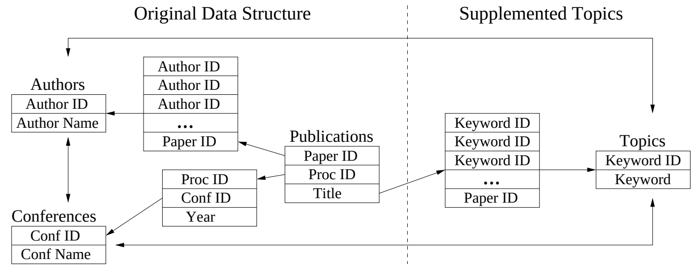

Figure 3: Our Data Structure extracted from DBLP

The scores for the author nodes are the last m elements of the vector. The result is achieved based on the following lemma and the RWR approach.

LEMMA 1. *Let c be the probability of restarting random walk from node α. Then the steady-state vector $\vec{u}_\alpha$ satisfies the following equation:*

$$\vec{u}_\alpha = (1-c)J\vec{u}_\alpha + c\vec{v}_\alpha$$

See [21] for proof of the lemma.

Algorithm 1 applies the above lemma repeatedly until $\vec{u}_\alpha$ converges. We set c to be 0.15 and ϵ to be 0.1, which gives the best convergence rate according to [22]. The bipartite structure of the graph model is used to save the computation of applying Lemma 1 in step 3. The last m elements of the result vector $\vec{u}_{\alpha(n*m+1:n*m+m)}$ contains the relevance score for all author nodes in A.

We extend algorithm 1 for the tripartite graph model (The proof for Lemma 1 on the tripartite graph is trivial and is omitted.). Similarly, we first form three adjacency matrices between any two of the three entity types (conference, author and topic), then apply the RWR approach following the visiting sequence until convergence, e.g., walk from author to conference, to topic, and back to author if we want to rank authors. There are $m + n + l$ elements for all nodes in the graph model in the result relevance score vector. The value of the corresponding node, either starting author, topic or conference, is initialized to 1. After the random walk algorithm terminates, scores for conference, author and topic nodes are recorded from 1 to n, from $n + 1$ to $n + m$ and from $n + m + 1$ to $n + m + l$ in the vector, respectively.

4. EXPLORING DBLP COMMUNITIES

In the academic world, since a researcher could usually belong to multiple related communities, e.g., Database and AI, it is unnecessary and improper to classify this researcher into any specific arbitrary communities. Therefore, in our experiment, we focus on investigating the closeness of researchers, i.e., we are interested at *how* and *why* two people are in the same community, instead of *which* community they are in.

4.1 DBLP Database

We downloaded the publication data for conferences from the DBLP website[7] in November 2006. Any publication after that date is not included in our experimental data. Moreover, we kept only conference proceedings and removed all journals and other publications. These were minimal

