An approach to build a cyber-community hierarchy

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 community structures in the Web by considering a community structure as a group of content creators that manifests itself as a set of interlinked pages. We abstract the community structure as a dense bipartite graph (DBG) over a group of Web pages and proposed an algorithm to extract the DBGs from the given data set. Also, a high-level community is abstracted as a DBG over a set of a low-level communities. Using the proposed approach, a community hierarchy can be constructed for the given data set that generalizes a large number of low-level communities into a few high-level communities. Using the proposed approach, we have extracted a a three-level community hierarchy from 10 GB TREC (Text REtrieval Conference) data set. We believe that the extracted community hierarchy facilitates easy analysis of the low-level communities, helps in reorganizing the Web sites, 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 infrastructure of 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. The Web lets people join communities across the globe by providing a vast opportunity to form associations among people. For instance, in the past one might have had a time to be a part of a neighborhood community and one or two social organizations. However, in the Web environment, one is limited only by the interests as the Web provides an opportunity to form connections with the entire world. An association in Web can formed 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. Due to the vast increase of the neighborhood domain, it is easy for a user to make the connections with the users of similar interest. In the context of the Web, we term such user groups as communities. The forming of the communities is an important activity in the Web. As a result, the Web contains several thousand well-known, explicitly defined communities. 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 [1].

In this paper we focus on the problem of finding community information in a given data set by performing hyperlink analysis on the Web pages and ignoring the text information. In the context of the Web, we consider a community structure 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 the communities by abstracting a community signature as a complete bipartite graph (CBG) over a group of Web pages. In [2], we have discussed a method to extract the communities by abstracting a community structure as a dense bipartite graph (DBG) over a group of Web pages. (In this paper the terms community, community structure, and community signature indicate the same.) We define a DBG by relaxing a link density criteria as compared to CBG. Due to the relaxation, if we extract the DBG structures from the given data set, we can also extract the additional community structures that fail to satisfy the CBG criteria. However, during the experiments on the real data set, it was observed that even though most of the extracted community structures are meaningful, however, a few big community structures that contain multiple topics are also being extracted. In this paper we report a community extraction algorithm by abstracting a community through an improved community definition (with constraints) that extracts the meaningful community structures. In addition, we show that by using the proposed algorithm, it is possible to build a community hierarchy. With the proposed approach, we extract a three-level community hierarchy from the 10 GB TREC [6] (Text REtrieval Conference) data set that contains 1.7 million pages and 21.5 million links.

*With International Institute of Information Technology (IIIT), Hyderabad (India), from April 2002.

We believe that the information regarding the existence of various communities and their relationships in a community hierarchy can be used for the following purposes: Firstly, note that even though we extract the thousands of community structures from a given data set, it is not a straightforward task to find the interesting communities by analyzing these communities by hand. Normally it is easy to analyze a few communities by hand. The community hierarchy generalizes a large number of the low-level communities into a few high-level communities. So it enables an easy analysis (detection of the interesting communities, for instance) of a large number of low-level communities by hand; one can start from a few top-level communities based on the broad topic or interest and reach the corresponding low-level communities by specializing the topic. Secondly, note that the classification of knowledge by search engines such as Yahoo [3] is done by hand. However, the proposed algorithm groups the extracted communities and forms a high-level communities, automatically. Observation of both the low-level communities and the successive high-level communities provides a new insights in reorganizing the information in the Web sites and search engine sites. And, thirdly, the community hierarchy provides a way to understand the sociology of the Web as it displays the implicit connections among the different communities.

The rest of the paper is organized as follows. In the next section, we define a dense bipartite graph and define a community structure in a hierarchy. 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.

2 Dense bipartite graph and community in a hierarchy

We first explain about dense bipartite graph. Next, we define the community structure in a hierarchy.

