Mining Communities on the Web Using a Max-Flow and a Site-Oriented Framework

Yasuhito Asano[1], Takao Nishizeki[1], and Masashi Toyoda[2]

[1] Graduate School of Information Sciences, Tohoku University, Aza-Aoba 6-6-05, Aramaki, Aoba-ku, Sendai, 980-8579, Japan,
asano@nishizeki.ecei.tohoku.ac.jp, nishi@ecei.tohoku.ac.jp
[2] Institute of Industrial Science, the Univeristy of Tokyo, Komaba 4-6-1, Meguro-ku, Tokyo, 153-8505, Japan, toyoda@tkl.iis.u-tokyo.ac.jp

Abstract. There are several methods for mining communities on the Web using hyperlinks. One of the well-known ones is a max-flow based method proposed by Flake *et al.* The method adopts a page-oriented framework, that is, it uses a page on the Web as a unit of information, like other methods including HITS and trawling. Recently, Asano *et al.* built a site-oriented framework which uses a site as a unit of information, and they experimentally showed that trawling on the site-oriented framework often outputs significantly better communities than trawling on the page-oriented framework. However, it has not been known whether the site-oriented framework is effective in mining communities through the max-flow based method. In this paper, we first point out several problems of the max-flow based method, mainly owing to the page-oriented framework, and then propose solutions to the problems by utilizing several advantages of the site-oriented framework. Computational experiments reveal that our max-flow based method on the site-oriented framework is significantly effective in mining communities, related to the topics of given pages, in comparison with the original max-flow based method on the page-oriented framework.

1 Introduction

A Web community is a set of sites or pages related to a topic. Several methods of mining communities have been recently developed by utilizing a graph structure of hyperlinks on the Web. The idea behind the methods is the following observation: if there is a link (u, v) from pages u to v, then v is considered to contain valuable information for the author of u. For example, HITS proposed by Kleinberg [10] regards a Web page as an *authority* of communities if it is linked from many pages, and as a *hub* if it has many links to authorities. Two of the other well-known methods are trawling proposed by Kumar *et al.* [11] and a max-flow based method proposed by Flake *et al.* [8]. The latter, in particular, is known as a useful method of mining communities related to the topics of given pages. Most of the methods of mining communities using hyperlinks [9], [14] adopt the following *page-oriented framework*:

(1) Collect data of URLs and links in Web pages. For example, the max-flow based method collects URLs and links of pages which are within two links from given pages.
(2) Construct from the data a Web graph $G = (V, E)$, where V is a set of Web pages and E is a set of links between pages in V.
(3) Find some characteristic subgraphs of the Web graph G as approximate communities. For example, the max-flow based method finds a dense subgraph as an approximate community.

It is often appropriate to regard a site, in place of a page, as a unit of information in the real Web. In fact, a *link* (u, v) *between sites*, i.e. a link (u, v) such that pages u and v are in the different sites, is created mostly as a reference to valuable information, while a *link* (u, v) *inside a site*, i.e. a link (u, v) such that both pages u and v are in the same site, may be created merely for convenience of Web browsing. However, the page-oriented framework deals with both kinds of links equally, and has some disadvantages for properly mining communities. Thus, Asano *et al.* [3], [4] proposed another framework for mining communities, named a *site-oriented framework*, as follows.

(1) Collect data of URLs and links in Web pages.
(2) Find "directory-based sites" by using the data.
(3) Construct an *inter-site graph*, whose vertices are directory-based sites, and whose edges are links between directory-based sites.
(4) Find some characteristic subgraphs in the inter-site graph as approximate communities.

This framework regards a site, instead of a page, as a unit of information, and is expected to be more useful for properly mining communities than the page-oriented framework. They implemented this framework by proposing a new model of a site, called a *directory-based site*, and by giving a method for finding directory-based sites from data of URLs and links. Asano *et al.* used data sets of `.jp` domain URLs crawled in 2000 and 2002 by Toyoda and Kitsuregawa [13], and verified that the method can correctly find directory-based sites in most cases. For the data sets above, they also showed [3], [4] that trawling on the site-oriented framework can find a number of communities which can never be found by trawling on the page-oriented framework. Thus, the site-oriented framework is effective in mining communities through trawling. However, it has not been known that whether the site-oriented framework is effective in mining communities through the max-flow based method.

