C4-1: Building a community hierarchy for the Web 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 propose an approach to extract and relate the communities by considering a community signature as a group of content creators that manifests itself as a set of interlinked pages. We abstract a community signature as a group of pages that form a dense bipartite graph (DBG), and proposed an algorithm to extract the DBGs from the given data set. Also, using the proposed approach, the extracted communities can be grouped to form a high-level communities. We apply the proposed algorithm on 10 GB TREC (Text REtrieval Conference) data set and extract a three-level community hierarchy. The extracted community hierarchy facilitates an easy analysis of low-level communities and provides a way to understand the sociology of the Web.

Keywords: Web mining, Communities, Trawling, Link analysis.

1 Introduction

The Internet (or Web) has rapidly grown into being an integral element of the infrastructure of the society. One of the most powerful socializing aspects of the Web is the ability to connect a group of like-minded people independent of geography or time zones. In the Web, community forming is one of the important activity as the Web lets people join communities across the globe by providing a vast opportunity to form associations among people; an association can be formed in the Web in many ways, such as by sending an e-mail message, browsing the Web site, or establishing a link with the other pages of interest. 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 resource collections in directories such as Yahoo and Infoseek, or the home pages of Geocities.

In this paper we focus on the problem of finding community information in a given data set by performing hyper-link analysis and ignoring the text information. In the context of the Web, we consider community as a group of content creators that manifests itself as a set of interlinked pages. Recently, Ravi Kumar et al. [1] proposed an approach to extract communities by abstracting a community signature[1] as a complete bipartite graph (CBG). In [2], we have discussed a method to extract the communities by abstracting a community signature as a dense bipartite graph (DBG). In this paper, we abstract the community through an improved DBG definition. Also, we show that by using the proposed algorithm, it is possible to build a community hierarchy. By applying the algorithm we extract a three-level community hierarchy from the 10 GB TREC (Text REtrieval Conference) data set that contains 1.7 million pages and 21.5 million links.

We believe that the community hierarchy information can be used for the following purposes: Firstly, it enables an easy analysis of the low-level communities (such as the detection of the interesting communities) by hand—reaching a few selected communities by staring from a few top-level communities based on the broad topic or interest. Secondly, note that the classification of knowledge by search engines such as Yahoo [3] is done in person. However, the proposed algorithm groups the extracted communities into communities of a high-level, automatically. Observation of the both low-level communities and corresponding high-level communities provides a new insights in reorganizing the information in the Web sites and search engines. And, thirdly, the community hierarchy provides a way to understand the sociology of the Web as it displays the connections among different communities.

The rest of the paper is organized as follows. In the next section, we discuss the community abstraction through bipartite graphs. In section 3, we present the community extraction algorithm. In section 4, we report the experimental results. In section 5, we discuss the related work. The last section consists of conclusions.

[1] In this paper the terms community, community structure, and community signature indicate the same.

2 Bipartite graphs and communities

We first explain how bipartite graphs can be used to abstract the community signatures. Next, we define the community structure based on DBG abstraction. Then, we explain about community hierarchy.

2.1 Dense bipartite graph

We use the following notations. The Web pages are denoted by P_i, P_j, ...; where i, j, ... are integers. A page is referred by its URL, which also denotes a node in a bipartite graph (BG). We refer a page and its URL interchangeably. If there is an hyper-link from page P_i to page P_j, we say P_i is a parent of P_j and P_j is a child of P_i. An hyper-link from one page to other page is considered as an edge between the corresponding nodes in a BG. For P_i, parent(P_i) is a set of all its parent pages and child(P_i) is a set of all its children pages. We use two set notations: fans(F) and centers(C), that are used to represent two groups in a BG as in [1]. Here, we give the definition of a bipartite graph.

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

In the context of Web we consider both F and C represent two groups of Web pages. In a BG, a fan is a Web page that has multiple centers as its children. Similarly a center is a Web page that has multiple fans as its parents. A Web page can be both a fan as well as a center. However as per the BG definition we overlook such occurrence.

By considering only links, a Web page can be abstracted as BG (Here, we ignore the links from a page to itself). A BG of P_i is denoted by BG(F,C), where F = $\{P_i\}$ and C= $\{P_j \mid P_j \in child(P_i)\}$. Each link from P_i to its children is reflected as a directed edge from F to C.

