Discovering Popular Co-Cited Communities in Blogspaces

Meng-Fen Chiang and Wen-Chih Peng[*]
Department of Computer Science
National Chiao Tung University
Hsinchu, Taiwan, ROC
E-mail:{wcpeng@cs.nctu.edu.tw,ankechiang@gmail.com}

Abstract

In a blogspace, citation behaviors reflect interests of bloggers. To fully get insight into the latent information in a blogspace, in this paper, we intend to mine popular co-cited communities consisting of core sets and follower sets. In such a co-cited community, bloggers in the core set are frequently cited by bloggers in the follower set and the co-citation behaviors among bloggers are very intensive. Through co-citations, not only the popular core-set nodes but also the followers can be discovered. As such, one could effectively obtain the trends of discussion among bloggers. Explicitly, two kinds of co-cited communities are exploited: perfect co-cited community and approximate co-cited community. Given a blogspace, we first transform this blogspace into a transaction database. Then, by exploring frequent closed itemset mining, we are able to discover perfect co-cited communities. Then, a greedy algorithm is proposed to derive approximate co-cited communities. To evaluate our community structures mined, we conduct extensive experiments on deli.icio.us dataset. The experimental results demonstrate the effectiveness of our proposed framework.

***Keywords**: Community extraction, co-cited community, Web 2.0.*

1. Introduction

With the fast growth in Web 2.0 services, such as blogspaces and bookmark sharing services (e.g., del.icio.us), mining community structures from these Web 2.0 services becomes flourishing. Community discovering in such Web services can be treated as a problem of social network analysis. A social network is usually modeled as a graph, where each vertex represents an individual and each edge between two related vertices indicates certain relationships. The context of relationships varies according to the features of Web 2.0 services. For example, in a blogspace, the relationship among individuals is the citation behavior. Given a set of individuals and their corresponding relationships, the goal of community discovery is to derive a set of communities in which community members share common interests. Intuitively, the common interests are represented as a set of keywords. However, discovering keywords for common interests are not easy and is hard to justify whether these keywords could fully reflect common interests of communities.

In this paper, we intend to extract co-cited communities that consist of a set of initiators (i.e., the core set) and a set of affiliated followers (i.e., the follower set). Figure 1 shows an example of co-cited community structure, where the core set contains blogger 3 and blogger 4 and the members in the follower set are blogger 1, 2, 5 and 6. Thus, a common interest is viewed as a set of nodes (i.e., the core set) instead of keywords. Moreover, in Web 2.0 services, an individual is not only a content-producer (i.e., author of some blogs) but also a content-consumer (i.e., reader of other blogger's content). Thus, in our proposed community model, an individual may belong to multiple communities as roles of either an initiator or a follower. Mining such popular co-cited communities is very important in that one could easily get insight to the hot news and trends discussed among bloggers. An example of existing Web 2.0 services only provides the exploration of the top-k most popular blogs. However, this only provides a view formed by an individual, not a latent view behind a group of related individuals. On the contrary, through investigating co-citations, we can reveal the latent view shared by a group of individuals. In this way, not only the popular content-providers(core-set nodes) but also the content-consumers(follower-set nodes) can be discovered. Consequently, one can easily capture the latent trends in a blogspace.

Given a blogspace, according to the co-citation behaviors of individuals, we first derive the "link-to" transaction database, in which each transaction represents a link-to re-

[*]The corresponding author of this paper.

