データベースシステム 125－66
（２００１．７．１９）

2部グラフ抽出に基づく関連コミュニティ発見の試み

P. Krishna Reddy、喜連川 優
東京大学生産技術研究所
〒153-8505 東京都目黒区駒場 4-6-1
{reddy, kitsure}@tkl.iis.u-tokyo.ac.jp

本論文では、大規模なWebページの集合に対してハイパーリンク解析を行うことによって、関連するコミュニティを抽出する手法を提案する。我々は、複数のコミュニティが共通の話題を持つとき、それらは関連していると見做す。本手法は、稠密な2部グラフ (dense bipartite graph (DBG)) による抽象化を2つの目的のために用いている。まず、DBGをコミュニティと見做して、Webページの集合から全てのコミュニティを抽出する。次に、コミュニティをノードとするDBGを用いて、関連するコミュニティを抽出する。我々は、170万ページ、2150万リンクを含む、10GBのTREC(Text REtrieval Conference) データセットを用いて実験を行った。その結果から、本手法が意味のあるコミュニティと、それらの間の関連を抽出できることを示す。

An Approach to Find Related Communities Based on Bipartite Graphs

P.Krishna Reddy and Masaru Kitsuregawa
Institute of Industrial Science, The University of Tokyo
4-6-1, Komaba, Meguro-ku, Tokyo- 1538505, Japan
{reddy, kitsure}@tkl.iis.u-tokyo.ac.jp

Abstract

In this paper we investigate the problem of extracting related community information from a large collection of Web-pages by performing hyperlink analysis. We consider a group of communities (say related community set) related if they have common interests on some topic. In the proposed approach, we employ dense bipartite graph (DBG) abstraction for two purposes. From the given page collection, we first extract all the communities by mathematically abstracting the community as a DBG over a set of pages. Next, our approach extracts related communities among the communities by abstracting related community set as a DBG over set of communities. We report experimental results on 10 GB TREC (Text REtrieval Conference) data collection that contains 1.7 million pages and 21.5 million links. The results demonstrate that the proposed approach extracts related community structures.

Index terms Web mining; Communities; trawling; Link analysis

1 Introduction

The Internet (or Web) has rapidly grown into being an integral element of infrastructure of the society. One of the most powerful socializing aspects of the Web is its ability to connect groups of like-minded people independent of geography or time zones. The Web lets people join communities across the globe by providing opportunity to form associations among the people. In the Web environment one is limited by only his/her interests. The Web has several thousand well-known, explicitly defined communities -- groups of individual users who share a common interest. Most of these communities manifest themselves as news groups, Web-rings, or as resources collections in directories such as Yahoo and Infoseek, or home pages of Geocities.

In the context of the Web, we consider community as a group of content creators that manifests itself as a set of interlinked pages. In [15], we have discussed a method to detect communities by abstracting a community as a dense bipartite graph (DBG). In this paper we investigate the problem of extracting related community information from a large collection of Web-pages by performing hyperlink analysis. Our interest is to find related communities among the extracted communities. We consider a group of communities related (called related community set), if they have common interests on some topic. In the proposed approach, we employ DBG abstraction for two purposes. From the given page collection, we first extract all the communities by mathematically abstracting the community as a DBG over a set of pages. Next, our approach extracts related communities among the extracted communities, by abstracting related community set as a DBG over set of communities. We report experimental results on 10 GB TREC (Text REtrieval Conference) data collection that contains 1.7 million pages and 21.5 million links. The results demonstrate that the proposed approach extracts meaningful community as well as related community structures.

The rest of the paper is organized as follows. In the next section, we review related research. In section 3, we discuss community abstraction through bipartite graphs. In section 4 we present community extraction algorithms. In section 5 we report experimental results conducted on 10GB TREC data. The last section consists of summary and future research.

2 Related work

We review the approaches proposed in the literature related to datamining and link analysis and, community

detection.

Data mining and link analysis

The data mining approach [1] focuses largely on finding association rules and other statistical correlation measures in a given data set. The notion of finding communities differs from the fact that, in our approach the relationship we exploit is co-citation whereas in data mining is performed based on the support and confidence.

One of the earlier uses of link structure is found in the analysis of social networks [17], where network properties such as cliques, centroids, and diameters are used to analyze the collective properties of interacting agents. The fields of citation analysis [10] and bibliometrics [22] also use citation links between works of literature to identify patterns in collections.

Most of the search engines perform both link as well as text analysis to increase the quality of search results. Based on link analysis many researchers proposed schemes [6, 7, 8, 5, 14, 13, 3] to find related information from the Web. In this paper we extend the concept of cocitation to the web environment to extract communities in the from a large collection of Web pages.