A community can be represented by a unique identifier and its members. Similar to a Web page, a community can be abstracted as a BG(F,C), where F contains the identifier and C contains the identifiers of its members. A directed edge is added from F to the each member of C.

Note that a BG is dense if many possible edges between F and C exist. In a BG, link-density between the sets F and C is not specified. Here, we define a dense bipartite graph by capturing the link-density between the sets F and C as follows.

Definition 2 Dense bipartite graph (DBG) *Let p and q be the nonzero integer variables. A DBG(F,C,p,q) is a BG(F, C), where (i) each node of F establishes an edge with at least p ($1 \leq p \leq \mid C \mid$) nodes of C, and (ii) at least q ($1 \leq q \leq \mid F \mid$) nodes of F establish an edge with each node of C.*

(Note that the notion of density of graph is non-standard. Goldberg [4] proposed that density of the graph is a ratio of number of edges to the number of vertices, and proposed an algorithm to extract a density graph by combining network flow techniques with binary search. However, our interest is to capture the link-density between the two groups in a BG. We will investigate the difference as a part of future work.)

Now we define a complete bipartite graph that contains all the possible edges between F and C.

Definition 3 Complete bipartite graph (CBG) *A CBG(F,C, p, q) is a DBG(F,C, p, q), where p = $\mid C \mid$ and q = $\mid F \mid$.*

It can be observed that in DBG(F,C, p, q), both p and q specify the link-density whereas in CBG(F, C, p, q) same specify both the number of nodes in C and F, and the link-density. Figure 1 shows the difference between a CBG(F,C, p, q) and a DBG(F,C,p,q).

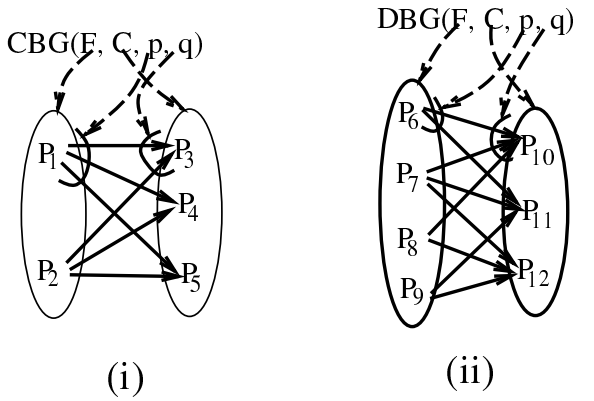

Figure 1. Bipartite graphs (i) CBG(F, C, p, q) (ii) DBG(F, C, p, q)

Theorem 1 *For a given data set and r and s, let dense bipartite graph set, DBGS(r,s) = {DBG(F,C, p, q) $\mid$ $p \geq r$ and $s \geq q$} and complete bipartite graph set, CBGS(r,s) = {CBG(F, C, p, q) $\mid$ $p \geq r$ and $s \geq q$}. Then, CBGS(r,s) $\subseteq$ DBGS(r,s).*

Proof: *Consider a CBG(F,C, p, q) $\in$ CBG(r,s). According to definition, DBGS(r,s) includes all the DBG(F,C, p, q) patterns such that $p \geq r$ and $q \geq s$. Consider a CBG(F,C, p, q). This implies that DBGS(r,s) includes a DBG(F,C, p, q) with p = $\mid C \mid$ and q = $\mid F \mid$ which is a CBG(F,C, p, q). So, CBGS(r,s) $\subseteq$ DBGS(r,s).*

All the CBGs are the instances of DBGs. That is, at fixed r and s values, if we extract all DBGS(r,s), all the CBGs in CBGS(r,s) are automatically extracted, i.e., the CBG structures in CBGS(r,s) are embedded in the corresponding DBGs. However, if there exist a DBG(F,C,r,s) pattern in a data set, there is

no guarantee that corresponding CBG(F,C,r,s) pattern is embedded. Because, the DBG definition covers more graph structures than the CBG definition. The embedding of a CBG in a DBG depends on the link-density between the pages in the data set. Overall, by extracting DBGs from a given data set, one can extract more graph strctures including the CBG structures. However, the converse is not true—if we extract all the CBG structures, corresponding DBG structures are not included.

2.2 Community abstraction

Our definition is based on the following intuition: *Web communities are characterized by DBGs.* In the Web environment, a 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. We made an effort to capture such a phenomena through by defining a DBG and proposed a simple and scalable algorithm to extract the DBGs.