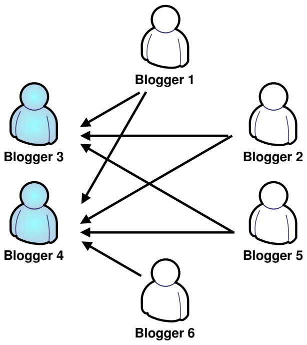
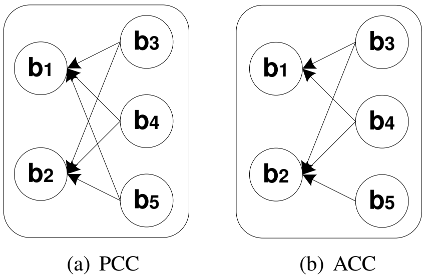


Figure 1. An example of co-cited community

lationship. Specifically, a transaction with its identification as b_i has a set of bloggers (e.g., b_j), which demonstrates a relationship that b_j is cited by b_i. Given a link-to transaction database, we utilize frequent closed itemset mining algorithm to derive the perfect co-citation communities (referred to as PCCs). In a PCC, the members in the core set are completely co-cited by the members in the follower set. Generally speaking, the co-citation behaviors in a perfect co-cited community is intensive and complete. It is possible that one may derive many perfect co-cited communities of smaller size in numbers of nodes, which cannot be efficiently utilized. This calls for mining approximate co-cited communities (referred to as ACCs). To discover approximate co-cited communities, we propose a greedy algorithm to iteratively merge co-cited communities. Specifically, a measurement (referred to as $incoherence$) is formulated to evaluate wether two communities are suitable for merging or not. The incoherence between co-cited communities capture the similarity of their members in each co-cited communities, the co-citation behaviors among two communities, and the popularity of a new core set merged from the core sets of two given co-cited communities. To evaluate our community structures mined, we conduct extensive experiments on deli.icio.us dataset. The experimental results demonstrate the effectiveness of our proposed framework.

A significant amount of research efforts have elaborated on mining communities in blogspaces [4][6][8][1][7][2][3]. The authors in [2] quantized a community in the context of sense of community. The authors in [4] viewed a community structure as a dense subgraph in blogspace and proposed a two-phases approach to identify dense subgraphs. By exploring the mutual awareness among bloggers, the authors in [6] model communities in terms of mutual-awareness and based on which a ranking-based algorithm is developed. The authors in [1] considered each community as a linear combination of subgraphs over time. In addition, the authors in [5] proposed a framework to generate overlapped communities, in which each blogger may belong to multiple communities. In this paper, through co-cited communities mined, the common interests are easily presented. By exploring frequent closed itemset mining and the formulation of incoherence, two kinds of popular co-cited communities are discovered. These features distinguish our study from others.



(a) PCC (b) ACC

Figure 2. Examples of co-cited communities.

The rest of the paper is organized as follows. The preliminaries are given in Section 2. A framework of mining co-cited communities is described in Section 3. Section 4 devotes to experimental results. This paper concludes with Section 5.

2. Preliminary

Without loss of generality, a blogspace is modeled as a graph $G = (V, E)$, where each node in V corresponds to a blogger and each directed edge (u, v) in E corresponds to a relation that blogger u cites blogger v. Note that relationships among individuals are application dependent and some research efforts are able to identify various relationships (e.g., mutual awareness). In this study, assume that a blogspace (i.e., G=(V, E)) is given. The objective of this paper is to discover co-cited communities with their corresponding interests represented as the set of nodes (referred to as a *core set*). Clearly, in a co-cited community, nodes in the core set are frequently co-cited by other nodes (referred to as a *follower set*). Thus, a community is formulated on the basis of co-citation behaviors between the core set and the follower set. In light of core sets and follower sets, the community mined is a bi-partite graph in which the nodes in the core set are referred by the nodes in the follower set. The community is thus defined as follows:

Definition 1: Co-cited Community: A co-cited community is denoted as C=(V, E), where V is the set of nodes in the community and E represents the citation behavior (edges) among nodes. Nodes in a co-cited community is partitioned into a core set and a follower set. A set of nodes is referred to as a *core set* (respectively, *follower set*), denoted as V_c (respectively, V_f), if $V_c \cap V_f = \phi$ and each node $v \in V_c$ is cited by nodes in V_f.