In this paper, we show that the site-oriented framework is significantly effective in mining communities through the max-flow based method in regard to both quality and quantity. The *quantity* of a found approximate community is measured by the number of sites that are truly related to the topics of given pages. The *quality* is measured by the ratio of the number of related sites to the total number of sites in the community.

The rest of this paper is organized as follows. In Section 2, we overview the results obtained by computational experiments. In Section 3, we introduce the

site-oriented framework proposed in [3], [4] In Section 4, we outline the max-flow based method proposed by Flake *et al.* In Section 5, we point out several problems of the max-flow based method on the page-oriented framework, and propose solutions to them by utilizing the site-oriented framework. We also analyze how the problems and the solutions affect the performance of the method. In Section 6, we present our concluding remarks.

2 Overview of Experimental Results

We implement our improved max-flow based method, and evaluate by computational experiments the performance of two methods: our method based on the site-oriented framework, and the original max-flow based method based on the page-oriented framework.

Table 1. Quality and quantity of found communities.

Topic	1	2	3	4	5	6	7	8
Page, original	0/2	0/12	0/5	3/126	7/29	2/4	8/12	1/5
Site, improved	17/26	16/28	6/11	4/9	15/23	36/42	7/7	11/21

Topic	9	10	11	12	13	14	15	16
Page, original	7/11	7/10	1/3	10/17	11/15	37/52	8/8	3/3
Site, improved	18/21	8/9	13/25	15/20	24/28	4/5	7/7	16/23

Table 1 depicts the results of the experiments using the data set collected in 2003. The topics indexed 1 to 16 are "Studio Ziburi (Japanese famous animation studio)," "Tactics Ouga (computer game)," "dagashi-ya (Japanese traditional papa-mama shops)," "The Lord of the Rings (movie and novel)," "swords," "cats," "Chinese history," "Spitz (Japanese singer)," "guns," "dogs," "army," "Hirohiko Araki (Japanese comic artist)," "train photos," "Monopoly (board game) clubs," "beetles," and "mountain photos," respectively. For each topic, we select three or four seed pages from the search results of the topic by Google. As seed pages, we pick up pages in personal sites on servers of ISPs (Internet Service Providers), servers of universities or rental Web servers as much as possible so that, for every topic, at most one seed page is not a page in such sites. In Table 1, "Page, original", abbreviated to PO, denotes the original max-flow based method on the page-oriented framework, while "Site, improved", abbreviated to SI, denotes our improved method on the site-oriented framework. "Page, original" and "Site, improved" rows show the results of applying PO and SI to the seed pages for the sixteen topics, respectively. These two rows consist of sixteen cells, each corresponding to a topic. Each cell contains a fraction; the denominator is the number of the sites in the found approximate community other than the seeds; and the numerator is the number of the sites in the community

that are truly related to the topic of the seeds. In other words, the numerator represents the quantity of the found community, and the fraction does the quality.

For example, we use three seed pages for topic 5, "swords." PO finds 29 sites other than the seeds, but only seven of them are truly related to the topics of the seeds. On the other hand, SI finds 23 sites, and 15 of them are truly related. Thus, the original method PO finds a community of quantity 7 and quality 7/29, while our method SI finds a community of quantity 15 and quality 15/23. It should be noted that the community found by SI contains nine personal sites in servers of ISPs, servers of universities or rental Web servers; these nine sites cannot be found by PO, although all the nine sites contain valuable information about the topic "swords."

The average number 13.6 of related sites found by SI is about 2.07 times as many as the average number 6.56 of related pages found by PO. The average quality 72.6% of the communities found by SI is significantly greater than the average quality 45.8% of those found by PO. In fact, SI achieves strictly better quality than PO for fourteen topics among all the sixteen topics. It still has relatively good quality for the remaining two topics 15 and 16; 100% and 69.6%, respectively. For topic 16, SI finds sixteen related sites, five times more than PO. As we will explain in Section 5, SI adopts several solutions utilizing some advantages of the site-oriented framework. Thus the site-oriented framework is fairly effective for improving the max-flow method both in quality and in quantity. Further analyses will be described in Section 5.

Table 2. The sizes of used graphs.