[7]http://www.informatik.uni-trier.de/~ley/db/

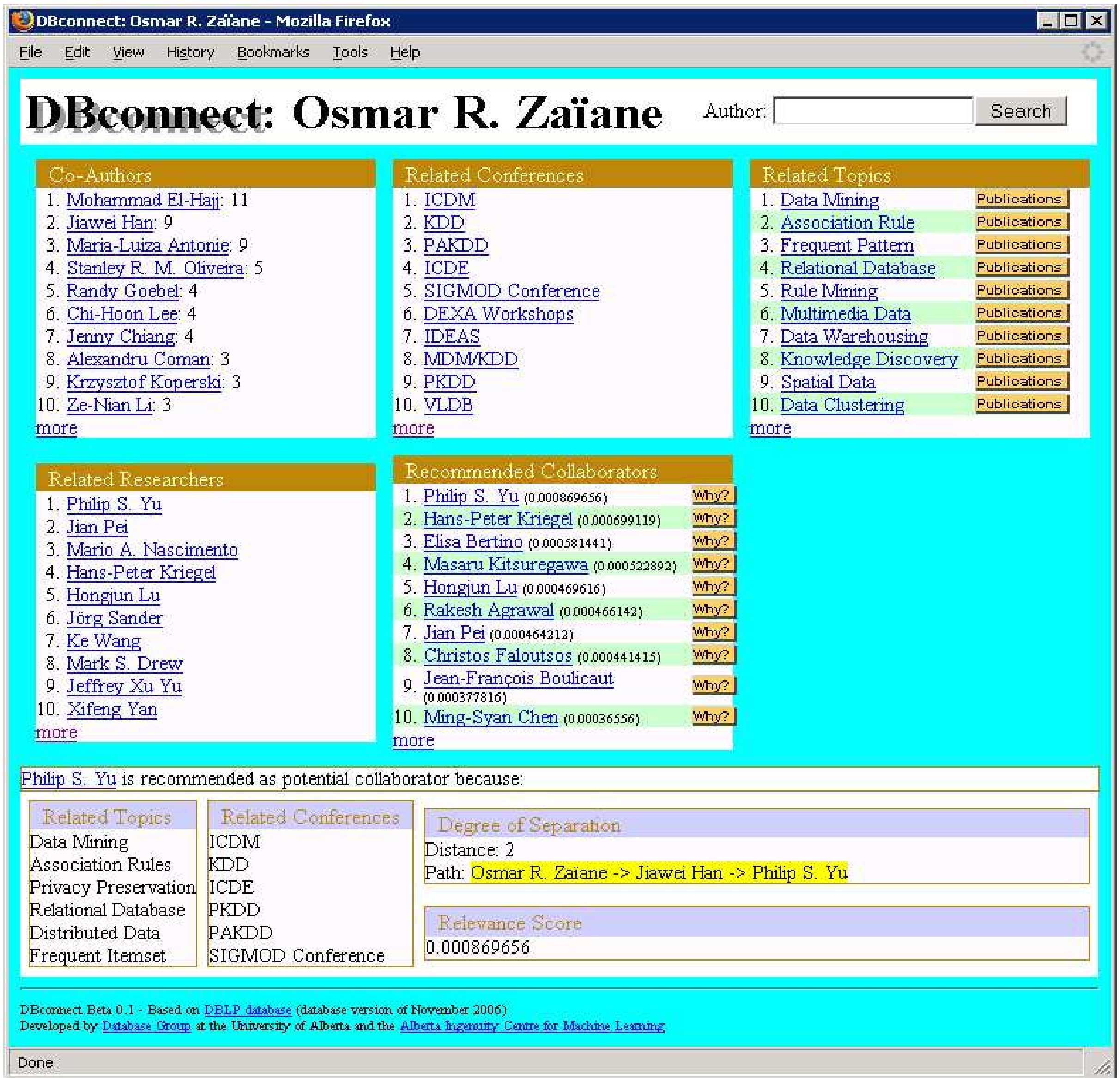

Figure 4: DBconnect Interface Screenshot for an author

compared to the conference publications. The data structure is shown in Figure 3. We extracted topics based on keyword frequency from paper titles, which are the only provided content related information. We first manually selected a list of stopwords to remove frequently used but non-topic-related words, e.g., "Towards", "Understanding", "Approach", etc. Then we counted frequency of every co-located pairs of stemmed words and selected the top 1000 most frequent bi-grams as topics. Additionally, we manually added several tri-grams, e.g. World Wide Web, Support Vector Machine, etc., since we observe both bilateral bi-grams to be frequent (e.g. World Wide and Wide Web). We chose to use bi-grams because they can distinguish most of the research topics, e.g, Relational Database, Web Service and Neural Network, while single keywords fail to separate different topics, e.g. "Network" can be part of "Social Network" or "Network Security".

Since the publication database is huge (it contains 367,654 authors, 3,290 conferences and the selected 1,000 N-gram topics), the entire adjacency matrix becomes too big to fit in main memory to make the random walk efficient. However, we can compute the result by first performing graph partitioning on the model and only running the random walk on the part where the given author is. This approach can only achieve an approximate result, since some weakly connected communities are separated, but it is much faster since we end-up computing with much smaller matrices. In this paper, we used the METIS algorithm [9] to partition the large graph into ten subgraphs of about the same size. Note that the proposed approach is independent of the selected partitioning method.