According to the definition of co-cited community, two kinds of community structures will be discovered in our paper. For example, Figure 2(a) shows a perfect co-cited community, where nodes in the core set: $\{b_1,b_2\}$ are cited by all nodes in its follower set :$\{b_3,b_4,b_5\}$. Such a tightly-coupled co-cited community is called a *perf*ect *co-cited community (PCC)*, which is a complete bi-partite graph with its vertex set represented by a union of V_c and V_f and the number of edges equals to $|V_c| \times |V_f|$. A perfect co-cited community is viewed as a community in which the nodes in the follower set share the interests presented by a core set. However, the tightly-coupled community, PCC, requires nodes in the follower set to refer to all nodes in the core set. Consider an example in Figure 2(b), where not all nodes in the follower set cite all nodes in the core set. Such a loosely-coupled community structure is called approximated co-cited community. An *approximate co-cited community (ACC)* is equivalent to an incomplete bi-partite graph, where nodes in the follower set do not refer to all nodes in the core set.

The proposed framework of mining co-cited communities point two insights: the members in a communities and the common interests associated with communities. For example, bloggers b_1, b_2 and b_6 in Figure 3 are viewed as similar because they both link to blogger b_3 and b_4. In this scenario, bloggers b_1, b_2 and b_6 share the interests of $\{b_2,b_6\}$. We borrow the concept of co-citation in the publication database and hence define the common interest of a community as a core set. In Figure 3, bloggers $\{b_1,b_2,b_6\}$ sharing the interests of $\{b_3,b_4\}$ jointly form a co-cited community. In Figure 3, a co-cited community $\{b_1,b_2,b_3,b_4,b_6\}$ is represented as a complete bi-partite graph in which $\{b_3,b_4\}$, a core set, is co-cited by each node in the follower set, $\{b_1,b_2,b_6\}$.

By exploiting citation behaviors among bloggers, we can avoid the abuse of keywords or semantic ontology for identifying the common interests associated with a community. On the other hand, since a blogspace contains a huge amount of bloggers, discovering all communities from a blogspace results in a considerable amount of co-cited communities. Thus, in this paper, we propose a greedy algorithm to derive popular co-cited communities whose members in the core set are approximately referred by the followers. Mining such popular co-citation communities are important and enables the investigation of the hot stories discussed in these communities.

3. Mining Popular Co-Cited Communities

Given a blogspace, we first transform the blogspace into a transaction database. After that, by exploring frequent closed itemset mining algorithms, we are able to discover perfect co-cited communities. Finally, we propose a greedy algorithm to derive approximate co-cited communities.

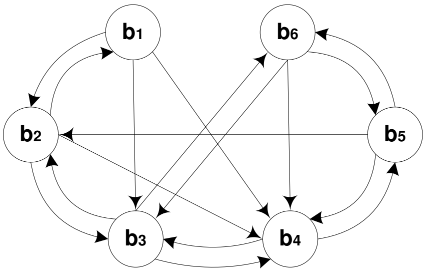


Figure 3. An example of blogspace topology.

3.1. Mining Perfect Co-cited Community

In order to identify core sets of perfect co-cited communities, we first transform a blogsapce into a "link-to" transaction databases. Explicitly, each blogger in a blogspace corresponds to a transaction, denoted as a tuple $< b_i, X >$, where b_i denotes blogger b_i and X refers to bloggers that are linked by b_i. X is also called an itemset which consists of items (i.e., bloggers). As prior works [9], *k-itemset* is referred to the itemset with k distinct items. Figure 3 illustrates an example of blogspace where each node in the graph represents a blogger and each arrow from b_i to b_j indicates an aggregated citation from b_i to b_j. Table 1 demonstrates the corresponding "link to" transaction database of the blogspace in Figure 3. For instance, the first transaction in Table 1 reflects the citation behavior of b_1, where $\{b_2,b_3,b_4\}$ are the bloggers cited by b_1. As pointed out early, in this paper, we intend to discover popular co-cited communities with their corresponding core sets and follower sets. Intuitively, the popularity of core-sets is determined by the number of transactions that contain the core set. To ignore redundant information, we use frequent closed itemsets to represent frequent itemsets in the database. Thus, given a transaction database transformed from a blogsapce, by exploring frequent closed itemset mining algorithms, core sets are exactly those frequent closed itemsets with supports no less than a minimum support (denoted as min_sup). On the other hand, the follower set of a frequent closed itemset is composed by the bloggers whose corresponding transactions contain the itemset. For example, in Table 1 itemset $\{b_2, b_4\}$ with support of 3 appears in transactions b_1, b_3, and b_5. If min_sup is set to 2, $\{b_2, b_4\}$ is a frequent itemset. Since there is no supper set of $\{b_2, b_4\}$ with identical support with $\{b_2, b_4\}$, the itemset $\{b_2, b_4\}$ is a frequent closed itemset and serves as a core set, whereas the follower set is $\{b_1, b_3, b_5\}$.