Topic	1	2	3	4	5	6
Page	345, 1422	687, 2518	160, 610	1472, 9020	2216, 7550	385, 1484
Site	2109, 5050	2605, 6216	1969, 4678	489, 1154	2143, 5208	4854, 11580

Topic	7	8	9	10	11	12
Page	1055, 4822	484, 1826	416, 1768	325, 1210	1987, 7584	1745, 7394
Site	1957, 4510	7602, 18436	1595, 3750	3196, 7642	5969, 14614	6497, 15942

Topic	13	14	15	16
Page	466, 2110	745, 3404	877, 4074	744, 3086
Site	590, 1472	2187, 5670	3116, 7302	2164, 5048

Table 2 shows the sizes of the Web graphs and the inter-site graphs used for the topics in the experiments. In each cell, the left number denotes the number of vertices, while the right number denotes the number of edges. For example, for topic 5, the number of vertices and edges in the Web graph used by PO is 2,216 and 7,550, respectively; the number of vertices and edges in the inter-site graph used by SI is 2,143 and 5,208, respectively.

3 Implementation of Site-Oriented Framework

It is easy to propose a site-oriented framework, but is very difficult to implement it, because a site is a vague concept. Asano *et al.* [4], [5] have implemented it by using a directory-based site model and a method of identifying directory-based sites from data of URLs and links. In this section, we outline the implementation, and compare the sizes of a Web graph and an inter-site graph obtained from a data set of 2003, a new one different from the data sets used in [4], [5].

3.1 Directory-based Sites

The phrase "Web site" is usually used ambiguously, and it is difficult to present a proper definition of a Web site.

Several methods including the max-flow based method, HITS, and trawling, have adopted an idea to use a Web server, instead of a site, as a unit of information. This idea works relatively well for a *single-site server*, which is a Web server containing a single site as in the case of official sites made by companies, governments or other organizations. On the other hand, the idea works poorly for a *multi-site server*, which is a Web server containing multiple sites. For example, a Web server of an ISP, a university's server or a rental Web server may contain a number of personal Web sites. Thus, adopting this idea would lose valuable information, since information about relatively minor or specialized topics is frequently held in such personal sites.

Li *et al.* proposed a *logical domain* as a unit of information smaller than a site [12]. However, the logical domain is used for clustering documents but not for mining communities.

Several researches [2], [6] and [7] defined a Web site as a set of Web pages that are written by a single person, company, or cohesive group, although they did not actually propose a method of finding sites on the Web according to their definitions. If every Web page contained information about its author, then one could identify a Web site according to this definition. Unfortunately, most pages contain no information about their authors, and hence it is difficult to identify Web sites according to this definition. Thus, we need a method of finding Web sites approximately according to some proper model.

Asano *et al.* [3], [4] proposed a new model of a site, called a *directory-based site*, and obtained a method to examine whether a given server is a single-site server or a multi-site server and to find a set of directories corresponding to the top directories of users' sites. Using their method, we have found 6,063,674 directory-based sites in 2,940,591 servers for the data set of 2003. The error rate for the results is only 1.9%, which has been estimated by random sampling.

3.2 Inter-Site Graph

Once the directory-based sites have been found for a given data set, we can construct an inter-site graph G, as follows:
(1) If there is a link from a page v in a directory-based site A to a page w in

another directory-based site B, we add to G a link from A to B, called a *global-link*.
(2) An *inter-site graph* G is a directed graph, whose vertices correspond to directory-based sites and whose edges correspond to global-links. A pair of oppositely directed global-links is called a *mutual-link*.

For the data sets of 2003, the numbers of vertices and edges of the Web (page) graph are 312,536,723 and 1,032,820,388, and the numbers of those of the inter-site graph are 6,063,674 and 43,197,249. The number of the edges of the inter-site graph is less than 5% of the number of the edges of the Web graph. One can thus deal with the inter-site graph more easily than the Web graph for mining communities especially in memory space.

4 Summary of Max-Flow Based Method

The max-flow based method proposed by Flake *et al.* [8] finds an approximate community from a Web graph composed of seed pages and their neighbor pages. The method finds a dense subgraph containing the seed pages, and regards it as a community, as outlined as follows.