2.1 Dense bipartite graph

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 the BG joins a node in F to a node in C.*

The set notations F and C in a BG refer the terms Fans and Centers, respectively (adopted from [1]). A fan is a page that has multiple centers as its children. Similarly a center is a page that has multiple fans as its parents. A Web page can be both a fan and center. However as per the BG definition, we overlook such occurrence.

In this paper Web pages are denoted by P_i, P_j, $\ldots$; where i, j, ... are integers. We refer a page and its *URL* interchangeably. If there is a hyperlink (or simply a link) from a page P_i to a page P_j, we say P_i is a parent of P_j and P_j is a child of P_i; also we say P_i has an out-link to P_j and P_j has an in-link from P_i. 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.

By considering only the links and ignoring the text, a Web page can be abstracted as BG. A page denotes a node in a BG and a link from one page to the other page is considered as an edge between the corresponding nodes in a BG. We ignore the links from a page to itself. A BG of a Web page, 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.

Note that a BG is dense if many possible edges between F and C exist. In this paper, for a BG, the term link density criteria is used to specify the minimum number of out-links (in-links) each member in F (C) establishes with the members of C (F). In a BG, link-density criteria between the sets F and C is not specified. Here, we define a dense bipartite graph by specifying the link-density criteria in a BG 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 each node of F establishes an edge with at least p ($1 \leq p \leq \mid C \mid$) nodes of C, and 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 a graph is nonstandard. Goldberg [4] proposed that density of the graph is a ratio of the 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.)

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

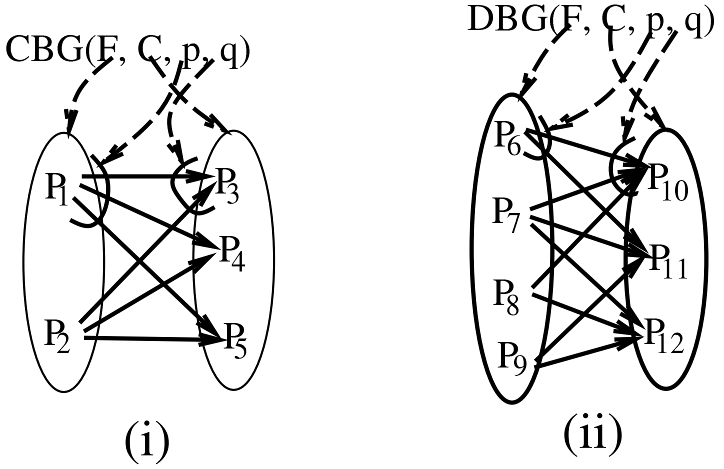

Figure 1. Bipartite graphs: (i) CBG(F, C, p, q) (ii) DBG(F, C, p, q)

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

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

Note that, given r and s values, if we extract DBGS(r,s), CBGS(r,s) is automatically extracted, i.e., the CBG structures in CBGS(r,s) are embedded in the corresponding DBG structures in CBGS(r,s). Figure 2 shows a sample DBG structure with an embedded CBG structure which is extracted from the TREC data set. However, there is no guarantee that for each DBG(F,C,r,s) structure in a data set, corresponding CBG(F,C,r,s) structure is being embedded into it. Because, DBGS(r,s) contains more graph structures than CBGS(r,s) as the DBG definition covers more graph strictures over the CBG definition due to the relaxation of the link density criteria. The possibility of a CBG being embedded in a DBG depends on the link-density among the pages in the data set. So not every DBG(F,C,r,s) contains an embedded CBG(F,C,r,s). (Figure 7 shows a DBG structure without an embedded CBG structure.)