By exploring frequent closed itemset, we are able to derive perfect co-cited communities in which the core-set is popular and the members in the core set are all cited together by the members in the follower set. Generally speaking, the co-citation behaviors in a perfect co-cited community is intensive and complete. Given a blogspace, one may derive many perfect co-cited communities with smaller sizes in numbers of nodes, which is not practical of use. This calls for the design of discovering loosely co-cited communities (i.e., approximate co-cited communities).

3.2. Mining Approximate Co-cited Community

In order to discover approximate co-cited communities, we propose a greedy algorithm to merge co-cited communities. Specifically, we first derive the measurement (referred to as $incoherence$) of merging co-cited communities. The incoherence between co-cited communities (e.g., $C_i, andC_j$) should capture the similarity of their members in each co-cited community (denoted as α_{ij}), the co-citation behaviors among two communities (denoted as β_{ij}), and the popularity of the new core set merged from two core sets of co-cited communities given (denoted as γ_{ij}). The core-set and the follower set in a community C_i is expressed by V_c^i and V_f^i, respectively. Thus, we have the following formula:

$$Incoherence_{ij} = 1 - (w_1\alpha_{ij} + w_2\beta_{ij} + w_3\gamma_{ij})$$

, where w_i is the weight for each factor considered and $\Sigma_i w_i = 1$.

To measure the similarity of two co-cited communities (e.g., C_i and C_j) in terms of community memberships, α_{ij} should consider the overlaps in their core-sets and follower sets. As such, α_{ij} is formulated as follows:

$$\alpha_{ij} = \tfrac{1}{2}\left(\tfrac{|V_c^i \cap V_c^j|}{|V_c^i \cup V_c^j|} + \tfrac{|V_f^i \cap V_f^j|}{|V_f^i \cup V_f^j| - |V_c^i \cap V_c^j|}\right)$$

The overlap of their core-sets is measured by the ratio between their common members in their core-sets over the union set of their core-sets. On the other hand, the overlap of their follower sets is measured by the ratio of the intersection set of the follower sets over the union of follower sets in the two co-cited communities. Given two identical co-cited communities, α_{ij} equals to 1. On the contrary, if two co-cited communities do not have any common nodes in the core sets and the follower sets, α_{ij} equals to 0. The higher overlaps between two co-cited communities, the larger similarity in terms of their community membership.

Table 1. An example of "link to" transaction database transformed from Figure 3

Tid	Items
b_1	b_2,b_3,b_4
b_2	b_1,b_3,b_4
b_3	b_2,b_4,b_6
b_4	b_3,b_5
b_5	b_2,b_4,b_6
b_6	b_3,b_4,b_5

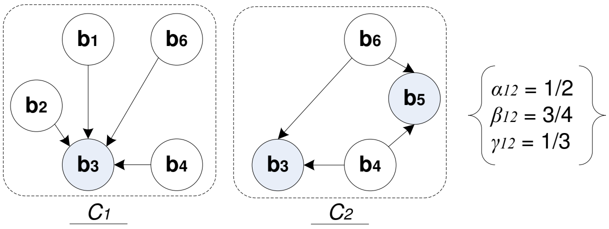


Figure 4. An example of incoherence calculation.

