PAPER Special Section on Information Processing Technology for Web Utilization

Finding Web Communities by Maximum Flow Algorithm using Well-Assigned Edge Capacities.

Noriko IMAFUJI[†], *Nonmember and* **Masaru KITSUREGAWA**[†], *Member*

SUMMARY A *web community* is a set of web pages that provide resources on a specific topic. Various methods for finding web communities based on link analysis have been proposed in the literature. The method proposed in this paper is based on the method using the maximum flow algorithm proposed in [7], [8]. Our objective of using the maximum flow algorithm is to extract a subgraph which can be recognized as a good web community in the context of the quantity and the quality. This paper first discusses the features of the maximum flow algorithm based method. The previously proposed approach has a problem that a certain graph structure containing noises (i.e., irrelevant pages) is always extracted. This problem is mainly caused by *edge capacities* assigned a constant value. This paper proposes an assignment of variable edge capacities that are based on hub and authority scores obtained from HITS calculation. To examine the effects of our proposed method, we performed experiments using a Japanese archive crawled in February 2002. Our experimental results demonstrate that our proposed method removes noise pages caused by constant edge capacities and improves the quality of web communities.

key words: *web community, web graph, maximum flow algorithm, HITS*

1. Introduction

Regarding web pages as nodes and hyperlinks as edges, the web can be recognized as a directed graph, which is designated as a *web graph*. The proposal in [3] has shown that approximately 92% of the nodes in the web graph are connected, the implication being that most web pages in the web are connected by hyperlinks. A new hyperlink (especially to another site) is typically added to a web page by the author intentionally (as opposed to randomly), which makes it likely that the web pages at both the endpoints of a hyperlink are related in some sense.

A *web community* is a subgraph of the web graph. Incidentally, edges in the web graph (i.e., hyperlinks) contain useful information. Some works focus on extracting a web community by analyzing only the link structure of the web graph so that we can obtain significant amounts of information on a specific topic from the member pages, thereby providing *topic-specific resources* to the users. In our case, finding web community is identical to finding a proper cut that separates a subgraph from the web graph i.e., solving a kind of graph partitioning problem.

Various methods for fining web communities based on link analysis have been proposed in the literature. Two works have been targeting a similar goal to ours i.e., the extraction of web communities as subgraphs separated from the web graph. The work [15] recognized web communities as bipartite graphs that contain at least one complete bipartite graphs and counted web communities existing in the web. Since the density of complete bipartite graph based communities becomes quite high, the community members are only limited to the nodes having rather many links. Subsequently to this work, the other work [7], [8] defined the graph structure of web communities so that the restriction by density was moderate. They recognized web communities as sets of nodes having more links in the communities than the outside and proposed an approach for extracting such web communities using the maximum flow algorithm.

We call a page that is irrelevant to members of the web community a *noise page* (or a *noise*). In other words, if the topic of a member page is not similar to (or same as) that of the web community, the page turns out to be a noise in the context of quality of web community. The previously proposed approach [7], [8] considered only edge capacities assigned a constant value. This assignment caused that a certain graph structure, in which noise pages are highly probably contained, is always extracted. This paper proposes an assignment of variable edge capacities that are based on hub and authority scores obtained from HITS calculation. To examine the effects of our proposed method, we performed experiments using a Japanese archive crawled in February 2002. We report the experimental results and show that our proposed method improves the quality of web community by removing noise pages.

The remainder of the paper is organized as follows. Section 2 provides an overview of related works and reviews the method for finding web communities using the maximum flow algorithm proposed by [7], [8]. Section 3 discusses the features and the problems that the previously proposed method has. In Section 4, we propose a method for making the best use of maximum flow algorithm i.e., an assignment of edge capacities using HITS algorithm. Section 5 reports our performance study. Finally, we conclude in Section 6 with directions for future work.

[†]Institute of Industrial Science, The University of Tokyo, Komaba 4-6-1, Meguro-ku, Tokyo

2. Related works: finding web communities with the maximum flow algorithm

As stated in the previous section, our objective is a topic specific resource discovery based on finding web communities. The method using the maximum flow algorithm for finding web communities was first proposed in [7]. In this section, we review the method as well as the maximum flow algorithm. Then, we discuss the effect of the method in the context of the quantity and the quality.

2.1 Maximum flow algorithm

Let $G = (V, E)$ be a directed graph where V and E are a set of nodes and edges respectively. Denote (u, v) be a directed edge from the node u to v. Suppose s, t be the fixed nodes in V and each edge $(u, v) \in E$ has an assigned edge capacity $c(u, v) \in \mathbf{Z}^+$. The flow from s to t of G is a non negative integer function f defined on E which satisfies the following conditions: $0 \leq f(u, v) \leq c(u, v)$ and $\sum_{(u_i,v)\in E} f(u_i, v) = \sum_{(v,u_o)\in E} f(v, u_o)$ for any $v \in V$. The latter condition implies the total flow out of s equals to the total flow into t. The value of the total flow is called a graph flow of G. The maximum flow algorithm finds the maximum graph flow.