1. Construct a Web graph $G = (V, E)$. The vertex set V is a union of three sets S, P and Q of pages; S is the set of seed pages given by a searcher, P is the set of all pages that either link to or are linked from pages in S, and Q is the set of all pages that are linked from pages in P. The edge set E consists of all the links from pages in S to pages in $S \cup P$ and all the links from pages in P to pages in $S \cup P \cup Q$. In Figure 1 all edges in E are drawn by solid lines.
2. Add to $G = (V, E)$ a virtual source s and a virtual sink t, and add to G a virtual edge (s, x) for each vertex $x \in S$ and a virtual edge (x, t) for each vertex $x \in V$. Let $G'_1 = (V', E')$ be the resulting network. Thus $V' = V \cup \{s\} \cup \{t\}$ and $E' = E \cup \{(s, x) \mid x \in S\} \cup \{(x, t) \mid x \in V\}$. In Figure 1 some of the virtual edges are drawn by dotted lines. The capacity $c(e)$ for each edge $e \in E'$ is given as follows:
   - $c(e) = k$ for each edge $e \in E$, where $k(> 1)$ is a given integer parameter, and is equal to the number of the seed pages in [8]. (Usually k is fairly small, say $2 \leq k \leq 4$.)
   - $c(e) = \infty$ for each edge $e = (s, x)$.
   - $c(e) = 1$ for each edge $e = (x, t)$.
3. For $i = 1$ to ℓ, do the following procedures (a)-(c), where ℓ is a given integer parameter. (Flake *et al.* usually set $\ell = 4$.)
   (a) Compute the min-cut of network G'_i, that separates s and t and is nearest to s, by finding a max-flow for G'_i. Let $E(X, V' \setminus X)$ be the set of edges in the min-cut, where $X \subseteq V'$ is a vertex set containing s.
   (b) If $i < \ell$, then find a vertex u of maximum degree in $X \setminus (S \cup \{s\})$, i.e. a vertex u such that $d(u) \geq d(u')$ for every vertex $u' \in X \setminus (S \cup \{s\})$, where $d(x)$ denotes the degree of vertex $x \in V'$ in G'_i, that is, $d(x)$ is

the sum of in-degree and out-degree of x. Regard u as a new seed, set $S := S \cup \{u\}$, and construct a new network G'_{i+1} for the new set S as in Steps 1 and 2. Increment i by one.

(c) Otherwise, i.e. $i = \ell$, output $X \setminus \{s\}$ as an approximate community.

Note that Flake *et al.* have adopted the following three policies in Step 1 [8].

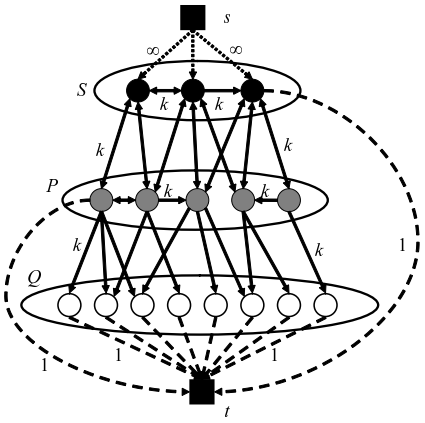

Fig. 1. A Web graph G and network G'_1.

1. Regard every page with relatively large degree, say 50 or more, as a portal, and ignore it.
2. Every edge joining a vertex in S and a vertex in P is regarded as a bi-directed edge in G even if there is only a uni-directional link between the two pages corresponding to the vertices. Without this policy, some pages that link to but are not linked from pages in S cannot be reached from the virtual source and consequently cannot be members of the found community although they are sometimes related to the seed pages on the real Web.
3. Ignore all the links between pages in the same server. This policy will be further discussed in the next section.

5 Problems of Max-Flow Based Method and Solutions to them

In this section, we first point out four problems of the max-flow based method, which mainly owe to the page-oriented framework, and then propose solutions to them by utilizing the site-oriented framework. We also analyze how the problems and the solutions affect the performance of the max-flow method. In subsections 5.1-5.4, we describe these four problems, which we call the *ignored link problem*, the *mutual-link problem*, the *missing link problem*, and the *capacity problem*, respectively.

5.1 Ignored Link Problem

The original max-flow based method adopts the policy of ignoring links in the same server. Consider a seed page belonging to an official site of companies, governments or other organizations. Such an official site usually consists of the whole pages in a Web server, and many links in the site are created merely

for convenience of the reader of the site. Thus, a Web graph G would contain a dense subgraph consisting of pages in the official site if the max-flow based method on the page-oriented framework did not adopt the policy of ignoring links in the same server. Thus the quantity of the found community, corresponding to the dense subgraph, is only one. Hence, the policy is necessary for finding a community with large quantity by the max-flow based method on the page-oriented framework. However, the policy suffers from the following two problems.