Community related research

In [11], communities have been analyzed which are found based on the topic supplied by the user by analyzing link topology using HITS (Hyper-link-Induced Topic Search) algorithm [13]. The HITS is one of the widely used algorithm in search engines to find authoritative resources in the Web that exploits connectivity information among the Web pages. The basic idea behind the community detection process using HITS is mutual reinforcement: good hubs point to good authorities; and good authorities are pointed by good hubs. HITS finds good authority pages given a collection of pages on same topic. Our motivation is to detect all the communities in a larger collection of pages that covers wide variety of topics.

Ravi Kumar et al. [16] proposed a trawling method to find potential communities by abstracting a core of the community as a group of pages that form a complete bipartite graph (CBG)(by considering web-page as a node and link as an edge between two nodes). Given a large collection of pages, the trawling algorithm extracts community cores by extracting all the potential CBGs. The proposed appraoch is different in that we use DBGs to extract and relate the potential communities.

In [9], given a set of crawled pages on some topic, the problem of detecting a community is abstracted to maximum flow /minimum cut framework, where as the source is composed of known members and the sink consist of well-known non-members. Given the set of pages on some topic, a community is defined as a set of web pages that link (in either direction) to more pages in the community than to the pages of outside community. The flow based approach can be used to guide the crawling of related pages.

In [4], an approach to find the related pages of a seed pages presented by specializing the HITS algorithm exploiting link weighting and order of links in a page. Companion first builds a subgraph of the Web near the seed, and extracts authorities and hubs in the graph using HITS. The authorities are returned as related pages. In [19] companion algorithm is extended to find related communities by exploiting the derivation relationships between pages.

In [15], we have proposed an algorithm to extract the members of the community by extracting the DBG patterns of potential communities. It has been shown that the DBG-based approach extracts significantly big community patterns as compared to the corresponding community patterns extracted by trawling approach.

The proposed approach differs from preceding approaches as we used DBG abstraction to extract the related communities.

3 Bipartite graphs and communities

We first explain terminology used in this paper. A page is referred by its URL, which also denotes a node in a bipartite graph. We refer a page and its URL interchangeably. If there is an hyper-link from page u to page v, we say u is a parent of v and v is a child of u. An hyper-link from one page to other page is considered as an edge between the corresponding nodes in the bipartite graph. For a page u, parent(u) is a set of all parent pages (nodes) of u and child(u) is a set of children pages of u.

The input to the community detection process is a large collection of pages, which is denoted by a set *page_set* (PS). The terms $targets$ (T) and $interests$ (I) denote the set of URLs which also denote two groups of nodes in a bipartite graph. Note that, $T \subset PS$ and $I \subset PS$. Also, the terms targets-count (tc) and interests-count (ic) denote the number of pages in T and I respectively.

3.1 Bipartite graphs

We first give definition of bipartite graph.

Definition 1 Bipartite graph (BG) *A bipartite graph BG(T,I) is a graph whose node-set can be partitioned into two non-empty sets T and I. Every directed edge of BG joins a node in T to a node in I.*

Note that BG is dense if many of possible edges between T and I exist. In a BG, the linkage denseness between the sets T and I is not specified. Here, we define a DBG which captures linkage denseness between the sets T and I.

Definition 2 Dense bipartite graph (DBG) *A DBG(T, I, α, β) is a BG(T,I), where (i) each node of T establishes an edge with at least α ($1 \leq \alpha \leq ic$) nodes of I, and (ii) at least β ($1 \leq \beta \leq ic$) nodes of T establish an edge with each node of I.*

Now we define the complete bipartite graph that contains all possible edges between the nodes of T and the nodes of I.

Definition 3 Complete bipartite graph (CBG) *A CBG(T,I) is a DBG(T,I, α, β), where $\alpha = ic$ and $\beta = tc$.*

It can be observed that in a DBG(T,I, p, q), both p and q specify the linkage denseness whereas in a CBG(p, q) same denote the number of nodes in T and I respectively. Figure 1 shows the difference between DBG(T,I, p, q) and CBG(p,q).

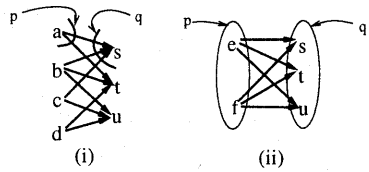

Figure 1. Graphs: (i) DBG(T,I, p, q) (ii) CBG(p,q)

3.2 Community and related communities

We consider a community as a collection of pages that form a linkage pattern equal to a DBG. Our definition is based on the following intuition: *Web communities are characterized by DBGs.* In the Web environment, page-creator (a person who creates the page) creates the page by putting the links to other pages of interest in isolation. Since a page-creator mostly puts the links to display his interests, we believe that if multiple pages are created with similar interests, at least few of them have common interests. Our intuition is that such a phenomena can be captured through a DBG abstraction.