Suppose $T = V \backslash S$ for a subset $S \subseteq V$. If $s \in S$, $t \in T$, then the edge set $\{(u, v) \in E | u \in S, v \in T\}$ is called a $s - t$ cut (or a cut) on a graph G, and an edge included in the cut set is called a cut edge. A minimum cut is a cut which minimizes the total capacities of the cut edges, and a maximum graph flow equals to the value of a minimum cut [13].

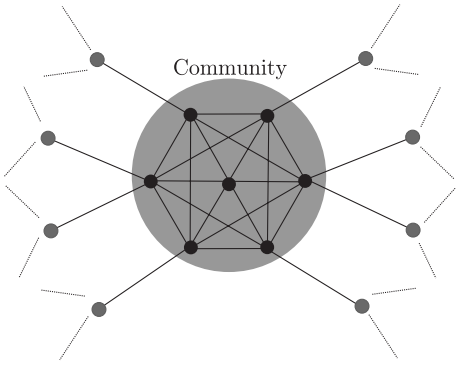

Fig. 1 An example of a web community defined in [7], [8]. Note that every node in the community connects with more nodes in the community than the outside.

2.2 Overview

A web community was defined in [7] as follows: Let V be a set of nodes. A community is a subset $C \subset V$ such that for any $v \in C$, the number of edges between v and the nodes in C is more than that of edges between v and the nodes in $V - C$ (Figure 1).

The definition can be expressed in different words as follows; a web community is a set of web pages having more links within the community than the outside. Figure 1 shows a simple example of a web community in the way of their definition. Every nodes in the circular frame has more links with the nodes in the frame than the outside, which satisfies their definition of a web community. The paper [7] has proved a theorem that after the $s-t$ maximum flow algorithm performed, all nodes accessible from s through unsaturated edges satisfy the definition of a web community. Note that the unsaturated edges mean the edges that the flow does not saturate the edge capacities.

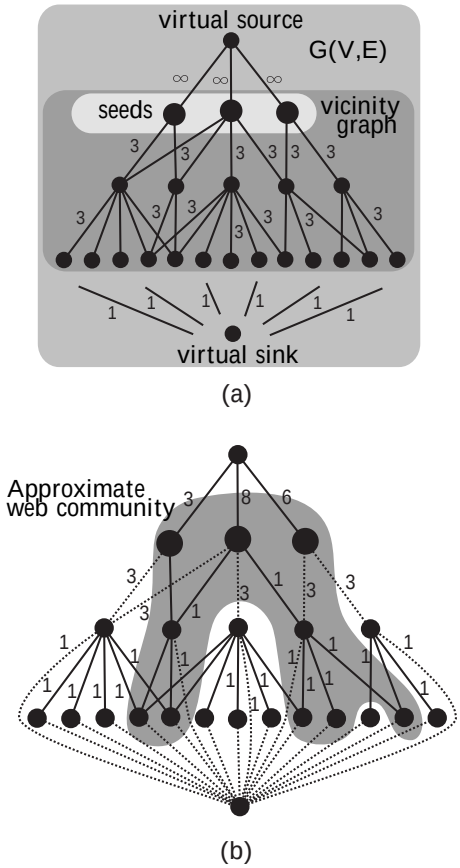

Fig. 2 An example of applying maximum flow for obtaining an *approximate* web community. For simplifying, suppose all the edges are going the same direction i.e., from the upper nodes to the lower nodes. The numbers in both (a) and (b) indicate the edge capacities and the final flow respectively. Edges to virtual sinks have been omitted in (a). The flow of the edges without any number is 0 in (b).

The theorem has supposed $s-t$ maximum flow algorithm is applied to the entire web graph, which is not entirely practical because of the size of the web graph. Therefore, a procedure to obtain the web community

that approximately satisfies the definition has been described in [8]. The procedure is as follows.

Suppose S is a set of seed nodes. $G = (V, E)$ is a subgraph of the web graph crawled within a certain depth from the nodes in S i.e., a certain in/out links away from the nodes in S. The subgraph $G = (V, E)$ is called a *vicinity graph*. The depth set to 2 in all the experiments in the paper [7], [8]. V and E are a set of nodes and edges respectively, and $S \subset V$. Suppose any edge $e \in E$ is dual directed with the edge capacity $c(e) = |S|$. Add a virtual source node s to V with the edges connecting to all nodes in S with the edge capacity $= \infty$, and add a virtual sink node t to V with the edges connected from all nodes in $V - \{S \cup s \cup t\}$ to t with the edge capacity $= 1$. Then perform $s - t$ maximum flow algorithm for G. All nodes accessible from s through unsaturated edges become new member pages of a web community. Add some nodes in the obtained web community to S, repeat the procedure until the desired-sized web community can be obtained.