Ignored Link Problem (1) Figure 2 illustrates two sites A and B; although there is a global-link going from A to B in an inter-site graph, the seed page $a \in A$ is isolated in a Web graph G since a link (a, b) is ignored and hence a link (b, c) cannot be found. Such a situation often occurs when a seed page is the top page of a site and all links to other sites are placed on other pages in this site. In such a case, the information of links from this site to the other sites is lost although it may be valuable for mining communities related to this site. There are a number of such cases in the real Web, particularly in personal sites in multi-site servers, and hence the lack of such information would be a serious problem. We call this problem the *ignored link problem (1)*.

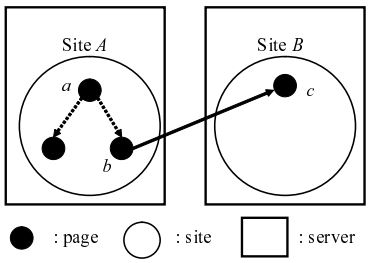

Fig. 2. Illustration for the ignored link problem (1).

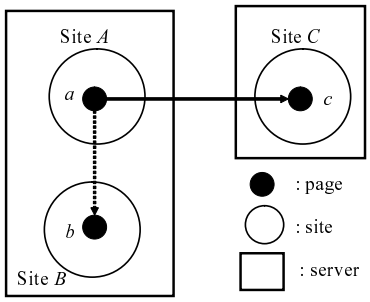

Fig. 3. Illustration for the ignored link problem (2).

Ignored Link Problem (2) If we adopt the policy of ignoring links in the same server, then we cannot utilize information of links between sites in the same server. Figure 3 illustrates three sites A, B and C, among which there are two global-links (A, B) and (A, C). The global-link (A, C) can be found but (A, B) cannot be found, because A and B are contained in the same server. In the real Web there are a number of multi-site servers such as rental Web servers, ISPs' servers and universities' servers, and hence the lack of such information would be a serious problem in mining communities. We call this problem the *ignored link problem (2)*.

Table 3 shows the number of ignored links, the number of edges in Web graphs, and their ratio for each of the sixteen topics used in the experiments described in Section 2. The average ratio 60.6% is fairly large, and ignored links

Table 3. The number of ignored links.

Topic	1	2	3	4	5	6	7	8
Ignored Links	706	793	1081	8265	5182	1361	1690	1271
Edges	1422	2518	610	9020	7550	1484	4822	1826
Ratio (%)	49.6	31.5	177.2	91.6	68.6	91.7	35.0	69.6

Topic	9	10	11	12	13	14	15	16
Ignored Links	75	855	2333	3933	2826	807	1577	1196
Edges	1768	1210	3086	7584	7394	2110	3404	4074
Ratio (%)	4.2	70.7	75.6	51.9	38.2	38.2	46.3	29.4

are frequently found around personal sites in multi-site servers. Thus, within the page-oriented framework, the loss of information due to ignored links would cause the poor quality of communities especially when given pages belong to personal sites in multi-site servers.

The site-oriented framework gives a natural solution to the ignored link problems. The site-oriented framework does not need the policy of ignoring links in the same server, since there is no links in the same site in an inter-site graph. The ignored link problem (2) is also solved by using an inter-site graph instead of a Web graph, because a link between sites in the same server appears as a global-link in an inter-site graph.

5.2 Mutual-Link Problem

A mutual-link between two sites is frequently created when these sites are related and the authors of the sites know each other site, and hence a mutual-link is more useful for mining communities than a single-sided link. Asano *et al.* have indeed proposed a method of mining communities by enumerating maximal cliques composed of mutual-links, and have verified that the method is more suitable for mining communities of personal sites than trawling [4], [5]. However, the original max-flow based method cannot correctly find mutual-links due to the following two reasons (1) and (2), illustrated for the example in Figure 4.

(1) Since the max-flow based method adopts the policy of ignoring links in the same server, all the links drawn by dotted lines in Figure 4 are ignored. Thus, even if both the top pages $a \in A$ and $c \in B$ are given, the link from $b \in A$ to c cannot be found, and hence the mutual-link between A and B cannot be found.
(2) In Figure 4, there are links from b to c and from c to a, but there is no pair of pages mutually linked. Thus, there is no mutual-link between two pages in a Web graph composed of all the pages and links in Figure 4.

In the real Web there are a number of situations similar to the example above, and thus it would be a serious problem that such mutual-links cannot be found.