A community phenomena can also be captured through a CBG abstraction[16]. A CBG abstraction extracts a small set of potential members to agree on some common interests. However, it is not possible to find large communities through CBG abstraction because page-creators put links in a page in an arbitrary manner. So it rarely happens that a page-creator puts links to all the pages of interest in particular domain.

Given a very large collection of pages, for each community there might exist few pages that could form a CBG. However, given the size of the Web it is not easy (impossible) to crawl a very large collection of Web pages. Collecting a very large collection of pages is a time consuming process. Also, for effective search, focused crawling is recommended that covers all the Web pages on few topics. In this situation, given a reasonably large collection of pages, there is no guarantee that each community formation is reflected as a CBG core. Because PS may not contain the potential pages to form a CBG.

Normally, each member in a community shares interests with few other members. Therefore, as compared to CBG abstraction, abstraction of a community pattern through a DBG matches well with real community patterns. In general community can be viewed as a macro-phenomena created by complex relationships exhibited by corresponding members. At micro-level, each member establishes relationships with few other members of the same community. Integration of all members and their relationships exhibit a community phenomena. In the context of Web, a DBG abstraction enables extraction of a community by integrating such micro-level relationships.

Note that not all DBGs are of interest in the context of communities. Now we give the community definition by fixing threshold values for both α and β in a DBG.

Definition 4 Community. *The set T contains the members of the community if there exist a dense bipartite graph DBG(T, I, α, β), where $\alpha \geq \alpha_t$ and $\beta \geq \beta_t$, where α_t and β_t are nonzero integer values which represent threshold.*

It can be observed that we have defined the community by keeping the number of nodes in both T and I unspecified. We specify only linkage denseness with both α and β for a given PS. The values of α_t and β_t are fixed with a feedback after examining the potential correspondence with the real community patterns. These values are fixed such that by extracting such patterns we should establish a close relationship among the members of T.

Related Community Set

A community is a set of related pages. We have considered DBG over a set of pages as an abstraction of a community. By extending same notion, we consider DBG over a set of communities as a abstraction of related communities. Let the input contains set of communities with its members. Here T contains identifiers of the communities and I contains the members of community (URL pages). Then, the related community set is defined as follows.

Definition 5 Related community set (RCS). *The set T contains the members of RCS if there exist a dense bipartite graph DBG(T, I, α, β), where $\alpha \geq \alpha_t$ and $\beta \geq \beta_t$, where α_t and β_t are nonzero integer values which represent threshold values.*

4 Cocite and relax_cocite relationsships

Web-page creators keep links in a page for different reasons. For example, one may put a link to other page to direct the relevant information, to promote the target page or as an index pointer. In this paper we consider the existence of a link from one page to another page as a display of interest by the former on the later page.

In the web environment, web pages can be grouped based on the type of relationship (association, pattern, or criteria) defined among pages. For example, in an information retrieval environment, the documents are searched based the notion of syntactic relationship that is measured based on the existence of number of common keywords. Similarly, one could define any type of relationship among the web pages and investigate the efficiency through experiments. In the Web environment researchers have defined different types of relationships to group web pages. Existence of a link, cocitation, coupling, number of paths between web pages are some examples of relationships.

In this paper we have investigated finding communities based on the $relax\_cocite$ relationship which is a relaxed version of the *cocitation* relationship. We first discuss about the *cocite* relationship to search related information in the Web. Next. after explaining $relax\_cocite$, we present the proposed algorithm.

4.1 Cocite

The fields of citation analysis [10] and bibliometrics [22] also use citation links between works of literature to identify patterns in collections. Co-citation [18] and bibliographic coupling [12] are two of the more fundamental measures used to characterize the similarity between documents. The first measures the number of citations in common between two documents, while the second measures the number of documents that cite both of two documents under consideration.

Also, in the information retrieval literature, relationship between documents can be established with keywords that exist in both documents. Similarly, in a web environment as we have considered link as a display of interest on the target page, by dealing with only links we can establish association among pages based on the existence of common children (or URLs). That is, we can establish association among pages through number of common children. We call this relationship *cocite* as in bibliographical terms if two documents [18] refer a collection of common references, we say, they cocite[1] them. We formally define the *cocite* relationship in the context of Web environment as below. Figure 2(i) depicts the cocite relationship between pages p and q with $cocite\_factor$ = 3.

Definition 6 Cocite *Let u and v are pages. Then, cocite(u,v)=true, if* $child(u) \cap child(v) \geq couple\_factor$, *where* $cocite\_factor \geq 1$.