Figure 4 illustrates an example of deriving incoherence value between two co-cited communities (i.e., C_1 and C_2). In Figure 4, the similarity in membership between C_1 and C_2 is as follows:

$$\tfrac{1}{2}\left(\tfrac{|\{\mathrm{b}_3\}|}{|\{\mathrm{b}_3,\mathrm{b}_5\}|} + \tfrac{|\{\mathrm{b}_4,\,\mathrm{b}_6\}|}{|\{\mathrm{b}_1,\,\mathrm{b}_2,\mathrm{b}_4,\,\mathrm{b}_6\}|}\right) = \tfrac{1}{2}.$$

β_{ij} is used to reflect the co-citation behaviors among the core sets and the follower sets of two co-cited communities. Given a blogspace graph, the number of edges between two vertex sets (i.e., V_i, and V_j) is denoted as E($V_i, andV_j$). Thus, one could derive the number of co-citation behaviors among two co-cited communities (i.e., $E(V_c^i \cup V_c^j, V_f^i \cup V_f^j)$). Without loss of generality, the maximal number of co-citation behaviors between two co-cited communities is $|V_c^i \cup V_c^j| * |V_f^i \cup V_f^j|$, which is in fact the complete bi-partite graph. Thus, we can formulate β_{ij} as follows:

$$\beta_{ij} = \frac{E(V_c^i \cup V_c^j, V_f^i \cup V_f^j)}{|V_c^i \cup V_c^j| * |V_f^i \cup V_f^j|}$$

It can be verified that in Figure 4, the ratio of co-citations and the maximal number of co-citations in complete bi-graph between $|V_c^i \cup V_c^j|$ and $|V_f^i \cup V_f^j|$ is derived as follows:

$$\tfrac{3}{4} = \tfrac{|\{(\mathrm{b}_1,\mathrm{b}_3),(\mathrm{b}_2,\mathrm{b}_3),(\mathrm{b}_4,\mathrm{b}_3),(\mathrm{b}_6,\mathrm{b}_3),(\mathrm{b}_4,\mathrm{b}_5),(\mathrm{b}_6,\mathrm{b}_5)\}|}{2 \times 4}.$$

γ_{ij} captures the support of a union set of two core-sets (i.e., $|V_c^i \cup V_c^j|$). The support of a union set of two core-sets is denoted as $\mathrm{sup}_{|V_c^i \cup V_c^j|}$. Then, we have γ_{ij} as follows:

$$\gamma_{ij} = \frac{sup_{|V_c^i \cup V_c^j|}}{|T|}$$

Based on the above formulas, we are able to derive incoherence values among co-cited communities. We propose a greedy algorithm to generate approximate co-cited communities. Specifically, in the beginning, the incoherence matrix is derived by calculating a set of perfect co-cited communities mined. Then, by selecting the minimal incoherence

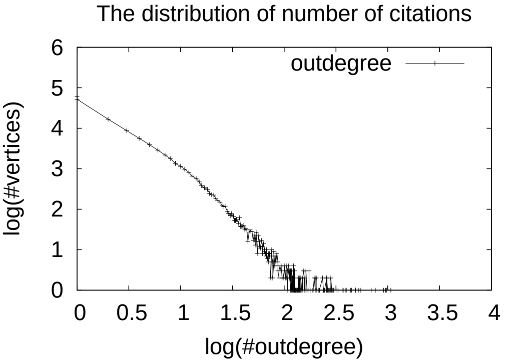


Figure 5. The distribution of out-degrees in the del.icio.us.

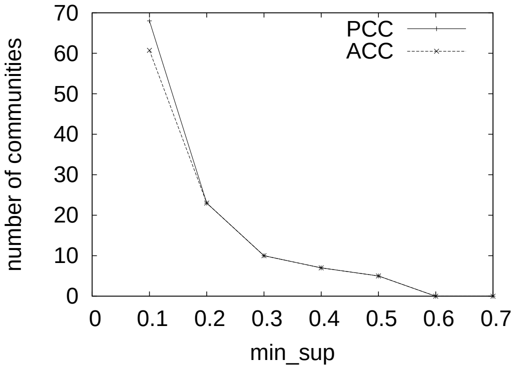


Figure 6. Number of co-cited communities with min_sup varied.