4.2 The DBconnect System

After the author-conference-topic data extraction from

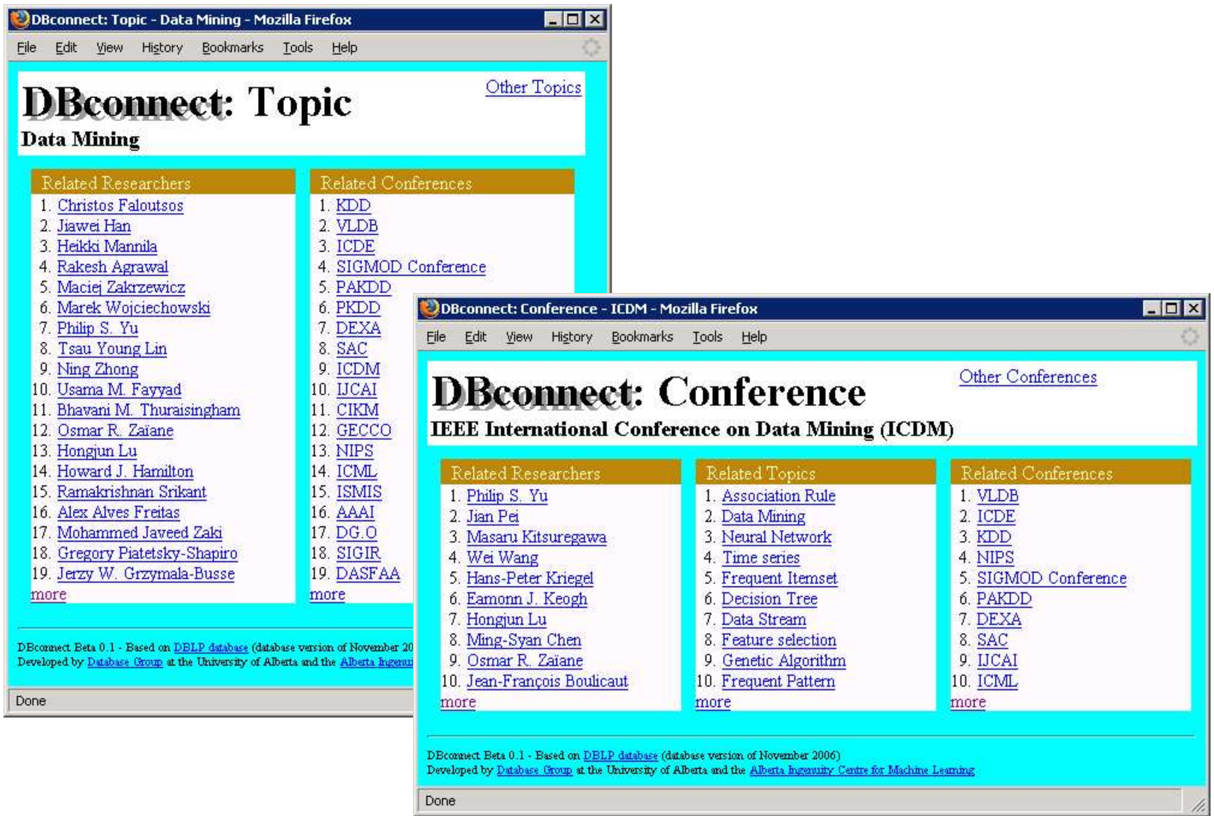

Figure 5: DBconnect Interface Screenshot for topic Data Mining and conference ICDM

the DBLP database, we generate lists of people with high relevance scores with respect to different given researchers. Our ongoing project *DBconnect*, which is a navigational system to investigate the community connections and relationships, is built to explore the result lists of our random walk approach on the academic social network. Figure 4 shows a screenshot of the interface of our *DBconnect* system. There are five lists displayed for a given author in the current version. Clicking on any of the hyper-linked names will generate a page with respect to that selected entity.

- *Co-authors* list reports the number of papers that different researchers co-authored with the given person. Since co-authors are treated as "observed collaborators", their names will not be seen in the other lists such as related researchers.

- *Related Conferences* list is generated by the random walk on an author-conference-topic model and is ordered by their relevance score, in descending order. These are not necessarily the conferences where the given researcher published but the conferences related to the topics and authors that are also related to the reference researcher. Clicking on the conference name leads to a new page with topics and authors related to the chosen conference. Figure 5 illustrates an example when the link "ICDM" is selected. Conferences have their own related conferences. Note that the topics here mean the most frequent topics used within titles of papers published in the given conference.

- *Related Topics* list is ordered by the relevance scores from a random walk on the tripartite model. Clicking on the button "Publications" after each topic provides the papers that the given author has published on that topic, i.e. the papers whose title contains the N-gram keywords. Clicking on the topic itself leads to a new page with conferences and authors related to the chosen topic. Note again that this relationship to topics comes from paper titles. Figure 5 shows an example when the topic "Data Mining" is selected.