4.2 Relax_cocite

According to *cocite*, a set of pages is related, if there exist a set of common children. Even though *cocite* is defined to establish relationship between two documents, it could form association among multiple documents in the following way. We consider two pages u and v in the PS are related if both have common links at least equal to $cocite\_factor$. Similarly, n ($n \geq 2$) pages are related under *cocite* if these pages have common children at least equal to $cocite\_factor$. If a group of pages are related according to *cocite* relationship, these pages form an appropriate CBG.

However, to extract a DBG, we have to retrieve a collection of pages loosely related. So we relax the *cocite* relationship to find loosely related pages in the following manner. We allow pages u,v and w to to group if cocite(u,v) and cocite(v,w) are true. This modification enables relationship between a page and multiple pages taken together. That is, if a page could not form association with another page according to *cocite*, it does not imply that they are different. Even though a page fails to satisfy a certain minimum criteria page-wise, however, it could satisfy minimum criteria with multiple pages taken together. We define the corresponding new definition, $relax\_cocite$ as follows.

[1]Note that we consider two documents are related as per *cocite* if they cite a group of documents and as per *couple* if a group of documents cite them. In this paper, we propose community extraction algorithm based on the relaxed form of cocitation.

Definition 7 Relax_cocite. *Let T be the set of pages and u be the another page (*$u \notin comm\_set$*). For any page* $p \in T$, $relax\_cocite(u,p)$=*true if* $child(u) \cap child(T) \geq relax\_cocite\_factor$. *Here, child(T) is a set that contains all the children of T.*

It can be observed that for a new page u, as compared to cocite, $relax\_cocite$ increases the probability of association with p as child(T) is larger than child(p). Figure 2(ii) depicts the $relax\_cocite$ relationship among web pages x, y and z, with $relax\_cocite\_factor$ equal to 1.

However, note that for a given page, $relax\_cocite$ may gather pages that are semantically different from the starting page. However, after collecting a reasonable number of pages we employ effective pruning methods to extract a DBG pattern by pruning non-potential pages.

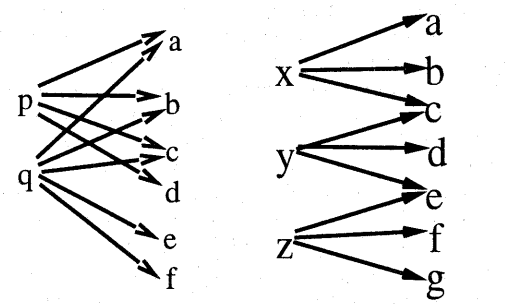

Figure 2. Depiction of *cocite* and $relax\_cocite$.

4.3 Proposed approach

We first present community extraction algorithm which extracts community structures from a large collection of web-pages. Next, we explain how related communities can be extracted using the same algorithm.

4.3.1 Community extraction

Given a large collection of pages, an algorithm to extract DBG structures consists of two steps: gathering related pages and DBG extraction. For each page, we gather related pages during gathering phase through $relax\_cocite$ relationship. We then apply the iterative pruning technique to extract a DBG(T, I, α, β). We now present corresponding routines.

1. **Gathering related pages**
   In this step for a given URL, p, we find T (set of URls). The default value for $relax\_cocite\_factor$ is 1. The variable $num\_iterations$ is an integer (> 0) variable. Set T = p.
   (a) While $num\_iterations \leq$ n ($n \geq 1$)
      i. At a fixed $relax\_cocite\_factor$ value, find all w's such that $relax\_cocite$(w, T) = true.

ii. T = w ∪ T.

(b) Output T.

2. **DBG extraction**
In this step the input contains T produced from the preceding step and the output contains a dense bipartite graph, DBG(T, I, α, β). Let $edge\_file$ be the set of elements $< p, q >$ where p is a parent (source) of child q (destination). Let, T1 and I1 be a set of tuples of the type $< URL, freq >$. Set T1, I1, and $edge\_file$ to ϕ.

(a) For each $p \in T$, insert the edge $< p, q >$ in $edge\_file$ if $q \in child(p)$.

(b) While $edge\_file$ is not converged repeat the following.

i. Sort the $edge\_file$ based on the source. Prepare T1 with $< source, freq >$. Remove $< p, q >$ from $edge\_file$ if $freq < \alpha$.

ii. Sort the $edge\_file$ based on the destination. Prepare I1 with $< q, freq >$. Remove $< p, q >$ from $edge\_file$ if $freq < \beta$.

(c) The resulting $edge\_file$ represents a DBG(T, I, α, β) where, T={ p | $< p, q > \in$ $edge\_file$ } and I= {q | $< p, q > \in$ $edge\_file$ }.

