2009 Advances in Social Network Analysis and Mining

Social Network Analysis as Knowledge Discovery process: a case study on Digital Bibliography

Michele Coscia, Fosca Giannotti, Ruggero Pensa
ISTI-CNR
Pisa, Italy
Email: name.surname@isti.cnr.it

***Abstract*—Today Digital Bibliographies are a powerful instrument that collects a great amount of data about scientific publications. Digital Bibliographies have been used as basis of many studies focused on the knowledge extraction in databases. Here we present a new methodology for mining knowledge in this field. Our approach aims to apply the potential of social network analysis techniques to accomplish this task, using a network representation of bibliography data. Besides we use some data mining techniques applied on social network representations in order to enrich this new point of view and to evolve our methodology towards a comprehensive local and global bibliography analysis workflow seen as a Knowledge Discovery process.**

***Keywords*-Social Network, Data Mining, Digital Bibliography**

I. Introduction

A social network is a set of (groups of) people with some pattern of contacts between them representable with a graph [7]. The recent research field of social network analysis aims at applying computer science analysis techniques to the sociological field and investigates phenomena in real world even of huge cardinality (such as Linkedin). These approaches came from complex modeling theory and are focused to the construction of models such as the Small-world [11] and the Scale Free model [2].

Interests towards this form of data are also emerging within the data mining research community. Data mining aims at the extraction of implicit knowledge from huge databases [1], by investigating patterns and regularities in data. Graph Mining [5] is the sub-area which focuses on mining from graph datasets. The contribution of data mining researchers to social network analysis, related to this work, focuses on two aspects (sometimes overlapping). The first one is the extraction of models by analyzing examples of networks [6], in which data mining is used to identify some critical parameters, based only on structural and not semantical properties, that describe the microscopic evolution of some social networks. The second aspect is focused on analyzing the temporal evolution of the network [3]. Some other issues are the graph mining problem applied to large networks [5] and the community detection in the structure [9]. The closest related work, at the best of our knowledge, can be found in [8], in which there is an attempt of building another data mining process on digital bibliographies, without providing a formal workflow nor some examples of analysis.

In this paper we introduce a methodology which uses the idea of the Knowledge Discovery process as the workflow for a social network analysis process which combines preprocessing, standard social network statistics and data mining analysis in an interactive and iterative way. This methodology enables the analysis of global and local behaviors of a social network and allows the definition of new application domain specific indicators as combination of underlying statistics and data mining tasks. We applied the proposed method to the case study on digital bibliography analysis. Digital bibliographies (such as NCBI, DBLP...) are a powerful instrument that collects a great amount of data about scientific publications. This data includes authors' information (e.g., name and institutions), and publication details (title, keywords, issue, publication date...). Starting from these data it is possible to construct a co-authorship network, whose nodes are authors and edges are created between authors if they have collaborated in at least one paper, which enables to model the underlying collaboration links among different researchers. We describe the entire digital bibliography analysis process from the construction of the network to the extraction of global statistical parameters and descriptor based on local regularities (patterns). Our contribution can be summarized as follows:

- We define the problem of analyzing digital bibliographies as the analysis of a dataset of graphs, instead of a unique large graph;
- We adopt existing data mining techniques to validate, compare and enrich the statistical parameters;
- We use co-clustering (i.e., simultaneous partitioning of rows and columns of a contingency matrix [4]) to assign a research profile (based on frequently used keywords) to each author;
- We show that using exploratory techniques, such as graph mining, allows to extract information that is hard to obtain by using standard query languages.

In particular, we apply our technique to DBLP, a well known computer science digital bibliography.

978-0-7695-3689-7/09 $25.00 © 2009 IEEE
DOI 10.1109/ASONAM.2009.65

IEEE computer society

The rest of the paper is organized as follows. In Section 2 we introduce some useful definitions on bibliography analysis, data mining techniques and social network analysis. In Section 3 we describe the workflow of our methodology. Sections 4 presents our case study. Section 5 concludes.

II. Preliminaries

Definition. We define bibliography analysis as a combination of social network analysis and data mining techniques applied to graphs representing co-authors' networks. A co-authors' network is a network where the nodes are authors and the edges represent one or more collaborations in publishing a (scientific) paper. These techniques are oriented to the semantic description of the network (make explicit the tendency of important authors to isolate newcomers, or attract most of the collaboration links, and so on).