Normally, each member in a community shares the interests with a few other members. Therefore, the abstraction of a community structure through a DBG matches comparatively well with the real community patterns. In general community can be viewed as a macro-phenomena manifested 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 the 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.

We abstract a community as a set of closely associated pages that form a DBG. Similarly, we abstract a high-level community as a DBG over a set of a low-level communities. The term *community* to represent both a low-level as well as a high-level community. So the members of the community can be Web pages or the communities, which depends on the level of the community. For the sake of simplicity, we consider a node to represent both a Web page as well as a community. Let the variable *num_levels* denote the number of the levels. A community is denoted by C_{ij}, where i ($1 \geq i \geq num\_levels$) is a nonzero integer value that denotes the level of the community and j is an integer value which denotes the unique community identifier at the level, i. A community, C_{ij}, is defined as follows:

Definition 4 Community C_{ij}**:** *Let p_t and q_t be the integer variables that represent the threshold values. Also, let rcf be a nonzero integer variable that represents the relaxation control factor. For some i ($i > 0$) and j ($j > 0$), C_{ij} = F, if there exist a DBG(F,C, p, q) over a set of nodes at level "i-1" with the following properties.*

- $p \geq p_t$ *and* $q \geq q_t$*; and*
- $\mid F \mid <= p_t * rcf$.

We consider a DBG pattern as meaningful if the pages in F agree on some topic. Note that not all the DBG(F,C, p,q) patterns, where $p \geq 1$ and $q \geq 1$, are meaningful community patterns. The filtering process is done with the two constraints in Definition 4. Firstly, we fix the threshold values for both p and q as p_t and q_t, respectively. And, secondly, from the experiment results on the real data, it was observed that some of the extracted DBG(F,C,p,q) patterns get a bigger F and C. In such cases it was observed that F contains pages with multiple topics. Through second constraint such cases are filtered out by bounding the number of nodes in C equal to p_t*rcf. If rcf=1, the number of nodes in F becomes equal to p. However as rcf increases, the size of C increases and F gets more members. Note that the values of p_t, q_t and rcf are fixed after examining a reasonable number of the DBG patterns for the given data set.

2.3 Community hierarchy

Also, it can be noted that the proposed approach can be extended to find a high-level communities among a low-level communities that forms a community hierarchy. Note that if i=1, the members of C_{ij} are the Web pages. Otherwise, if $i > 1$, the members of C_{ij} are the communities of level "i-1". That is, the set F in DBG(F,C, p, q) over the Web pages forms a community C_{1j}. So, the members of C_{1j} are the Web pages. However, the set F in DBG(F,C,p,q) over a set of communities of type C_{1j}, j=1 ... L1, (there are L1 communities at 1-level) forms a community at 2-level, i.e., C_{2j}. In this way the low-level communities become the members of the successive high-level communities that forms a community hierarchy. This is interesting in the sense that if we extend from bottom to top, we can build an hierarchy of communities for a given data set. In general, given a set of nodes (community or a page) and association information among them, the proposed DBG abstraction helps to extract the interesting groups from the given set of nodes.

3 Extraction of DBGs

In the information retrieval literature, the documents are related based the notion of syntactic relationship that is measured based on the existence of number of common keywords. Similarly, in the Web environment we consider a link as an association between pages. So by dealing with only links we establish an association among pages based on the existence of the common links (or URLs). In the context

of the Web environment, we call this relationship *CommLink* and is defined as below.

Definition 5 CommLink *Let P_i and P_j be pages. Then CommLink(P_i, P_j) = $\mid child(P_i) \cap child(P_j) \mid$. We say both P_i and P_j are related if CommLink(P_i, P_j) is greater than the predefined threshold value.*

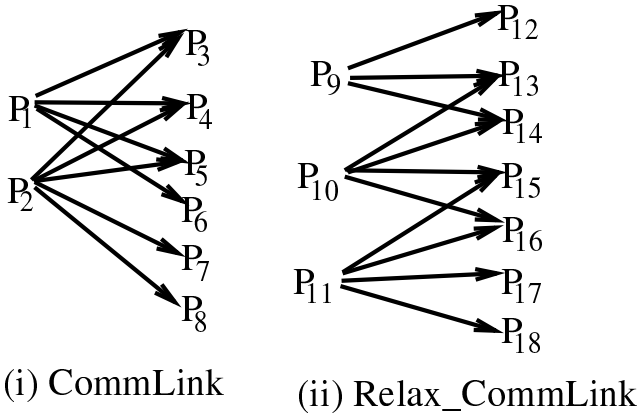

Figure 2. Depiction of the relationships.