4.3.2 Related community set extraction

We now explain how the community extraction algorithm can be used to extract RCS. In this we input a collection of communities, instead of collection of pages. This step consists of forming a collection of communities and extraction of DBGs from these communities. The community extraction algorithm, gives a set of $edge\_files$ as output which represent the DBGs of corresponding communities. Here, from the $edge\_file$ we explain a procedure to extract community information which will be input to extract RCSs. The explanation of $edge\_file_i$ and $community_i$ is as follows.

- **Edge_file$_i$**. Let both r and s be URLs and i be an integer value. An $edge\_file_i$ consists of edges of the form $< r, s >$ where there is an hyperlink from r to s. Here, r is a member of community and i is a unique identifier of the community.

- **Community$_i$**. Let r and s be URLs and i be an integer value. Then, $community_i = \{r \mid < r, s > \in edge\_file_i\}$.

By giving communities as an input instead of webpages, the RCSs are extracted by using community extraction algorithm.

5 Experiment results

5.1 Preprocessing and link-file preparation

We report experimental results conducted on 10 GB TREC [21] (Text Retrieval Conference [20]) data collection. It contains 1.7 million web pages.

For a given page collection, link-file contains all the links of the form $< p, q >$ where $p \in parent(q)$. We prepare a link-file through the following steps (for details see [16]): extracting all the links, eliminating the duplicates and removing both popular and unpopular pages.

The pages are in the text format with html marking information. We have extracted links by ignoring all the text information. We then created a link-file for entire page collection in the following manner. We employed 32 bit fingerprint function to generate a fingerprint for each URL. Each page is converted into a set of edges of the form <source , destination>, where source represents the title URL and destination represents the other URL in the page. The total number of pages and edges comes to 1.7 million and 21.5 million respectively.

Next, we removed the possible duplicates by considering two pages as duplicates if they have a common sequence of links. We employed the algorithm proposed in [2] to remove the duplicates. We have selected shingle window size as four links. We kept at most three shingles per page. We have considered two pages as duplicates even one shingle is common between them. We found that considerable number of pages are duplicates. After the duplicate elimination, the total number of edges comes to 18 million.

Next we have removed edges derived from both extreme popular and unpopular pages. The popular pages are those which are highly referred in the Web such as WWW.yahoo.com. Also the unpopular pages are those which are least referred. We considered a page as popular if it has more than 50 parents (we have adopted this threshold from [16]). We considered a page as unpopular if it has less than two parents. After sorting the link-file based on the destination, those pages having number of parents greater than fifty and less than two are removed. Also, we removed pages with one child by considering that these do not contribute to community finding. So, after sorting based on the source, the links which have number of children less than two are removed. The above two steps are performed repetitively until the number of edges converge to a fixed value. After this step the number of pages and corresponding edges comes to 0.7 million and 6.5 million respectively.

This link-file is used to retrieve both parents and children of a given page during community extraction.

(α, β)	# of DBG(T, I, α, β)	(avg(T), avg(I))
(2,3)	110422	(36.21, 162.6)
(2,4)	81135	(36.98, 109.65)
(2,5)	61566	(36.15, 83.465)
(3,3)	90129	(32.86, 192)
(3,4)	59488	(32.26, 140.56)
(3,5)	40708	(30.17, 114.93)
(4,3)	66670	(34.29, 244.81)
(4,4)	49051	(27.75, 159.62)
(4,5)	32309	(24.97, 134.33)
(5,5)	28296	(21.07, 145.09)
(6,6)	17335	(19.03, 161.67)
(7,7)	10960	(18.97, 198.17)

Figure 3. Graph details: # of DBG(T,I, α, β) patterns, average # of pages in T and I.

5.2 Community extraction results

We first report the results during gathering phase. We then discuss community extraction using proposed approach. Next, we show some examples of real community patterns extracted using proposed approach from the TREC data collection.

In the gathering phase, it has been observed that with number of iterations beyond 1, the pages in T are found to be too loosely related. Since our aim is to find all communities, we extracted communities by restricting number of iterations to one. Among these pages, we extract DBG(T,I, α, β).

Figure 3 shows the number of DBG(T,I, α, β) patterns for all the pages that constitute link-file. The total number of pages that constitute link-file is around 0.7 million. For a DBG(T,I, α, β), the column ''(avg(T), avg(T))'' indicates average number of pages in T and I. (Note that these include duplicate communities.) In this, the node set T contains members of the community.