A Co-Authors Warehouse is a collection of edges $\mathcal{E} = \{e_1, e_2, ..., e_n\}$ where each edge $e_i = \{I_{1,i}, I_{2,i}, ...I_{k,i}\}$ in which $I_{1,i}, I_{2,i}, ...I_{k,i}$ are k attributes of the collaboration such as the year, the conference or the title. Applying some filters on the attributes it is possible to obtain the Co-Authors' Network $\mathcal{N}_i = \{ V_i, E_i, I_i\}$ in which there are some attributes attached with any $v \in V_i$ (such as degree, number of publications in the network and class of author) and with any $e \in E_i$ (such as number of publications between the two authors). I_i are the attributes used to obtain the network and that describe it as a whole (such as an interval of years and conferences).

We now briefly introduce the two main analysis instruments that we use within our framework. We first list a number of classical social network analysis descriptors and then we give a short explanation about the data mining techniques of graph mining and co-clustering.

A. Social Network Analysis

Here we present some basic concepts about the network analysis. This is a very brief explanation that focuses only on the aspects that will be used in the bibliography analysis framework. See [7], for an exhaustive survey for the classical social network analysis.

Vertex Degree: Number of edges connected to a vertex.

Component: The set of vertices that can be reached from a chosen vertex by paths running along edges of the graph.

Clustering: is the triadic closure ratio, in other words the number of possible triangles of vertices in the network that have the third edge on the number of all possible triangles (it's the 1-neighborhood clustering). It can be seen also as 2-neighborhood clustering[1] considering the ratio of the number of edges in 1-neighborhood on the number of edges in 2-neighborhood.

B. Data Mining Techniques

Here we present a brief explanation about the two main data mining techniques used in the bibliography analysis framework: *graph mining* and *co-clustering*.

Graph Mining is a particular frequent pattern mining problem, in which the data are represented as a dataset of (labeled) graphs, and a pattern is subgraph common to a sufficient number of graphs [10]. A basic concept of Graph Mining is the graph isomorphism. A graph isomorphism is a bijection between the vertex, edge and label sets of two graphs. Graph Mining is to find every subgraph g in a graph dataset such that the number of its isomorphic graphs is more or equal to a given minimum support threshold.

Co-Clustering [4]. Clustering is a fundamental tool in unsupervised learning that is used to group together similar objects. Given a contingency matrix, co-clustering enable to simultaneously cluster both dimensions of the matrix. A co-clustering is a pair of maps from rows to k row-clusters and from columns to ℓ column-clusters, and the optimal co-clustering is one that leads to the minimum loss in mutual information between the two sets of clusters.

III. Bibliography Analysis: Our Approach

We now explain the workflow (Figure 1) that enable the knowledge discovery process of the bibliography analysis.

The starting point is the database of a digital bibliography. This database is at the basis of two distinct operation. The first one is the extraction of the Co-Authors Warehouse: a data structure in which we record all the information about the scientific collaborations that are present in the database.

The second operation that uses the raw data of the digital bibliography is the building of an author-keywords matrix, in which rows are authors and columns are keywords, obtained with a cleaning process from titles and/or abstracts of their publications. This data structure is used in order to obtain a classification of authors based on the keywords that can describe their works. This process can be implemented in two ways: fully-automatic or with the intervention of an expert of the domain, that can make the classification process more accurate and close to the actual authors' classes. These two structures, are used in the query process. Using a query language it is possible to specify various kind of constraint in order to extract from this data warehouse a graph representation of a particular co-authors' network.

Once the social network is obtained, it is possible to perform two different kinds of analysis. At a global level, one can compute new or standard statistical parameters which are semantically related to the bibliography analysis framework. Thus, we obtain a direct knowledge of the global behavior of the network (**global analysis**). At a local level, we can perform a complete data mining process. Using a graph mining algorithm it is possible to extract local patterns that can be analyzed individually to study some frequent regularities that exist in the network and that can describe

[1] A definition of Batagelj V. for Pajek tool, http://vlado.fmf.uni-lj.si/pub/networks/pajek/

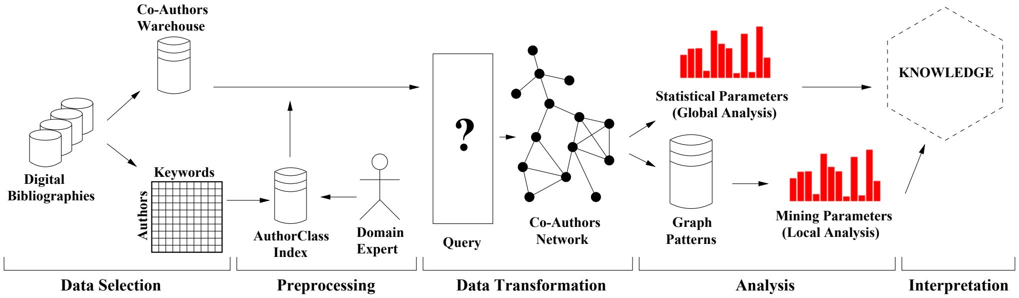

Figure 1. Bibliography Analysis Workflow.

its micro-behavior (**local analysis**). This enable to obtain a different point of view of the knowledge mined from the network. Alternatively, the collection of pattern can be post-processed to compute pattern-based indicators.