Figure 2(i) depicts the *CommLink* relationship between the pages P_1 and P_2, with CommLink(P_i, P_j)= 3. Under *CommLink*, n ($n \geq 2$) pages are related if these pages have common children at least equal to a threshold value. If n pages are related according to the *CommLink* relationship at a certain threshold value, say t, these pages form a CBG(F,C,p,q), with $\mid F \mid$=q=n and $\mid C \mid$=p=t.

To extract a DBG pattern, we have to retrieve a collection of pages that are loosely related. To find the pages that are loosely related, we relax the *CommLink* relationship in the following manner. We allow pages P_i, P_j and P_k to group if both pairs, P_i and P_j, and P_j and P_k, are related under CommLink. This modification enables relationship between a page and multiple pages taken together. That is, if a page could not form an association with the another page according to *CommLink*, it does not imply that they are different. Even though a page fails to satisfy a certain minimum criteria pagewise, however, it could satisfy the minimum criteria with multiple pages taken together. We define the corresponding new definition, *Relax_CommLink*, as follows.

Definition 6 Relax_CommLink. *Let T be the set of pages and P_i be the another page ($P_i \notin T$). Then, Relax_CommLink(P_i,T)= $\mid child(P_i) \cap child(T) \mid$. We say both P_i and the pages in T are related if Relax_CommLink(P_i, T) is greater than the predefined threshold value. Here, child(T) contains the pages that are the children of T's members.*

It can be observed that, as compared to *CommLink*, P_i can be associated with more number of pages using *Relax_CommLink*, as T contains more members. Figure 2(ii) depicts the three pages which were grouped using the *Relax_CommLink* relationship with the threshold value equal to 2. Note that P_{11} is grouped with both P_9 and P_{10} which is not possible with the CommLink relationship. However, unrelated pages can be clustered together under *Relax_CommLink*. So after collecting a reasonable number of pages we employ iterative pruning methods to extract a DBG pattern.

3.1 The algorithm

Given a large collection of nodes (all the nodes are either the Web pages or the communities), the algorithm to extract DBG structures consists of two steps: gathering the related nodes and the extraction of DBGs. For each node P_i, we gather related nodes during the gathering phase through the *Relax_CommLink* relationship. We then apply the iterative pruning technique to extract a DBG(F,C, p, q) given the threshold values p_t and q_t. We now present corresponding routines.

1. **Gathering related nodes**
   For a given node, P_i, we find the set *rel_set*, which is a set of nodes related to P_i. We use three integer variables, *threshold*, *num_iter*, and n, which are set to 1, 0, and the maximum number of iterations (> 0).
   (a) Set *rel_set* = { P_i }.
   (b) While *num_iter* $\leq n$
      i. Find all P_j such that *Relax_CommLink*(P_j, rel_set) $\geq$ *threshold*.
      ii. *rel_set* = { T_j } $\cup$ *rel_set*.
      iii. *num_iter* = *num_iter* + 1;
   (c) Output *rel_set*.
2. **Extracting a DBG**
   In this step we extract a DBG(F, C, p, q) from *rel_set*. Let the variable *edge_file* be the set of elements $< P_i, P_j >$ where P_i is a parent (source) of child P_j (destination). The *edge_file* is set to ϕ. (In the following steps, notations child(P_i) indicates the set of nodes P_j such that $< P_i, P_j > \in$ *edge_file*. Similarly, parent(P_j) indicates the set of nodes P_i such that $< P_i, P_j > \in$ *edge_file*.) The values of p_t, q_t and rcf are given. Let p and q be theinteger variables.
   (a) p= p_t, q=q_t.
   (b) For each $P_i \in$ *rel_set*, insert the edge $< P_i, P_j >$ in *edge_file* if $P_j \in$ *child*(P_i).
   (c) While *edge_file* is not converged repeat the following.
      i. Sort the *edge_file* based on the destination. Remove $< P_i, P_j >$ from *edge_file* if $\mid$ *parent*(P_j) $\mid < q$.
      ii. Sort the *edge_file* based on the source. Remove $< P_i, P_j >$ from *edge_file* if $\mid$ *child*(P_i) $\mid < p$.
   (d) Let C ={$P_j \mid < P_i, P_j > \in$ *edge_file* }. If $\mid C \mid > p \times rcf$, then q= q+1 and go to

(c). (This routine tests whether the number of nodes in C is less than or equal to $p \times rcf$. Other wise, the value of q is incremented, forcing the each member of C to have in-links with more members of F that reduces the number of nodes in C.)