Community Examples

Here we provide three potential communities extracted from 10GB TREC data collection. The set $Targets$ represents the potential members of the community (corresponding topics are indicated in the brackets) and the set $Interests$ represents the potential children of the community. In the figures, an edge (x,y) is represented as an arrow from x to y.

All the graphs represent DBG(T, I, 3, 3); i.e., each member of T has at least 3 children in I and at least 3 members in T have one common child in I. The corresponding graphs are shown in Figure 4.

Example 1. Topic: Comedy

Targets

1. http://www.tnef.com/jim_carrey.html (Jim Carrey - (15 links) Actors)
2. http://www.comedyweb.co.uk/cwlinks.htm (Comedy Web Links Page)
3. http://www.starcreations.com/abstract/laughriot/lr-fam01.htm (LAUGH RIOT - FAMOUSLY FUNNY)
4. http://www9.yahoo.com/Business_and_Economy /Companies/Entertainment/Comedy/Comedians/Carrey_Jim/ (Yahoo! - Comedy:Comedians:Carrey, Jim)
5. http://www.scar.utoronto.ca/ 93kolmeg/starp.html (Personalities on Chog)
6. http://www.allny.com/comedy.html (New York Comedy Clubs)

Interests

1. http://q.continuum.net/ scout/jimpage.htm
2. http://www.halcyon.com/browner/
3. http://www.nd.edu/ jlaurie1/dmhome.html
4. http://www.cheech.com/
5. http://meer.net/ mtoy/steven_wright.html
6. http://www.en.com/users/bbulson/jim.html

Example 2. Topic: Environment and safty

Targets

1. http://www.saul.com/env/index.html (Saul, Ewing, Remick & Saul - 10: Environmental Law (PA, NJ, DE))
2. http://www.crystalcity.org/cfd/sitelinks.html (CFD links to other sites)
3. http://www.safetylink.com/ (Safety Link)
4. http://wwell.com/safety-resources/related-links.html (Safety Resources on the Web)
5. http://www.pixelmotion.ns.ca/WCB/links.html
6. http://www.mcaa.org/safety.htm (Safety & amp; Health)

Interests

1. http://atsdr1.atsdr.cdc.gov/toxfaq.html
2. http://www.ccohs.ca/
3. http://turva.me.tut.fi/ oshweb/
4. http://atsdr1.atsdr.cdc.gov/hazdat.html
5. http://www.wpi.edu/ fpe/nfpa.html
6. http://www.osha-slc.gov/

Example 3. Telecommmunications related community

Targets:

1. http://gatekeeper.angustel.com/links/l-mfrs.html (Telecom Resources: Manufacturers)
2. http://gemini.exmachina.com/links.shtml (Wireless Links)
3. http://millenniumtel.com/ref-voic.htm (Millennium Telecom:References)
4. http://www.buysmart.com/phonesys/phonesyslinks.html (BuyersZone: Phone systems)
5. http://www.commnow.com/links.htm (WirelessNOW Links Page)
6. http://eserver.sms.siemens.com/scotts/010.htm (http://www.smutking.com:80/)
7. http://www.searchemploy.com/research.html (Search & Employ)
8. http://www.electsource.com/elecoem.html (Electronics OEM's)

Interests

1. http://www.harris.com/
2. http://www.nb.rockwell.com/
3. http://www.cnmw.com/
4. http://www.mpr.ca/
5. http://www.brite.com/
6. http://www.pcsi.com/
7. http://www.ssi1.com/
8. http://www.mitel.com/
9. http://www.centigram.com/
10. http://www.adc.com/
11. http://www.dashops.com/
12. http://www.octel.com/

13. http://www.isi.com/

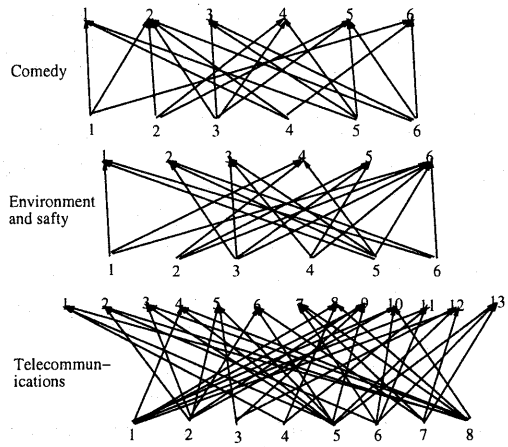

Figure 4. Community examples: Kids, environment and safty, and telecommunitycations.

5.3 Related community results

Comm_link_file preparation

For community extraction, we have prepared $link\_file$ which consists of all the hyper edges of TREC data collection. Similarly, for related community extraction, we prepare $comm\_link\_file$, which consists of edges that denote relationship between a community and its members, as follows. Let N be number of communities extracted from TREC data. For i=1 to N, an edge $< i, p >$ is inserted in $comm\_link\_file$, if p is a member of $community_i$. With $comm\_link\_file$ as input, the community extraction algorithm is used to find the related communities.