- *Related Researchers* list is based on the bipartite graph model with only conference and author entities, but the co-authorship emphasized i.e. we use our extended bipartite graph. The result implies that the given author, Osmar Zaïane, is related to the same conferences and via the same co-authors as these listed researchers.

- *Recommended Collaborators* lists related researchers based on the tripartite graph author-conference-topic. The result implies that the given author, Osmar Zaïane, shares similar topics and conference experiences with these listed researchers, hence the recommendation. The relevance score calculated by our random walk is displayed following the names. Clicking on the "why" button brings the detailed information of the relationship between the two authors. For example, in Figure 4, relations between Philip Yu and Osmar Zaïane are described by the six top topics and conferences they share, and the degree of separation in the co-authorship chain ($A \to B$ means A and B are co-authors).

Note that while there is some overlap between the list of related researchers and the list of recommended collaborators, there is a fundamental difference and the difference by no means implies that collaboration with the missing related researchers is discouraged. They are simply two different communities in the network even though they overlap. The list of related researchers is obtained from relationships derived from co-authorships and conferences by a RWR on an extended bipartite graph with returning relations. The result is a quasi-obvious list due to the closeness from co-authors. Co-authors themselves are not included. This list could create a sort of trust in the system given the clear closeness of this community. The list of recommended collaborators could be perceived as a more distant community and thus as an interesting discovery. It is obtained without co-authorship but with relations from topics. We use a RWR on a tripartite graph authors/conferences/topics. The explanation on the why collaborators are recommended (i.e. common conferences and topics, and degree of separation) establishes more trust in the recommendation. A systematic validation of these lists is difficult but the cases we manually substantiated were satisfactory and convincing. We are considering a user feedback for validation and a more methodical corroboration based on chronological entries within DBLP data.

5. CONCLUSIONS AND FUTURE WORK

In this paper, we extend the traditional bipartite graph model to incorporate co-authorship, and propose a random walk approach on the new model to find related conferences, authors, and topics for a given author. The main idea is to use a random walk with restarts on the bipartite or tripartite model of DBLP data to measure the closeness between any two entities. The result, the relevance score, can be used to understand the relationship between entities and discover the community structure of the corresponding data. We basically use the relevance score to rank entities based on importance given a relationship.

We also present our ongoing work DBconnect, which can help explore the relational structure and discover implicit knowledge within the DBLP data collection. Not all of the more than 360,000 authors are indexed in DBconnect at the time of printing as the random walks are time consuming. A queue of authors is continuously processed in parallel and authors can be prioritized in the queue by request.

The work we presented in this paper is still preliminary. We have implemented a prototype. However, more work is needed to verify the value of the approach. The lists of related conferences, topics and researchers to a given author are interesting and can be used to help understand the entity closeness and research communities. While the output of DBconnect is satisfactory and the manual substantiation confirms acceptable and suitable lists (as opposed to lists provided by DBLife), some systematic evaluation is still desired. However, validation of the random walk is difficult and we are considering devising methods to confirm the accuracy of the relevance score and the generated lists. Moreover, it is hard to extract correct topics for researchers since the only available information is the title of the paper, which usually does not suffice to describe the content. Some titles are even unconventionally unrelated to the content of the paper only to attract attention or metaphoric. We are considering implementing a hierarchy of topics to group similar topics and ease the browsing of the long list of related topics in computer science. We also plan to address the issue of acronyms in titles that are currently discarded. For example HMM for Hidden Markov Model is currently eliminated due to infrequency while relevant as a topic. In addition, the matrix multiplications in the random walk process make it expensive to compute. Improving the efficiency of the random walk without jeopardizing its effectiveness is necessary since the computations for relevance score estimation need to be redone continuously as the the DBLP database never ceases to grow.

6. ACKNOWLEDGMENTS

Our work is supported by the Canadian Natural Sciences and Engineering Research Council (NSERC), the Alberta Ingenuity Centre for Machine Learning (AICML), and the Alberta Informatics Circle of Research Excellence (iCORE).

7. REFERENCES

[1] Sergey Brin and Lawrence Page. The anatomy of a large-scale hypertextual web search engine. In *Seventh International World Wide Web Conference*, pages 107–117, Brisbane, Australia, 1998.

[2] DBLP Bibliography database. http://www.informatik.uni-trier.de/~ley/db/.