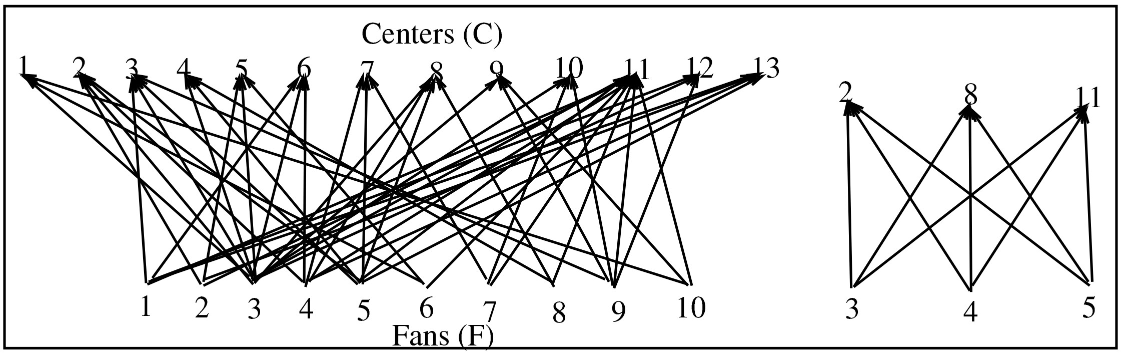

Figure 2. A DBG(10,13,3,3)structure with an embedded CBG(3,3,3,3) structure (extracted during the experiments on the TREC data set). The integer values denote Web pages. An arrow mark from a fan to a center depicts a hyperlink.

2.2 Community in a hierarchy

In general, a community can be viewed as a macro-phenomena manifested by complex relationships exhibited by the corresponding members. At a micro-level, each member establishes a relationship with a few members of the same community. By knowing one member of the community, the other members can be known through the relationship information between the member and other members. Similarly, in the context of the Web, we made an effort to extract the community structures through the link information, automatically. This is done by considering an existence of a DBG structure over a group of Web pages as a community structure.

In the Web environment, a page-creator (a person who creates the page) creates the page in isolation by putting the links to the other pages of interest. If multiple pages are created with similar interests, we believe that, these pages form a community structure. Our intuition is that such Web communities can be characterized by DBGs. We use a DBG to characterize the pages that contain links with similar interests. We capture such a phenomena by defining a DBG by capturing a link density in a BG, and proposed a simple and efficient algorithm to extract the DBGs from the given data set.

Community hierarchy:

Apart from the extraction of the communities from the Web pages, the notion of community abstraction through DBG can be extended to find a high-level communities over the extracted communities. Note that a community can be represented with unique identifier and the members. So 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 all the members of C. Also, the notations parent, child and the other explanation in Section 3 apply to the communities, accordingly.) By considering a low-level community identifiers as its members, a high-level community can be abstracted as a DBG over a set of the low-level communities. In this way, low-level communities become the members of the successive high-level communities. By extending this process, a community hierarchy can be extracted for a given data set. In general, given a set of nodes of any type and association information among them, the proposed DBG abstraction helps to extract the hierarchy of interesting groups.

Definition of a Community:

Here, we define the community that covers the communities of all the levels in a hierarchy. The members of the communities of 1-level are the Web pages and the members of the communities at the other levels are the communities of the predecessor level in the hierarchy. We use the term node to represent a Web page as well as a community. The nodes at the level 0 are the Web pages, and the nodes at the other levels are the communities. A community is defined as follows:

Definition 4 Community C_{ij}**:** *Let the variable num_levels denote the number of the levels in the hierarchy. Let* p_t *and* q_t *be the integer values that represent the threshold values, and* rcf *be a nonzero integer variable that represents the relaxation control factor. A community at level i* ($1 \geq i \geq num\_levels$) *is denoted by* C_{ij}*, where j* ($j > 0$) *denotes the community identifier at the level i. Then,* $C_{ij} = F$*, if there exist a DBG(F,C, p, q) over a* **set of nodes** *at the level "i-1" with the following constraints:*

1. $p \geq p_t$ *and* $q \geq q_t$*; and*
2. $\mid C \mid <= p_t * rcf$.