With α = 3, β =4, we have extracted 59488 communities using community extraction algorithm. After removing the duplicates among these communities, we extracted related communities using proposed approach. Here, due to space limitation, we show only two examples of related communities extracted from TREC data set.

Example 4. The following two communities are about cultural diversity and astrology.

Arts and culture

1. http://maple.lemoyne.edu/ bucko/indian.html (Fr. Bucko's Mighty Home Page)
2. http://fin.firstnations.ca/Arts/s-index.html
3. http://www.artnatam.com/links.html (ArtNatAm - Links to Other Sites)
4. http://minyos.its.rmit.edu.au/ leo/muserob.html (ART GALLERIES AND MUSEUMS ... at ROB'S)
5. http://www.wvmccd.cc.ca.us/wvc/la/gbarrett/html /Culture.html (Gordon Barrett's Cultural Diversity Page)
6. http://www.sosc.osshe.edu/stu_affa/clubs/ nasu/index.htm
7. http://americanwest.com/pages/indians.htm (The American West - Native Americans)

Artifacts

1. http://www.eruditepub.com/list/all.html (To all? Gateway)
2. http://www.stenhammar.net/aliens/ (Stenhammar.Net - Aliens)
3. http://www1.tpgi.com.au/users/ron/alien_other.html (Alien Information)
4. http://www.links.net/astro/recs.html (the Underground Astrologer: resources)
5. http://www.ramtha.com/links.html (LINKS)
6. http://www.mastery.net/links/newage.htm (Enchantment-Metaphysics and Religion Resources)
7. http://www.squishy.com/music/artists/mystical_sun/ arcstuff.html (Mystical Sun Research, Artifacts, Resources and Links)
8. http://home1.inet.tele.dk/fenger/link.html (Erling Fenger)
9. http://pegasus.cc.ucf.edu/ jrk61265/newage.html (Links List)
10. http://www.winternet.com/ robin/occult.pages.html (Robin's Links to the Mystical Internet - Web Pages)

Example 5. The following three community structures are about AIDS, Medicine and Health information.

AIDS

1. http://aspe.os.dhhs.gov/datacncl/dcmembrs.htm (HHS Data Council Membership)
2. http://www.lawmiami.net/lbry/govagenc.htm (LawFlorida.net -- Government Agencies)
3. http://server2.madcon.com/mad-fed.htm (MADCON Consultation Services' Home Page for Environmental Services)
4. http://smpnet.i2k.net/links/fedgov.htm (SMPNET - Federal Govt.)
5. http://www.lib.loyno.edu/subjects/government.htm (Government) 6. http://sis.nlm.nih.gov/aids/aidstrea.html (HIV/AIDS Treatment Information Service (ATIS))
7. http://www.aacn.nche.edu/GovtAff/galinks.html (AACN Government Affairs Links)
8. http://www.planet-link.com/category/medical.shtml (Planet Link: Health and Medical)

Medicine

1. http://public.nlm.nih.gov/publications/staff_publications/rodgers/internet_course/biomed.html (Biomedical Information Resources on the Web)
2. http://www.keenesentinel.com/ clinic/medlinks.shtml (Medical WWWW Links)
3. http://info.neoplastics.mssm.edu/SitesofInterest.html (NPD Sites of Interest)
4. http://medicineonline.com/misc.htm (Misc)
5. http://www.graylab.ac.uk/cancerweb/sitenet/ hospamer.html (Hospitals: North America)
6. http://wings.buffalo.edu/medicine/med/medical.html (Medical Links)
7. http://his-solutions.com/res.htm (HIS Solutions Resources & Links)

Health information

1. http://www.gotech.com/skb/links.htm (Link-O-Rama)
2. http://www.mandala.com/bookmarks.html (Jeff's Bookmarks)
3. http://fs01.hwp0.ocps.k12.fl.us/health.html (WPHS health)
4. http://www.dsmo.com/archive.htm (Not So New on the Web)
5. http://tavernmiami.com/links.htm (Tavern SuperLinks Page)
6. http://mookie.ontime.com/bookmark.htm (Smart Bookmarks)
7. http://scratchy.hcrhs.hunterdon.k12.nj.us/othersites /health.html (health.html)
8. http://smiley.logos.cy.net/CHARLIE/nutr.html

9. http://pptnet.com/wwwgate/favorite/science.html (Favorite Places - Science & Health)
10. http://www.coxnet.com/tech.html (The Big List Of Links)

6 Summary and conclusions

In this paper we proposed a simple and efficient approach to extract and relate communities from a large collection of web pages by performing hyperlink analysis. The proposed approach gathers related information based on the $relax\_cocite$ relationship and then follows iterative pruning technique to extract a potential DBG pattern. We have employed DBG abstraction for two purposes. From the given page collection, we first extract all the communities by mathematically abstracting the community as a DBG over a set of pages. Next, our approach extracts related communities among the extracted communities by abstracting related community set as a DBG over set of communities. We have conducted experimental results on 10 GB TREC data collection that contains 1.7 million pages and 21.5 million links. The results show that proposed approach extracts meaningful community as well as related community structures.