(e) The resulting $edge\_file$ represents a DBG(F, C, p, q) where, F={ $P_i \mid < P_i, P_j > \in edge\_file$ } and C= {$P_j \mid < P_i, P_j > \in edge\_file$ }.

4 Experiment results

We conducted experiments on 10GB TREC [5] (Text Retrieval Conference [6]) data set. It contains 1.7 million Web pages. We first explain briefly about the preprocessing steps and then discuss the results.

4.1 Preprocessing

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 [1]): 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 [7] 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 [1]). 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.

4.2 Community hierarchy results

By applying the proposed algorithm we have extracted a 3-level community hierarchy. At all the levels, we have set p_t=3, q_t=3 and rcf = 5, and applied the proposed algorithm to find the DBG structures.

From the TREC data set we have extracted 67698 DBG strctures of 1-level. Each DBG(F,C,p,q) pattern is considered as a community (with unique identifier) and its members are the elements of F (URLs). So we ignore the elements of C. Next, the duplicate communities are removed. We consider two communities as duplicates if the members of the one community are equal to or the subset of the members of the other community.It was found out that most of them are duplicates. After eliminating the dublicates the number of communities at 1-level comes to 15857.

To extract the 2-level communities, we formed $< id, member >$ pairs of 1-level communities and applied the proposed algorithm. In this case we found 14698 communities. Note the the members of the each 2-level community are the identifiers of the 1-level communities. After eliminating the duplicate communities, the number of communities comes to 2010.

We repeated the process and extracted the 3-level communities that comes to 1734. After eliminating the duplicates, the number comes to 332. The results are summarized in Figure 3.

Level	**# of DBG**	**# of DBG (after duplicate elimination)**
1-level	67698	15857
2-level	14698	2010
3-level	1734	332

Figure 3. Community hierarchy statistics

The community hierarchy results are displayed at [8]. For each community, the potential key words are displayed. One can catch the topic of the community through these key words. Note that the members of each 1-level community are URLs (Web pages). We merged the title text of these pages and selected top 20 frequent words. Similarly to represent a 2-level community, top 30 key words are selected after merging the key words of the corresponding 1-level

communities. Similarly, each community of the 3-level is represented with top 50 frequent words after merging the key words of the corresponding 2-level communities. We provide three sample communities (one each from 1-level, 2-level and 3-level) extracted from the TREC data set.

- **A 1-level community; topic: telecommunications** The community structures are extracted by extracting the corresponding graphs of type DBG(F, C, 3, 3). Figure 4 shows a DBG(8, 13, 3, 3) graph. Through the title text of the Fans, it can be observed that the topic is *Telecommunitications*

  Fans:
  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
  7. http://www.searchemploy.com/research.html (Search & Employ)
  8. http://www.electsource.com/elecoem.html (Electronics OEM's)

  Centers:
  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/

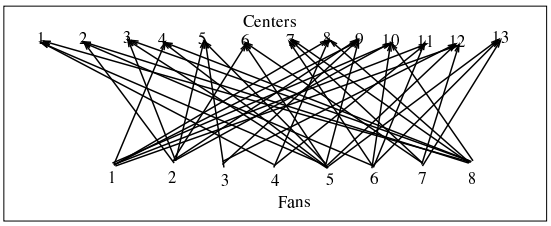

Figure 4. A 1-level community pattern: a DBG(8,13,3,3). The topic is telecommunications. A arrow mark from a fan to a center depicts an hyperlink.