The constraints in Definition 4 are imposed to filter the meaningful communities. We consider a DBG structure meaningful if the pages in F are related to the same topic. However, it was observed that not all the DBG(F,C, p,q) structures, where $p \geq 1$ and $q \geq 1$ (i.e., without any constraint) are meaningful. So through the first constraint, we fix the threshold values for both p and q as p_t and q_t, respectively. From the experiment results on the real data, it was observed that some of the extracted DBG(F,C,p_t,q_t), (at p_t=3 and q_t=3) structures get a bigger F and C. In such cases it was observed that the pages in F are not related to the same topic. The second constraint is used to filter such cases by bounding the number of nodes in C equal to p_t*rcf. If rcf=1, the number of nodes in C becomes equal to p_t. Through rcf, the size of C can be controlled, which in turn affects the size of F (see the algorithm). Note that the values of p_t, q_t and rcf are fixed after examining a reasonable number of the DBG structures for the given data set.

3 Extraction of community structures

In the information retrieval literature [5], the documents are related based the notion of a syntactic relationship that is measured based on the existence of number of common keywords. Similarly, we consider a link as an association between pages in the Web environment. So by dealing with only links, we establish a relationship among pages based on the existence of the common links (or URLs). In the context of the Web environment, we call this relationship *CommLink* which 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 at least equal to the predefined threshold value.*

Figure 3(i) depicts the *CommLink* relationship between the pages P_1 and P_2, with CommLink(P_1, P_2)= 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 under *CommLink* 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.

However, to extract a DBG structure, we have to retrieve a collection of pages that are loosely related. To find the pages that are loosely related, we relax *CommLink* 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 under *CommLink*, they might be related, however. Here we try to relate such pages by broadening the criteria of relationship with the notion that even though a page fails to satisfy a certain minimum criteria with other related pages in a page-wise manner, it could satisfy the minimum criteria with multiple pages taken together. We define the corresponding new relationship, *Relax_CommLink*, that captures such related pages 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 at least equal to the predefined threshold value. Here, child(T) contains a set of 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 under *Relax_CommLink*, as T contains more members. Figure 3(ii) depicts three pages which were grouped under *Relax_CommLink* with the threshold value equal to 2. Note that P_{11} is grouped with P_9 which is not possible with the CommLink relationship. However, unrelated pages can be grouped together under *Relax_CommLink*. So after collecting a reasonable number of related pages, we employ iterative pruning methods to extract a DBG structure.

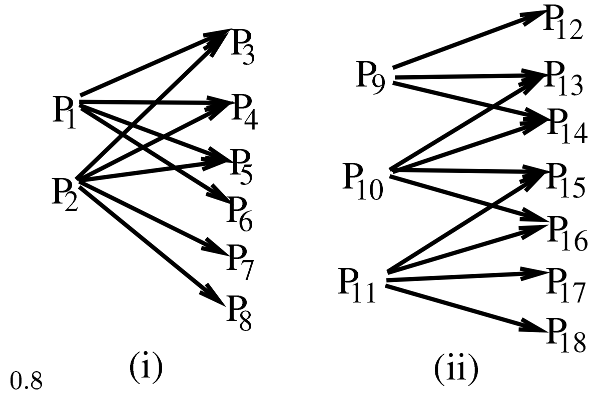

Figure 3. Depiction of the relationships: (i) CommLink (ii) *Relax_CommLink*.

3.1 Community (DBG) extraction algorithm

The input is *node_set*, which is a set of either the Web pages or communities, and the output is the DBG(F,C, p_t, q_t) structures at the given values of p_t, q_t and rcf. For each node $P_i \in$ *node_set*, we first gather the related nodes. Next, we extract a DBG structure from these nodes.