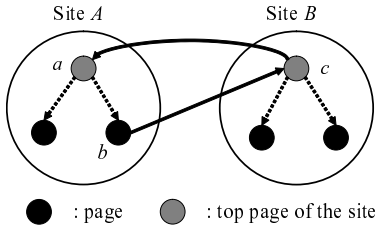

Fig. 4. Illustration for the mutual-link problem.

Moreover, the max-flow based method cannot fully utilize the found mutual-links, as follows.

(I) The method regards all the links emanating from seeds as bi-directed edges even if they are single-sided links. Hence, the mutual-links and the single-sided links emanating from seeds are equally treated in the max-flow based method.

(II) Most of the "path flows" [1] of the max flow in a network G' pass through vertices in the following order: vertices in S, vertices in P, and vertices in Q. That is, most of the path flows go along an edge from S to P and along an edge from P to Q, while very few of the path flows go along an edge from Q to P and along an edge from P to S. Thus, even we use a bi-directed edge to represent a mutual-link and use a uni-directed edge to represent a single-sided link in order to make a difference between a mutual-link and a single-sided link, the mutual-link would not contribute to finding good approximate communities.

Thus, the original max-flow based method on the page-oriented framework can neither find all mutual-links nor fully utilize the found mutual-links. We call this problem the *mutual-link problem*.

We now propose a solution to the mutual-link problem using the site-oriented framework. First, we construct an inter-site graph as G from a given set of seed sites and use a site, instead of a page, in the max-flow based method described in Section 4. Second, when we construct the last network G'_ℓ, let R be the set of vertices connected to one or more seed sites by mutual-links, and we add all the sites in R to the given set of seed sites, where ℓ is the parameter in Step 3 of the method. Some preliminary experiments reveal that adding the sites in R to the set of seed sites when constructing intermediate networks G_i, $1 \leq i \leq \ell - 1$, does not contribute to finding good approximate communities in quality. The solution above is based on the observation in [4], [5] that a mutual-link between two sites represents a strong relation between them.

Table 4 shows the numbers of mutual-links in the graphs used for the experiments in Section 2. For each topic, "Page" row shows the number of mutual-links in the Web graph, while "Site" row shows the number of mutual-links in

Table 4. The numbers of mutual-links.

Topic	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
Page	0	0	1	9	0	9	7	5	4	0	7	8	5	3	8	10
Site	12	19	10	0	8	14	7	17	5	19	10	28	15	3	18	11

the inter-site graphs. The average number 12.3 of mutual-links in an inter-site graph is about 2.58 times as many as the average number 4.75 of mutual-links in a Web graph, and hence the site-oriented framework is suitable for mining communities using mutual-links. Only for topic 4 "The Lord of the Rings," the Web graph has more mutual-links than the inter-site graph, although the quality of the community found by PO is worse than that by SI, as described in Table 1. In the Web graph, some of the seed pages for topic 4 have links to pages concerning a well-known HTML grammer checker *htmllint*, and all the mutual-links in the Web graph join two of the pages on *htmllint*. Hence, even if a page becomes a member of the community owing to the mutual-links, the page is related to *htmllint* but not to topic 4, and hence these mutual-links do not contribute to the quality of the community for topic 4.

For the topics other than topic 4, although the number of mutual-links is fairly small compared with the number of edges, their effect for mining communities is not small on the site-oriented framework. For example, for topic 6 "cats", the following five sites in the community found by SI would not be found without the solution above: `homepage2.nifty.com/yazu/`, `www10.ocn.ne.jp/~ne-help/`, `www5e.biglobe.ne.jp/~moyochu-/`, `www.kuma-kuma.net/~kylie/`, and `www.kuma-kuma.net/~nekoneko/`. Each of these five sites is a personal site in a multi-site server, and contains attractive contents about cats, and thus our solution based on the site-oriented framework is effective for mining communities containing personal sites in multi-site servers.

5.3 Missing Link Problem

In this paper, a link from pages a to b is called a *missing link* if a corresponds to a vertex in set Q and b corresponds to a vertex in set $P \cup Q$. The original max-flow based method ignores all these missing links in the construction step of a Web graph $G = (V, E)$. However, ignoring the missing links would lose valuable information and consequently be a serious problem for mining communities. We call this problem the *missing link problem*.

Using experiments with the same inputs described in Section 2, we have verified the fact that the average number of missing links for the site-oriented framework is larger than the average number of those for the page-oriented framework.