- **A 2-level community; topic: Nutrition and food habits** Kitchen Wine Food Link Missouri Country Resources Sites Nutrition Health Cooking Consumer Recipes Right Page Drink Interest Eating MedLinks EuroNet Quincy San Home Virtual Excite Miscellaneous Medical MESA1 CIMC Other interNet Diego Magazine OnLine Notes! Online96 David Lists Menus Restaurant NetDirectory Reviews TuBears Helpful ZIA Directory Recreation amp Diet (The corresponding 1-level communities are as follows.)
  - Recipes Food Cooking Wine Kitchen David interNet Notes! San Diego Magazine OnLine Home Page Lovers Forum Virtual Quincy
  - Health Nutrition Wine MESA1 Page Miscellaneous Medical The Kitchen Link CIMC Other Sites Strathcona Consumer MedLinks Eating Right Kids
  - Phamacology Falk Library University Kitchen Link Health Nutrition Apoteket Pharmacy NYC Pharmacies Consumer MedLinks Pharmaceutical Eating Right Universities
  - Recipes Wine Kitchen Link Sites Assorted Food Virtual Quincy Directory Recreation Cooking amp Country Missouri
  - Cooking Recipes Food Wine Links Excite NetDirectory Recipe Grab Bag David interNet Notes! The Daily bLink San Diego
  - Kitchen Medical Robertson Websites Health Nutrition CIMC Other Sites Food Related Consumer MedLinks Excite
  - Nutrition Sites Health Wine Consumer Healthcare Kitchen Resources TuBears Diet Federal Information Yahoo! Country Missouri Department Food Religion
  - Kitchen Wine Virtually California San Francisco Delectable Specialty Foods Online96 Food Chefs Foodservice Cooking Schools Country Missouri Randy
  - Health Kitchen Nutrition Boulder Community Network Center Resources WPHS health TuBears Diet Consumer MedLinks Eating Right Helpful Millard
  - Kitchen Link Restaurant Menus Reviews Wine EuroNet Interest Food Drink Country Missouri Business Factory Restaurants
  - Wine EuroNet Interest Food Drink Kitchen Beverages Chrisbac Missouri
  - Kitchen Restaurant Menus Reviews Wine EuroNet Interest Drink Chris Davy Bookmarks HomeArts Hot Country Missouri
  - Wine Food Rich Herman BBQ ZIA Barbecue Wings Missouri
  - Recipes Wine ZIA Religion Islam Resources Food College Islamic WWW Sites Virtual Quincy Directory Recreation Cooking Missouri
  - Health Resources MESA1 Kitchen Nutrition Maryland Sea Grant Extension HACCP Sites Eating Right Helpful Excite NetDirectory Games
- **A 3-level community; topic: Bio-chemistry** Biology Resources University Genetics Rockefeller Research VIRION Chris Computing Library Molecular Lab Chemistry MPBC Burge Bio Biomedical Servers Connections Favorite Virtual Wide Macromolecular Interest Journals Crysta Biochemistry Structure Sequence Microbiology (The corresponding 2-level communities are as follows.)
  - Biology Resources Information University Sciences Biological Genetics Rockefeller Research VIRION Servers Computing Web Science Center Sites Services Library Molecular Lab Biomedical Chris Chemistry MPBC Bio Burge Chemical World Scientific Virtual Macromolecular Darst
  - Scientific Biology Rockefeller University Information Computing Online Library Sites Journals CrystaLinks Lab Services Chemistry Useful Publication html Science journals Sciences Darst Biological VIRION links Electronic Web Center Research Page MPBC Chris Other News Genetics Biochemistry Structure Sequence Burge WUSTL Microbiology Molecular publicat Chemical
  - Biology Information Biological Web Research html Genetics VIRION Library bioref2 Lab Protein Chemistry Center University Chemical Useful Rockefeller Servers Sites Darst Department Other Scientific Online MPBC Chris update Last Connections 95 Biomedical Special 1 Burge Computational Biochemistry Favorite Macromolecular Computing Hospital

- Sciences Biological Library Biology Science Web Macromolecular Favorite Information BioTech Chemistry Research Microbiology WUSTL Genetics University Rockefeller Protein Scientific Biotech IMV BioInformatics Scientifics Computational SUNY CMI Journals
- Biology Biological Library Sciences Macromolecular Research Genetics BioTech Biotech Rockefeller WUSTL Microbiology Virtual Computing Scientifics Wide CMI Chemistry Lab Biomedical Scientific Molecular MPBC Bio Companies VIRION

5 Related work

In this section, we review the approaches related to data mining and link analysis, and community detection.

Data mining and link analysis: The data mining approach [9] focuses largely on finding association rules and other statistical correlation measures in a given data set. The notion of finding communities in the proposed approach differs from data mining since we exploit co-citation whereas 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 [10], where network properties such as cliques, centroids, and diameters are used to analyze the collective properties of interacting agents. The fields of citation analysis [11, 12, 13] and bibliometrics [14, 15] also use citation links between the works of literature to identify patterns in collections. Also, most of the search engines perform both link as well as text analysis to improve the quality of search results. Based on link analysis many researchers proposed schemes [16, 17, 18, 19, 20, 21, 22] to find related information from the Web. Chakrabarti [23] surveys research works in the area of hypertext mining.