1. **Gathering the related nodes**
   The variable *max_iter* (> 0) denotes the maximum number of iterations. The variables *threshold* and *num_iter* are the integer variables. The variables *rel_set* and *temp_set* denote a set of nodes. Given the node, (P_i) and the value of *max_iter*, the following routine finds a set of related nodes, *rel_set*.

```
(a) threshold=1; num_iter =0; rel_set = { $P_i$ }; temp_set = $\phi$.
(b) While num_iter $\leq$ max_iter
    i. For all $P_j \in$ node_set such that Relax_CommLink($P_j$, rel_set) $\geq$ threshold, temp_set = { $P_j$ } $\cup$ temp_set.
    ii. rel_set = rel_set $\cup$ temp_set.
    iii. temp_set = $\phi$; num_iter = num_iter + 1.
(c) Output rel_set.
```

2. **Extracting a DBG**
   We extract a DBG(F, C, p_t, q_t) from *rel_set*. Let the variable *edge_set* be the set of elements $< P_i, P_j >$ where P_i is a parent (source) of child P_j (destination). The *edge_set* is set to ϕ. Both p and q are integer variables. (The values of p_t, q_t and rcf are fixed.)

(a) p= p_t, q=q_t, $edge\_set = \phi$.

(b) For each $P_i \in rel\_set$, the edge $< P_i, P_j >$ is inserted in $edge\_set$ if $P_j \in child(P_i)$. (In the following steps, notations child(P_i) indicates the set of nodes P_j such that $< P_i, P_j > \in edge\_set$, and parent($P_j$) indicates the set of nodes P_i such that $< P_i, P_j > \in edge\_set$.)

(c) While $edge\_set$ is not converged repeat the following steps:

  i. The $edge\_set$ is sorted based on the destination. Each edge, $< P_i, P_j > \in edge\_set$, is removed if $\mid parent(P_j) \mid < q$.

  ii. The $edge\_set$ is sorted based on the source. Edge edge, $< P_i, P_j > \in edge\_set$, is removed if $\mid child(P_i) \mid < p$.

(d) Let C =$\{P_j \mid < P_i, P_j > \in \; edge\_set \}$. 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 exceeds $p \times rcf$. Otherwise, the value of q is incremented. This step forces each member of C to have an in-link with more members of F, and therefore turn reduces the number of nodes in C.)

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

4 Experiment results

We conducted experiments on 10GB TREC [6] (Text Retrieval Conference [7]) data set. It contains 1.7 million Web pages. We first explain briefly about the preprocessing. Next we discuss about the extraction of DBGs from the link–file. Then, we explain about the community hierarchy results.

4.1 Preprocessing

The preprocessing is done through the following steps (for the details see [1]): extracting all the links, eliminating the duplicates and removing the both popular and unpopular pages.

The pages are in the text format with html marking information. We extracted links by ignoring the text information. 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 **link-file** is prepared that contains the edges of all the pages in a data set. 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 [8] 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 a 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 the extreme popular and unpopular pages: the popular pages are those which are highly referred in the Web such as WWW.yahoo.com and 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 the number of parents greater than fifty and less than two are removed. Also, we removed the pages with one child by considering that these do not contribute to community formation. So, after sorting based on the source, the pages that have the 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.

We use the link-file to extract the communities.

4.2 DBG extraction results

For each source page in the link-file (by sorting based on the source), we first extracted the related pages during the gathring phase to extract a DBG. Figure 4 shows how the number of pages in rel_set increases during the gathering phase with the number of iterations (the variable max_iter in the algorithm). This figure represents the trend of four URLs that are selected randomly. It can be observed that the number of pages grow exponentially (y-axis is as per the log scale) with the number of iterations. If the value of max_iter exceeds three, the number of pages explode.

It can be noted that if *max_iter* exceeds one, a fairly large number of pages are extracted in *rel_set*. Also, it was observed that, the pages in the corresponding DBG structures are too loosely related. So, we extracted communities by restricting *max_iter* to one.

The total number of pages that constitute link-file is around 0.7 million. Figure 5 shows the number of DBG(F,C, p_t, q_t) structures extracted from link-file at the different values of p_t and q_t (no duplicate elimination). Note that, for these results, the rcf factor is not taken into account; the proposed algorithm is applied without the second constraint in Definition 4 (without the step 2(d) in the algorithm). For a DBG(F,C, p_t, q_t), the third column indicates the average number of pages in F and C. The results demonstrate that the proposed approach extracts a fairly big DBG (community) structures. (Figures 2 and 7 shows a sample DBGs.)