IV. Case study

In this section we present an instance of our framework on DBLP[2], a well-known digital bibliography that stores publications (journal articles, conference papers, books...) related to the computer science research domain. We now show some examples of analysis following the introduced workflow: from preprocessing to mining, statistical and combined analysis.

The giant component problem - Preprocessing: A social network is a big, unique and unconnected graph [7], but traditional graph mining algorithms work on datasets of many, small and connected graphs.

The first step of the solution is to consider the unconnected social network as a graph dataset in which every component is an entry of the database. This is a step without any loss of information and solves the issue of the many and connected graphs, but it is not sufficient for the requirement of the dimension of the graphs. In fact in a Social Network an extensive fraction of all vertices (from 30% to 100%) is all linked in a giant component. However, if we consider a single co-authors' network of a single conference in a single year *regardless its size* it won't have any giant component. Then, as shown in Figure 2, a co-authors' network of a single conference and a single year can be easily represented with a dataset of many small connected graphs. But if we consider different conferences in different years an author is now able to connect many different research groups has worked with during his careers.

Indeed, in order to apply graph mining to Social Networks we must define a procedure to obtain a graph dataset of small graphs from the union of networks that generate a giant component. Let $\mathcal{N}_a$ and $\mathcal{N}_b$ be two network that satisfied our set of constraints i and must be used to obtain the network $\mathcal{N}_i$. Let x be an author present in both networks. We define the join operation on networks ($\mathcal{N}_i = \mathcal{N}_a \cdot \mathcal{N}_b$) the operation that gives as result a single vertex x in $\mathcal{N}_i$ connected to all his neighbours from the two networks. We also define the union operation on networks ($\mathcal{N}_i = \mathcal{N}_a \cup \mathcal{N}_b$) the operation that gives as result two distincts vertices x_a and x_b (see Figure 3). We call the first network the Co-Authors Social Network (similar to the network in Figure 2b) and we use it for the statistical analysis, and we call the second one the Co-Authors Graph Dataset (similar to the network in Figure 2a) and we use it for the graph mining analysis.

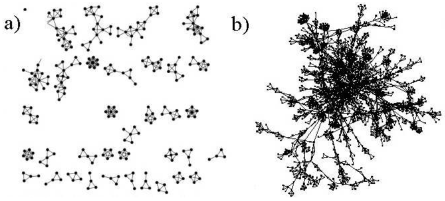

Figure 2. A representation of a graph dataset obtained from KDD conference in a single year (a) and from its junction of six years (b).

Class Graphs - Mining step: With this analysis we want to find, in a given conference or in general, what classes are likely to be joint together, or what competences tend to be useful if joint together. With ”class” we refer to the results of the co-cluster phase, applied to an Author-Keywords matrix that join together authors whose works can be described with similar keywords (there are 120 classes in our work set, see Table 1 for some statistics about the most important considered in this work). Then we can draw graphs like that in Figure 4 for every conference, with classes of competences as nodes and edge as collaboration (weighted with the frequency of that collaboration in the network).

Network Communicative Structure - Statistical step: Computing the Pearson correlation coefficient of degrees and ages of the vertices at both sides of each edge in the network it is possible to find how much the authors with many publications or an high seniorship (many years of publication) tend to isolate the least prolific or young ones. We call these parameters *isolation degrees*. Let d_s and d_t

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

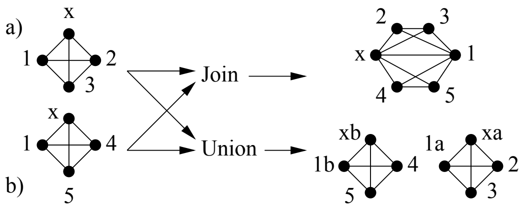

Figure 3. Representation of the two network operations of join and union.

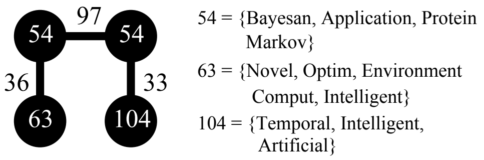
97
54
54
36
33
63
104
54 = {Bayesan, Application, Protein Markov}
63 = {Novel, Optim, Environment Comput, Intelligent}
104 = {Temporal, Intelligent, Artificial}

Figure 4. A class graph for ICML Conference.