[3] AnHai Doan, Raghu Ramakrishnan, Fei Chen, Pedro DeRose, Yoonkyong Lee, Robert McCann, Mayssam Sayyadian, and Warren Shen. Community information management. *IEEE Data Engineering Bulletin, Special Issue on Probabilistic Databases*, 29(1), 2006.

[4] Michelle Girvan and M. E. J. Newman. Community structure in social and biological networks. In *Proceedings of the National Academy of Science USA, 99:8271-8276*, 2002.

[5] Taher H. Haveliwala. Topic-sensitive pagerank. In *WWW: Proceedings of the 11th international conference on World Wide Web*, pages 517–526, 2002.

[6] Jingrui He, Mingjing Li, Hong-Jiang Zhang, Hanghang Tong, and Changshui Zhang. Manifold-ranking based image retrieval. In *MULTIMEDIA: Proceedings of the 12th annual ACM international conference on Multimedia*, pages 9–16, 2004.

[7] P. Holme, M. Huss, and H. Jeong. Subnetwork hierarchies of biochemical pathways. *Bioinformatics*, 19:532–538, 2003.

[8] Glen Jeh and Jennifer Widom. Simrank: a measure of structural-context similarity. In *KDD*, 2002.

[9] George Karypis and Vipin Kumar. Multilevel k-way partitioning scheme for irregular graphs. *Journal of Parallel and Distriuted Computing*, 48(1):96–129, 1998.

[10] B. W. Kernighan and S. Lin. An efficient heuristic procedure for partitioning graphs. *Bell System Technical Journal*, 49:291–307, 1970.

[11] Stefan Klink, Patrick Reuther, Alexander Weber, Bernd Walter, and Michael Ley. Analysing social networks within bibliographical data. In *DEXA*, pages 234–243, 2006.

[12] Michael Ley. The DBLP computer science

bibliography: Evolution, research issues, perspectives. In *SPIRE*, pages 1–10, 2002.

[13] Sìlvio César Cazella and Luis Otávio Campos Alvares. An architecture based on multi-agent system and data mining for recommending research papers and researchers. In *Proc. of the 18th International Conference on Software Engineering and Knowledge Engineering (SEKE)*, pages 67–72, 2006.

[14] Mario A. Nascimento, Jörg Sander, and Jeffrey Pound. Analysis of sigmod's co-authorship graph. *SIGMOD Record*, 32(2):57–58, 2003.

[15] M. E. J. Newman. The structure and function of complex networks. *SIAM Review*, 45(2):167–256, 2003.

[16] Lawrence Page, Sergey Brin, Rajeev Motwani, and Terry Winograd. The pagerank citation ranking: Bringing order to the web. In *Technical report, Stanford University Database Group*, 1998.

[17] Jia-Yu Pan, Hyung-Jeong Yang, Christos Faloutsos, and Pinar Duygulu. Automatic multimedia cross-modal correlation discovery. In *KDD*, pages 653–658, 2004.

[18] A. Pothen, H. Simon, and K. P. Liou. Partitioning sparse matrices with eigenvectorsof graphs. *SIAM J. Matrix Anal. Appl.*, 11:430–452, 1990.

[19] Filippo Radicchi, Claudio Castellano, Federico Cecconi, Vittorio Loreto, and Domenico Parisi. Defining and identifying communities in networks. *PROC.NATL.ACAD.SCI.USA*, 101:2658, 2004.

[20] A. F. Smeaton, G. Keogh, C. Gurrin, K. McDonald, and T. Sodring. Analysis of papers from twenty-five years of sigir conferences: What have we been doing for the last quarter of a century. *SIGIR Forum*, 36(2):39–43, 2002.

[21] Gilbert Strang. Introduction to linear algebra, Wellesley-Cambridge Press, 3 Edition, 1998.

[22] Jimeng Sun, Huiming Qu, Deepayan Chakrabarti, and Christos Faloutsos. Neighborhood formation and anomaly detection in bipartite graphs. In *ICDM*, pages 418–425, 2005.

[23] Hanghang Tong, Christos Faloutsos, and Jia-Yu Pan. Fast random walk with restart and its applications. In *ICDM*, pages 613–622, 2006.

[24] Joshua R. Tyler, Dennis M. Wilkinson, and Bernardo A. Huberman. Email as spectroscopy: automated discovery of community structure within organizations. *Communities and technologies*, pages 81–96, 2003.

[25] S. Wasserman and K. Faust. Social network analysis: Methods and applications, Cambridge University Press, 1994.