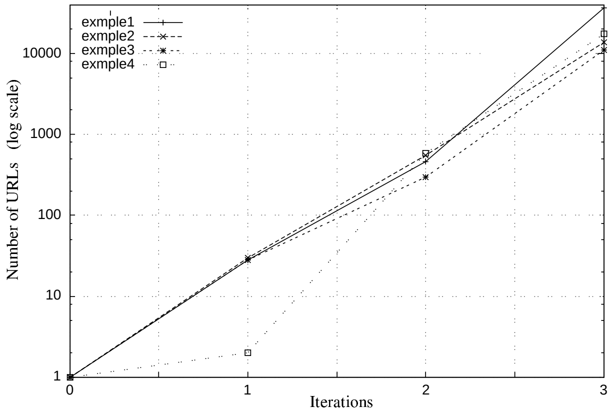

Figure 4. Expansion of pages during the gathering phase.

(p_t, q_t)	**# of DBG(F, C, p_t, q_t)**	**(avg(F), avg(C))**
(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 5. DBG extraction results.

4.3 Community hierarchy results

In this experiment, the threshold values (see Definition 4) for p_t and q_t are fixed at three after observing a few DBG structures. If the value of p_t as well as q_t is less than three, it was observed that the pages in F are not related (or meaningful) for the most of the DBGs. The value of rcf is fixed at 5, intutively. We currently investigating for methods to determine the appropriate rcf value.

For all the levels we fix p_t=3, q_t=3, rcf = 5, and *max_iter*=1. From the link-file, we extracted 67698 DBG(F,C,3,3) structures of 1-level. Each DBG(F,C,p,q) structure is considered as a community (with unique identifier) and its members are the members 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 one community are equal to or the subset of the members of the other community. It was found out that most of them were duplicates. After eliminating the duplicates the number of communities at 1-level comes to 15857.

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

We repeated the process and extracted 3-level communities from the 2-level communities. The number of 3-level communities comes to 1734. After eliminating the duplicates, the number comes to 332. The results are summarized in Figure 6.

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

Figure 6. Community hierarchy statistics

The community hierarchy results are displayed at [9]. For each community, the potential key words are displayed. One can understand 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 the 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 3-level community is represented with the 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 7 shows a DBG(8, 13, 3, 3) structure. Through the title text, it can be observed that the pages in F are related to the topic *telecommunications.*

  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/

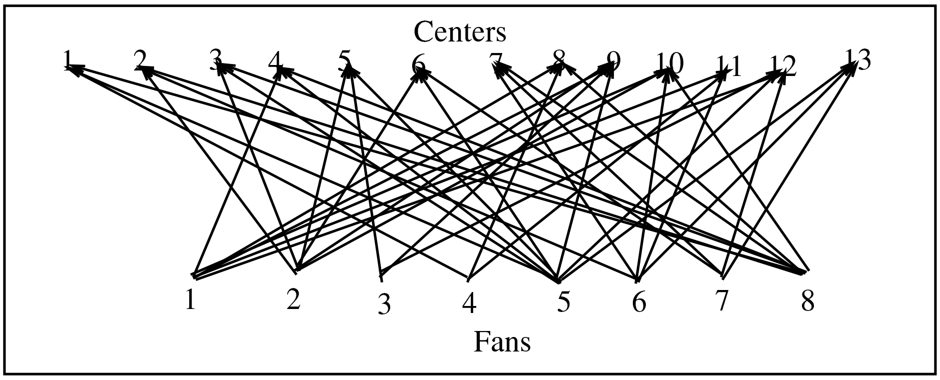

Figure 7. A 1-level community structure: a DBG(8,13,3,3). The Fans are related to telecommunications. (Note that an arrow mark from a fan to a center depicts a hyperlink.)