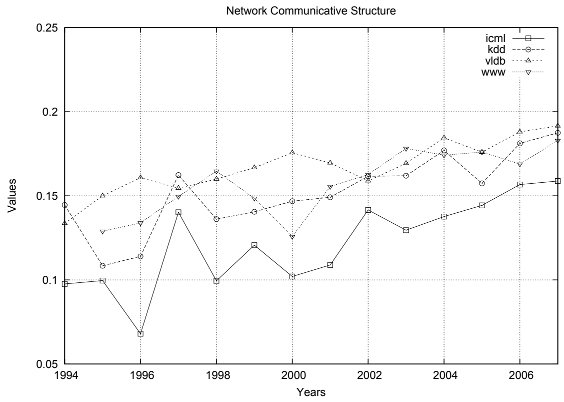

Figure 5. Network Communication Structure parameter for KDD, ICML, VLDB and WWW from 1994 to 2007.

be the degrees of source and target node of each edge and a_s and a_t be the ages (in these series each undirected edge must be considered in both its directions, so N is the double of the number of edges in the network or $N = 2e$):

$$I_d = \frac{\sum d_s d_t - \frac{\sum d_s \sum d_t}{N}}{\sqrt{\left(\sum d_s^2 - \frac{(\sum d_s)^2}{N}\right)\left(\sum d_t^2 - \frac{(\sum d_t)^2}{N}\right)}},$$

$$I_a = \frac{\sum a_s a_t - \frac{\sum a_s \sum a_t}{N}}{\sqrt{\left(\sum a_s^2 - \frac{(\sum a_s)^2}{N}\right)\left(\sum a_t^2 - \frac{(\sum a_t)^2}{N}\right)}}.$$

Being a variation of Pearson Correlation Coefficient, these parameters take values from -1 (anti-correlation = minimum isolation) to 1 (total correlation = maximum isolation).

Using these parameters it is possible to see how much the network's structure is good in creating a virtuous flow of informations. This is inversely proportional to the isolation degrees and directly proportional to the average clustering of the network ($\bar{C}$, the more links the more contacts can be used to reach a particular information):

$$CS = \frac{2\bar{C}}{4 + I_a + I_d},$$

that spans from 0 to 1. When the network has a great isolation the CS tends to $\frac{\bar{C}}{3}$, it means that high values of clustering can still trigger a certain flow of information, but it decreases very rapidly if also clustering is low. When both kinds of isolation tend to 1 CS tend to be equal to the clustering coefficient: it means that the structure is so perfect (each senior member collaborate only with young ones) that the flow of new ideas is very easy and automatic if the authors are well connected each others.

In Figure 5 we can show this parameter computed for KDD, ICML, VLDB and WWW conference from 1994 to 2007. What we see is that all conferences have the tendency of improving their communicative structure, although with different rates. But there are also irregularities in this improvement that can provide some interesting analysis.

KDD conference was born in 1994 with few clusterized authors. In 1997 there is the explosion of data mining with a remarkable peak. After that more and more authors enter in this conference lowering its clustering coefficient. Then the structure stabilizes on a regular improvement. Also WWW was born in 1994/1995. But its structural evolution was heavily disturbed by the economic flop of World Wide Web in the first years of the century. Then the introduction of new research ideas (the Web 2.0) carried a new life blood to this conference, which sign another peak in 2003 and continues its more regular evolution.

Conference's Dominant Class - Statistical and Mining analysis combined: With this analysis we want to find what class for a given conference is the most important. This analysis can be done at two different levels, representing the main example of the combination of mining and statistical approaches in bibliography analysis.

With the mining approach we identify the author's classes present in every connected component. This is different from simply counting the labels of the nodes in a network, because there can be a label that is very frequent but only in few big components. We want to find instead the class that is present in every single graph of the dataset, and we consider that class the most important.

This can tell us if the clustering algorithm was precise by checking the most important cluster of an high specialized conference, like GIS, and also quantify the entity of a class relevance in a conference. The class relevance is the number of graphs with at least one author belonging to it ($n_{c,1}$) on

Id	Important Keywords	# Authors	# Keywords	GIS R_c(Stat)	GIS R_c(Mining)	KDD R_c(Stat)	KDD R_c(Mining)
9	spatial, spectrum, space, map	6358	1682	3.3*	0.36*	0.03	0.05
10	pattern, fast, novel, gene	3209	3663	0.05	0.04	3.63*	0.29*
15	digit, multimedia, metadata, ip	4278	2510	0.19**	0.07	0.05	0.05
114	relat, access, multimedia, spatial	3110	3548	0.07	0.11**	0.76**	0.19**

Table I
CLUSTERS STATISTICS.

the number of total graphs in the graph dataset N:

$$R_c = \frac{n_{c,1}}{N}$$

which takes values from 0 to 1.