Community detection:

In [24], cocitation analysis technique [25] has been extended to cluster the Web pages by considering that a hyperlink provides a semantic linkages between pages in the same manner that citations link documents to other related documents. The principle component of cocitation analysis measures the number of documents that have cited a given pair of documents together. It has been shown that citation analysis shown to create better formed and more meaningful clusters of documents.

In [26], 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 [21]. 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. The motivation behind the HITS is to find good authority pages given a collection of pages on the same topic. Our motivation is to detect all the communities in a larger collection of pages that covers a wide variety of topics.

Ravi Kumar et al. [1] proposed an approach to find the potential community cores by abstracting a core of the community as a group of pages that form a complete bipartite graph (CBG). A CBG abstraction extracts a small set of potential members to agree on some common interests. Given a very large collection of pages, for each community there might exist a 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. Also, it rarely happens that a page-creator puts links to all the pages of interest in particular domain. Because, a data set may not contain the potential pages to form a CBG.

In [27], 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 [28], the Companion algorithm is proposed 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. The Companion algorithm first builds a subgraph of the Web near the seed, and extracts the authorities and hubs in the graph using HITS. The authorities are returned as related pages. In [29], the companion algorithm is extended to find the related communities by exploiting the derivation relationships between pages.

The proposed approach is different due to the fact that we abstract a community through a DBG and we use the notion of transitive page similarity based on common links.

6 Summary and conclusions

In this paper, we proposed an algorithm to extract community signatures by mathematically abstracting the community as a DBG over a set of pages. Next, our approach extracts the related communities among the extracted communities by abstracting a high-level community as a DBG over a set of communities. We experimented on the TREC data set and built a three-level community hierarchy. The community hierarchy enables an easy analysis of the low-level communities by hand and provides a way

to understand the sociology of the Web.

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] Ravi Kumar, Prabhakar Raghavan, Sridhar Rajagopalan, Andrew Tomkins, Trawling the Web for emerging Cyber-communities, in proc. of 8th WWW Conference, May 1999.

[2] P.Krishna Reddy and Masaru Kitsuregawa, An approach to relate the web communities through bipartite graphs, in proc. The Second International Conference on Web Information Systems Engineering (WISE2001), Dec. 2001, IEEE Computer Society.

[3] Yahoo! (http://www.yahoo.com), November 2001.

[4] A.Goldberg, Finding a maximum density subgraph, University of California, Berkeley, Technical report, CSD-84-171, 1984.

[5] http://www.ted.cmis.csiro.au/TRECWeb/access_to_data.html, Nov. 2001.

[6] TREC: Text REtrieval evaluation (http://trec.nist.gov), August 2000.

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

[8] Community hierarchy, http://www.tkl.iis.u-tokyo.ac.jp/~reddy/community.html, Feb. 2002.

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

[10] John Scott, Social Network analysis : a handbook, Sage Publications, 1991.

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

[12] H.G.Small. 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.

[13] D.H.White, and G.C. Belver. 1980. Author cocitation: A literature measure of intellectual structure. Journal of American Society for Information Science, 28, no. 5, pp.345-354, 1980.

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

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

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

[17] 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.

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

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

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

[21] J.Kleinberg, Authoritative sources in a hyper linked environment, in proc. of ACN-SIAM Symposium on Discrete Algorithms, 1998.

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

[23] S.Chakrabarti, Data mining for hypertext: A tutorial survey, ACM SIGKDD Explorations, 1(2), pp. 1-11, 2000.

[24] James Pitkow, Characterizing World wide ecologies, Ph.D Thesis, Georgia Institute of Technology, June 1997.

[25] H.Small and B.Griffith. The structure of scientific literatures I, Identifying and Graphing Specialties. Science Studies, 4(17), pp. 17-40, 1974.

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

[27] G.W.Flake, Steve Lawrence, C.Lee Giles, Efficient identification of web communities, in proc. of 6th ACM SIGKDD, August 2000, pp.150-160.

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

[29] Masashi Toyoda and Masaru Kitsuregawa, Creating a Web community chart for navigating related communities, in proc. of 12th ACM Conference on Hypertext and Hypermedia, August 2001, pp. 103–112.