- **A 2-level community; topic: Nutrition and food habits**

  Key words: 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 key words of 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**

  Keywords: 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 key words of 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 [10] 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 the association among the pages 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 [11], where network properties such as cliques, centroids, and diameters are used to analyze the collective properties of interacting agents. The fields of citation analysis [12, 13, 14] and bibliometrics [15, 16] also use citation links between the works of literature to identify the 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 [17, 18, 19, 20, 21, 22, 23] to find related information from the Web. Chakrabarti [24] provides a survey of the research works in the area of hypertext mining.

Community detection:

The principle component of the cocitation analysis measures the number of documents that have cited a given pair of documents together. In [25], a cocitation analysis technique [26] has been extended to cluster the Web pages by considering that a hyperlink provides a semantic linkages between the pages in the same manner that citations link documents to other related documents. It has been shown that citation analysis shown to create a better formed and more meaningful clusters of documents.

In [27], communities have been analyzed which are found based on the topic supplied by the user by analyzing the link topology using the HITS (Hyperlink-Induced Topic Search) algorithm [22]. 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). After extracting a community signature, the real community can be extracted using the HITS algorithm [22]. A CBG abstraction extracts a small set of potential pages that have common links. 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 which 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 a link to all the pages of interest in particular domain. Because, a data set may not contain the potential pages to form a CBG. In this paper we focused on the issue of extracting the community structures based on the DBG abstraction and then relating them. Since we could extract a fairly big community (or DBG) structures, a high-level communities are being extracted from these structures without resorting to a community expansion phase using HITS as in [1]. Note that the abstraction of a community structure through a DBG matches well with the real community structures. Because, in a DBG(F,C, p,q), each node in F is allowed to form an edge with a few other nodes of C, in a similar manner as a member in a community forms a relationship with a few other members. This differs from a CBG(F,C,p,q), in which each node in F is forced to form an edge with all the nodes of F.

In [28], given a set of the 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 [29], the Companion algorithm is proposed to find the related pages of the seed pages presented by specializing the HITS algorithm exploiting the weight of the links and the order of the 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 [30], the companion algorithm is extended to find the related communities by exploiting the derivation relationships between the 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 the community structures by mathematically abstracting a community as a DBG over a set of pages. Also, a high-level community is abstracted as a DBG over a set of low-level communities. Using the proposed approach, a community hierarchy can be constructed for the given data set that generalizes a large number of low-level communities into a few high-level communities. From the experiment results on the TREC data set show that the proposed approach extracts a fairly big community structures. Also, a 3-level community hierarchy is being constructed for the TREC data set. It displays connections among the communities and can be be used to find the interesting communities by hand, reorganize the Web sites, and understand the sociology of the Web.

As a part of future work we planning to investigate the techniques to improve the community definition and analyze the connections among the extracted communities. Also, we will investigate how the text information improves the community extraction results.

In general a community is a macro phenomena manifested by complex relationships exhibited by corresponding members. At a micro level, each member establishes a relationship with a few other members of the same community. Integration of all the members and their relationships with other members exhibits a community phenomena. Note that the abstraction of a community structure through a DBG matches well with the real community structures. Because, each node in a DBG forms an edge with a few other nodes similar to a member in a community. In general, given any data set that consists of nodes and association information among the nodes, the proposed approach can be used to extract the interesting groups.

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. of 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] R.Baeza-Yates and B.Ribeiro-Neto, Modern information retrieval, Addison Wesley, 1999.

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

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

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

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

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

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

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

[13] 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.

[14] 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.

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

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

[17] 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.

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

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

[20] 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.

[21] 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.

[22] J.Kleinberg, Authoritative sources in a hyperlinked environment, in proc. of ACM-SIAM Symposium on Discrete Algorithms, 1998.

[23] K.Bharat and M.Henzinger, Improved algorithms for topic distillation in hyperlinked environments, in proc. of 21st SIGIR, 1998.

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

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

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

[27] 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.

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

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

[30] 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.