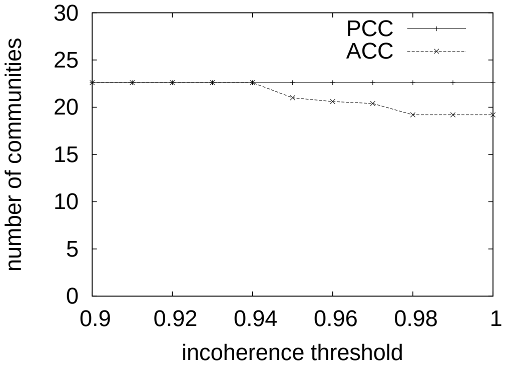


Figure 7. Number of co-cited communities with $incoherence$ varied.

value, two co-cited communities are merged as one new community. Then, following the same operation, co-cited communities are iteratively merged until no pair of co-cited communities remain with their incoherence values smaller than the threshold given.

Algorithm 1 A greedy algorithm for mining approximate co-cited communities

```
Input: Set of $PCCs$, an $incoherence$ threshold $\delta$.
Output: Set of $ACCs$.
Compute the $incoherence$ matrix;
while minimum $incoherence \leq \delta$ and unstable do
    Merge the two co-cited communities with the minimum $incoherence$ value;
    Update the $incoherence$ matrix
end while
return Set of $ACCs$;
```

4. Experimental Results

To evaluate our proposed algorithm for community extractions, we conduct extensive experiments on the real dataset, del.icio.us, the most popular bookmark sharing site.

4.1. del.icio.us dataset

The social network of del.icio.us is constructed based on the citations (i.e., fan links and network links) among users. We built the social network along with users' citations by searching popular tags in del.icio.us and retrieving active users for each popular tag. We refer to these active users as the seeds. Then, the entries related to these seeds via either network or fan links are continuously crawled until more than 100,000 users are collected. For each user, we retrieved recent 1,000 bookmarks in which user names, tags, URLs and bookmark time can be referred. The graph we used for experiments contains about 165,002 users and 431,721 citations. Each node represents a user and each edge from node u to v represents a citation that the user u is interested in user v's bookmarks. Given a social network from del.icio.us, we transform the social network as a ”link to” transaction database. Figure 5 illustrates the distribution of number of citations(out-degree) for all users in our dataset. Most users have lower than 100 citations. To make the experiment results more realistic and convincing, we conduct experiments on a real data set. The default parameters are set as follows: min_sup=0.1% - 0.7%, w_1=0.8, w_2=0.1, w_3=0.1, and δ=0.9 - 1.

4.2 The impact of minimum support and the incoherence threshold

In this section, we investigate the impact of min_sup and the $incoherence$ threshold. In Figure 6, the number of communities for both PCC and ACC decreases as min_sup increases. With larger min_sup, the number of frequent closed itemsets decreases. Hence, the number of co-cited communities is reduced. Note that the number of approximate co-cited communities drastically de-

Table 2. Examples of co-cited communities discovered from del.icio.us.

support count	top-15 tags
134	video, autobookmarked, youtube webdev, awesome, photography, flickr funny, via:reddit, music, osx, commerce humor, videogames, via:mefi
46	music, interviews, politics, reviews pop, opinion, uk, us, fashion, culture humor, london, photos, celebs, media videos.
60	make, 43folers, quick_post, press mbwideas, mac, apple, osx, howto secondlife, music, art, diy, eletronics wordpress

creases, showing the compactness of approximate co-cited communities. Figure 7 shows the impact of *incoherence* threshold to the number of communities. Note that with a larger *incoherence*, more co-cited communities are merged. Therefore, the number of communities significantly decreases when the *incoherence* threshold is larger.