Our results are shown in Table 1, columns GIS and KDD R_c Mining, where the most important class is highlighted with a mark and the second one with two marks. Looking at the different rations between the two most important classes we can identify the high specialized conferences (with one very dominant class) and the conferences that collects many different kinds of abilities.

The global approach is slightly different. We consider two aspects of a class in a conference network. A class is important if the authors belonging to it have an high local cluster value (in other words many other authors want to collaborate with them because they need their competence) and, of course, if there are many authors belonging to it.

The first part, the local cluster of a node, can be computed with the following criteria. Let *deg(v)* denotes the degree of vertex *v*, $CC_2(v)$ the local 2-neighborhood clustering of vertex *v* (the ratio between the number of lines among vertices in its 1-neighborhood and the number of lines among vertices in its 1 and 2-neighborhood, see Section 2.1) and *MaxDeg* maximum degree of vertices in the network. The clustering of the 2-neighborhood of vertex *v* can be normalized as follows:

$$CC_2^{'}(v) = \frac{deg(v)}{MaxDeg} CC_2(v).$$

We use the normalized version to compute the class relevance. For each class *c*, we sum the $CC_2^{'}(v)$ for each vertex *v* belonging to *c*. We then normalize this result with the ratio between the number of vertices belonging to *c* (n_c) and the total number of vertices (*N*).

$$R_c = \frac{n_c}{N} \sum_{v}^{v \in c} CC_2^{'}(v).$$

This parameter spans from 0 (the class is irrilevant for the conference) to $+\infty$ (its authors have an high clustering coefficient and are the most present in the conference).

This parameter confirms the previous mining approach, see Table 1, columns GIS and KDD R_c Stats. The very high ratio difference came from the high weight of clustering in this parameter. So the two levels of analysis confirm each others and mining techniques can be used to support the statistical analysis.

V. CONCLUSION

In this paper we have introduced a framework for bibliography analysis. We have showed the potential impact of combining global analysis tasks and local pattern mining approaches on a concrete (though preliminary) application to a well-known computer science digital bibliography. Both statistical parameters and mining process can evolve in novel analysis. It might be possible to define some global analysis that enable a direct confrontation of different research groups, in order to understand their impact rank or composition. Besides, the mining process should work not only on classes of research domains, but also on the single researcher, in order to refine the grain of our analysis. This obviously require the design of new scalable algorithms that are able to mine huge graphs, and exploit constraints other than the minimum support we used in this work. Finally, it is worth investigating how to embed our approach in a graph OLAP framework such as the one recently introduced in [3]. This may enrich the set of analysis provided, and also provide an efficient instrument to make these analysis quick and easy for everyone.

REFERENCES

[1] R. Agrawal, T. Imielinski and A. Swami, *Mining association rules between sets of items in large databases*, SIGMOD, 1993.

[2] A.-L. Barabasi and R. Albert, *Emergence of scaling in random networks*, Science 286, 1999.

[3] C. Chen, X. Yan, F. Zhu, J. Han and P. S. Yu, *Graph OLAP: Towards Online Analytical Processing on Graphs*, ICDM, 2008.

[4] I. S. Dhillon, S. Mallela and D. S. Modha, *Information-Theoretic Co-clustering*, KDD, 2003.

[5] M. Kuramochi and G. Karypis, *An efficient algorithm for discovering frequent subgraphs*, TKDE, 2004.

[6] J. Leskovec, K. Lang and A. Dasgupta, *Microscopic Evolution of Social Networks*, KDD, 2008.

[7] M. E. J. Newman, *The structure and function of complex networks*, SIAM Review 45, 2003.

[8] S. Nicholson, *The basis for bibliomining: frameworks for bringing together usage-based data mining and bibliometrics through data warehousing in digital library services*, IPM, 2006.

[9] S. Papadimitriou, J. Sun, C. Faloutsos and P. S. Yu, *Hierarchical, Parameter-Free Community Discovery*, ECML PKDD, 2008.

[10] X. Yan and J. Han, *gSpan: Graph-Based Substructure Pattern Mining*, ICDM, 2002.

[11] D.J. Watts and S. H. Strogatz, *Collective dynamics of 'small-world' networks*, Nature 393, 1998.