It can be observed that the quality of extracted communities and related communities depends on the methods and scope of data collection. As a part of future work, we will experiment the proposed ideas on the data collections that differ in size and scope. Also, we will extend proposed approach to build hierarchy of communities for a given data collection.

Acknowledgments
This work is supported by "Research for the future" (in Japanese Mirai Kaitaku) under the program of Japan Society for the Promotion of Science, Japan.

References

[1] R.Agrawal and R.Srikant. Fast algorithms for mining association rules, in proc. VLDB Chile, 1994.

[2] Andrei Z.Broder, Steven C.Glassman, Mark S.Manasse, and Geoffery Zweig, Syntactic clustering of the Web, 6th International WWW conference, 1997.

[3] K.Bharat and M.Henzinger, Improved algorithms for topic distillation in hyper-linked environments, in: proc: 21st SIGIR Conference, Australia, 1998.

[4] Jeffrey Dean, and Monica R.Henzinger, Finding related pages in the world wide web. 8th international WWW conference, 1999.

[5] S.Brin and L.Page, The anatomy of a large scale hypertextual web search engine, in proc. of 7th WWW Conference, April 1998, pp. 107-117.

[6] J.Carriere and R.Kazman. Web query: Searching and visualizing the web through connectivity. In proceedings of 6th WWW Conference, pp. 107-117, April 1997.

[7] S.Chakrabarti, B.Dom, D.Gibson, J.Kleinberg, P.Raghavan and S.Gopalan, Automatic resource compilation by analyzing hyper-link structure and associated text, in proc. of 7th WWW conference, 1998, pp. 65-74.

[8] Ellen Spertus. Parasite: Mining structural information on the Web. In proceedings of 6th WWW Conference, pp. 587-595, April 1997.

[9] G.W.Flake, Steve Lawrence, C.Lee Giles, Efficient identification of web communities, The Sixth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, August 2000, pp.150-160.

[10] E.Garfield. Cocitation analysis as a tool in journal evaluation, Science, 178, 1772.

[11] D.Gibson, J.Kleinberg, P.Raghavan. Inferring web communities from link topology, in proc. ACM Conference on hypertext and hyper-media, 1998, pp. 225-234.

[12] M.M.Kessler. Bibliographic coupling between scientific papers. American Documentation, 14, 1963.

[13] J.Kleinberg, Authoritative sources in a hyper linked environment, proc. of ACN-SIAM Symposium on Discrete Algorithms, 1998. Also, appears as a IBM Research Report RJ 10076(91892) May 1997, and at http://www.cs.cornell.edu/home/kleinber/.

[14] Loren Terveen and Will Hill. Evaluating emergent collaboration on the Web. In Proceedings of ACM CSCW'98 Conference on Computer Supported Cooperative Work, Social Filtering, Social Influences, pp. 355-362, 1998.

[15] P.Krishna Reddy and Masaru Kitsuregawa, Inferring web communities through relaxed cocitation and dense bipartite graphs, In the proceedings of 2001 Data Engineering Workshop (DEWS'2001), Tokyo, March 8-10, 2001.

[16] Ravi Kumar, Prabhakar Raghavan, Sridhar Rajagopalan, Andrew Tomkins, Trawling the Web for emerging Cybercommunities, 8th WWW Conference, May 1999.

[17] John Scott. Social Network analysis : a handbook. SAGE Publications, 1991.

[18] Small, H.G. Co-citation in the scientific literature: A new measure of the relationship between two documents. Journal of American Society for Information Science, 24, no. 4, pp.265-269, 1973.

[19] Masashi Toyoda and Masaru Kitsuregawa, Creating a Web Community Chart for navigating Related Communities, ACM Hypertext 2001.

[20] TREC: Text REtrieval evaluation (http://trec.nist.gov).

[21] http://pastime.anu.edu.au/TAR/vic2.html

[22] H.D.White and K.W. McCain, Bibliometrics, in: Annual Review of Information Science and Technology, Elsvier, 1989, pp. 119-186.