4.3 Community Extraction

Under the parameter settings with *incohernce* threshold 0.9, w_1=0.8, w_2=0.1, w_3=0.1, and min_sup=0.1%, three examples of approximate co-cited communities mined from del.icio.us dataset are demonstrated in Table 2. Instead of showing bloggers in a community, the common interests are represented as the top-15 tags used by the members in the core set of each corresponding community. Some communities demonstrate cohesive topics while some are more diverge which shows the advantage of our community model. Explicitly, mining communities via co-citation is more effective compared to keyword-based models. When the latent story behind a group of individuals diverges into few topics, mining communities according to co-citation behaviors is required. For example, the three tags ”video”, ”music” and ”humor” are frequently used in the first two co-cited communities. Our work is able to find co-cited communities with tags frequently discussed but differ in semantic context. Moreover, discovered co-cited communities with some diverse interests is allowable. For example, both the first two co-cited communities are interested in ”video”, ”music” and ”humor. However, the concept as a whole demonstrates different semantic context.

5. Conclusion

In this paper, given a blogspace, we proposed a framework of mining popular co-cited communities consisting of core sets and follower sets. In a popular co-cited community, bloggers in the core set are very frequently cited by bloggers in the follower set and the co-citation behaviors among bloggers are very intensive. Two kinds of co-cited communities are exploited: perfect co-cited community and approximate co-cited community. Given a blogspace, we first transformed a given blogspace into a transaction database. Then, by exploring frequent closed itemset mining, we are able to discover perfect co-cited community. Finally, we proposed a greedy algorithm to derive approximate co-cited communities. To evaluate our community structures mined, we conducted extensive experiments on deli.icio.us dataset. The experimental results demonstrate the effectiveness of our proposed framework.

References

[1] Y. Chi, S. Zhu, X. Song, J. Tatemura, and B. L. Tseng. Structural and temporal analysis of the blogosphere through community factorization. In *Proc. of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 163–172, 2007.

[2] A. Chin and M. H. Chignell. A social hypertext model for finding community in blogs. In *Proc. of the 17th ACM Conference on Hypertext and Hypermedia*, pages 11–22, 2006.

[3] K. Ishida. Extracting latent weblog communities: A partitioning algorithm for bipartite graphs. In *Proc. of the 2nd Annual Workshop on the Weblogging Ecosystem: Aggregation, Analysis and Dynamics*, 2005.

[4] R. Kumar, J. Novak, P. Raghavan, and A. Tomkins. On the bursty evolution of blogspace. In *Proc. of the 12th International World Wide Web Conference*, pages 568–576, 2003.

[5] X. Li, B. Liu, and P. S. Yu. Discovering overlapping communities of named entities. In *Proc. of the 10th European Conference on Principles and Practice of Knowledge Discovery in Databases*, pages 593–600, 2006.

[6] Y. R. Lin, H. Sundaram, Y. Chi, J. Tatemura, and B. Tseng. Discovery of blog communities based on mutual awareness. In *Proc. of the 3rd Annual Workshop on the Weblogging Ecosystem: Aggregation, Analysis and Dynaics.*, pages 487–499, 2006.

[7] M. McGlohon, J. Leskovec, C. Faloutsos, M. Hurst, and N. Glance. Finding patterns in blog shapes and blog evolution. In *Proc. of the 1st International Conference on Weblog and Social Media*, 2007.

[8] B. Tseng, J. Tatemura, and Y. Wu. Tomographic clustering to visualize blog communities as mountain views. In *Proc. of the 2rd Annual Workshop on the Weblogging Ecosystem: Aggregation, Analysis and Dynaics.*, 2005.

[9] S. B. Yahia, T. Hamrouni, and E. M. Nguifo. Frequent closed itemset based algorithms: a thorough structural and analytical survey. *SIGKDD Explorations*, 8(1):93–104, 2006.