Table 5 shows the numbers of missing links in the experiments. For each topic, "Page" row shows the number of missing links for the page-oriented framework, while "Site" row shows the number of missing links for the site-oriented framework. The average number 93.8 of missing links for the site-oriented framework

Table 5. The numbers of missing links.

Topic	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
Page	18	24	2	80	34	56	88	34	46	6	80	88	132	30	210	138
Site	26	42	142	6	34	134	40	200	34	86	20	190	170	12	284	80

is about 1.41 times as many as the average number 66.6 of missing links for the page-oriented framework.

The fact implies that using missing links on the page-oriented framework would not much affect finding good approximate communities, but utilizing missing links on the site-oriented framework would be useful for mining communities. Thus, in order to solve the missing link problem, we include the missing links in a graph G on the site-oriented framework.

For example, for topic 6 "cats", the following two sites in the community found by SI could not be found without the solution above: `www.kt.rim.or.jp/~taya/`, and `www.iris.dti.ne.jp/~mya-/`. Each of these two sites is a personal site in a multi-site server, and contains attractive contents about cats, and thus our solution based on the site-oriented framework is effective for mining communities containing personal sites in multi-site servers.

5.4 Capacity Problem

Let $G' = (V', E')$ be a network used to find an approximate community. Let C be the set of vertices corresponding to the approximate community. Let $(X, V' \setminus X)$ be the min-cut of G' corresponding to C, where $s \in X \subseteq V'$. Thus $C = X \setminus \{s\}$. For a vertex v, we denote by $d^+(v)$ and $d^-(v)$ the out-degree and in-degree of v in G', respectively. We consider a situation in which the following condition (1) hold:

(1) G' has no edge (u, v) for any pair of vertices u and v such that $u, v \in P$ or $u, v \in Q$.

We then claim that the following (A) and (B) hold for the approximate community found by the original max-flow based method.

(A) A vertex $v \in P$ is contained in C if $d^-(v) > 1/k + d^+(v) - 1$, where k is the given parameter in Step 2 of the method.
(B) A vertex $v \in Q$ is contained in C if there is an edge $(u, v) \in E'$ for some vertex $u \in P \cap C$.

Proof of (A): Suppose that $d^-(v) > 1/k + d^+(v) - 1$ for a vertex $v \in P$. Then the sum $k \cdot d^-(v)$ of capacities of the edges entering v is larger than the sum $1 + k(d^+(v) - 1)$ of capacities of the edges emanating from v. The capacity of every edge connecting the virtual source s and a vertex in S is infinite. Therefore, v can be reached from the virtual source s via a vertex in S in the "residual network" [1] of G' for the max-flow, and hence v is contained in C. □

Proof of (B): Suppose that there is an edge $(u, v) \in E'$ for some vertex $u \in P \cap C$. Let $P'(v) = \{u' \in P \cap C \mid (u', v) \in E'\}$, then $P'(v) \neq \emptyset$. Since G' satisfies the condition (1), vertex $v \in Q$ has excatly one outgoing edge (v, t), and $c(v, t) = 1$. On the other hand, the total amount of capacities of all edges going from vertices in $P'(v)$ to v is more than 1 since $c(u, v) = k > 1$. Thus, edge (u, v) for $u \in P'(v)$ cannot be saturated by any max-flow in G', and hence the vertex v can be reached from the virtual source s through u in the residual network. Thus $v \in C$. □

(A) and (B) above immediately imply the following two facts (I) and (II).

(I) Even if a vertex v in Q has only one incoming edge (u, v), the max-flow based method regards v as a member of the community when $u \in P \cap C$.
(II) Even if a vertex v' in P has incoming edges from all the vertices in S, corresponding to the seed pages, the max-flow based method does not regard v' as a member of the community when $d^+(v') > |S|$.

In the real Web, a page corresponding to a vertex v in Q with only one incoming edge is usually not related to the topic of seed pages, while a page corresponding to a vertex v' in P linked from all the seeds is usually related to the topic. However, (I) and (II) imply that v often becomes a member of a community but v' often becomes a non-member of a community. Thus, (A) and (B) cause a problem for mining communities. This problem is serious, because many graphs used in the max-flow based method satisfy the condition (1); even if they do not satisfy the condition (1), the graphs obtained from them by removing few edges often satisfy the condition (1). We call this problem the *capacity problem*.

To solve the capacity problem, we propose to increase the capacity of every edge joining a vertex in S and a vertex in P. That is, we modify the original capacity function as follows:

- $c(e) = k'$ for every edge $e = (u, v)$, $u \in S$ and $v \in S \cup P$, where $k'(> 1)$ is a new parameter.
- $c(e) = 1$ for every edge $e = (u, v)$, $u \in P$ and $v \in P \cup Q$.
- $c(e) = \infty$ for each edge $e = (s, x)$.
- $c(e) = 1$ for each edge $e = (x, t)$.

We have made some preliminary experiments varying the parameter $k' \in \{3, 5, 10, 15, 20, 25\}$ for the sixteen topics described in Section 2, and confirmed that the method on the site-oriented framework finds approximate communities with good quality when $k' = 10$ or $k' = 15$ for these topics.

6 Concluding Remarks

We have pointed out several problems of the max-flow based method on the page-oriented framework, and proposed solutions to them utilizing the site-oriented framework. Our improved max-flow based method using the solutions on the site-oriented framework has achieved much better results, both in quality and in

quantity, than the original max-flow based method on the page-oriented framework. Our solutions are relatively simple, but are particularly effective for mining communities containing personal sites in multi-site servers. Other methods for mining communities using hyperlinks suffer from the problems described in this paper, and hence the site-oriented framework and the proposed solutions in the paper would be useful for improving such methods.

References

1. R. K. Ahuja, T. L. Magnanti, and J. B. Orlin, "Network Flows – Theory, Algorithms, and Applications," Prentice Hall, New Jersey, 1993.
2. B. Amento, L. G. Terveen, and W. C. Hill, "Does "authority" mean quality? predicting expert quality ratings of web documents," *Proc. 23rd Annual International ACM SIGIR Conference*, pp. 296–303, 2000.
3. Y. Asano, H. Imai, M. Toyoda, and M. Kitsuregawa, "Applying the site information to the information retrieval from the Web," *Proc. 3rd International Conference on Web Information Systems Engineering*, IEEE CS, pp. 83–92, 2002.
4. Y. Asano, "A New Framework for Link-Based Information Retrieval from the Web," Ph.D. Thesis, the University of Tokyo, March 2003.
5. Y. Asano, H. Imai, M. Toyoda, and M. Kitsuregawa, "Finding neighbor communities in the Web using an inter-site graph," *Proc. 14th International Conference on Database and Expert Systems Applications*, LNCS 2736, pp. 558-568, 2003.
6. K. Bharat, B. W. Chang, M. Henzinger, and M. Ruhl, "Who links to whom: mining linkage between Web Sites," *Proc. 1st IEEE International Conference on Data Mining*, pp. 51–58, 2001.
7. N. Craswell, D. Hawking, and S. Robertson, "Effective site finding using link anchor information," *Proc. 24th Annual International ACM SIGIR Conference*, pp. 250–257, 2001.
8. G. W. Flake, S. Lawrence, and C. L. Giles, "Efficient identification of Web communities," *Proc. 6th ACM SIGKDD KDD2000*, pp. 150–160, 2000.
9. N. Imafuji and M. Kitsuregawa, "Finding Web communities by maximum flow algorithm using well-assigned edge capacity," *IEICE Trans. Special Section on Information Processing Technology for Web Utilization*, Vol. E87-D, No. 2, pp. 407-415, 2004.
10. J. Kleinberg, "Authoritative sources in a hyperlinked environment," *Proc. 9th Annual ACM-SIAM SODA*, pp. 668–677, 1998.
11. R. Kumar, P. Raghavan, S. Rajagopalan, and A. Tomkins, "Trawling the Web for emerging cyber-communities," *Computer Networks*, 31(11-16), pp. 1481–1493, 1999.
12. W. S. Li, N. F. Ayan, O. Kolak, and Q. Vuy, "Constructing multi-granular and topic-focused Web site maps," *Proc. 10th International World Wide Web Conference*, pp. 343–354, 2001.
13. M. Toyoda and M. Kitsuregawa, "Extracting evolution of Web communities from a series of Web archives," *Proc. 14th Conference on Hypertext and Hypermedia (Hypertext 03)*, ACM, pp. 28–37, 2003.
14. X. Wang, Z. Lu and A. Zhou, "Topic exploration and distillation for Web search by a similarity-based analysis," *Proc. 3rd International Conference of Advances in Web-Age Information Management (WAIM 2002)*, LNCS 2419, pp. 316–327, 2002.