Figure 2 shows a simple example of the procedure using a vicinity graph of depth 2. For simplification, all the edges are in the direction from the upper nodes to the lower nodes. Note that the number in (a) and (b) represent the edge capacities and the value of flow respectively. The nodes accessible from the seed nodes through unsaturated edges are extracted as member pages of a web community.

3. Understanding features and problems of previously proposed approach

In this section, we will discuss the features and the problems of the approach proposed in [7], [8].

3.1 Community size dependency on edge capacities

In the paper [7], [8], that the edge capacities are heuristically chosen was mentioned. This means that the edge capacities are first set to the number of the seed nodes as in the procedure and if the size of obtained web community (i.e., the number of member pages) becomes too small, say a few nodes around the seed nodes, it is suggested that the edge capacity should be made larger [7], [8]. Especially, in the case the number of the seed nodes is very small, the edges connecting to (or connected from) the seed nodes can be easily saturated. The size of the obtained web community becomes extremely small, say a few nodes around the seed nodes, so therefore the edge capacities have to be raised.

Since no clear statement about edge capacities and its effects were given in the papers in [7], [8], we performed an experiment to examine the effects of the edge capacities raised one by one on the size of the obtained web community. Figure 3 shows the relation between the value of edge capacities and the size of the obtained web community for the two seed nodes; www.orgel.com/, www.lares.dti.ne.jp/~jubal/home-j.html.

We have examined the effects of variations of the edge capacities on the size of web community using the same data set as the one used in Section 5. The two seed pages about pipe organ are used. Figure 3 shows the relation between edge capacities and the size of the obtained web community.

As can be seen in the graph chart, although the size of the obtained web community becomes larger with increment of the edge capacities, the behavior of the increases is discrete. In the case of this example, with raising the edge capacities from 9 to 10, 14 to 15, and 20 to 21, the size becomes drastically larger i.e., from 9 to 36, 36 to 50, and 50 to 630 respectively. Even if another seed set is given, the similar behavior was observed. In the worst case, only one quantum jump extracts all nodes in the vicinity graph. In other words, the number of the nodes in the obtained web community changes from just a few nodes to all nodes in the vicinity graph before and after the quantum jump (as can be seen in the topic No.11, No.12, and No.17 of the experiments in the Section 5). This fact implies that the increase and decrease of the edge capacities cannot always be a solution for obtaining web communities whose size is neither too small (i.e., a few nodes around the seed node) nor too large (i.e., most nodes in the vicinity graph).

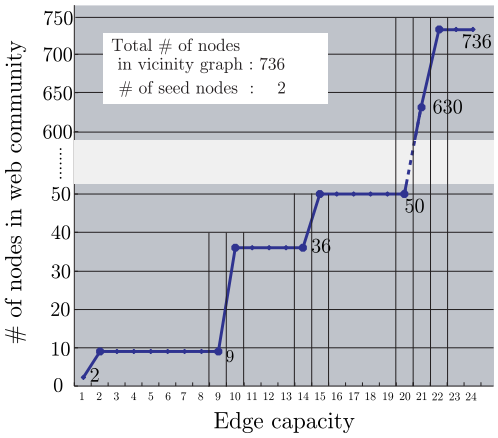

Fig. 3 Relation between edge capacities and size of a web community. The same data set as in Section 5 is used. The seed urls are; www.orgel.com/, www.lares.dti.ne.jp/~jubal/home-j.html, and both pages are about pipe organ.

3.2 Identification of possible cause of noise

Now, we examine the quality as a web community. In the context of web community, the irrelevant pages that are not related to the topic of the given seed pages are regarded as *noises*. The less the number of included

noises is, the higher the quality of the web community is. As we mentioned above, the edge capacities are assigned a same constant value for all edges in their approach. Accordingly, the graph patterns, which cause noise, can be roughly estimated as follows:

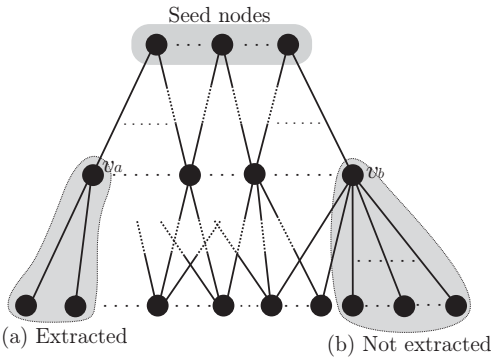

Fig. 4 A rough estimation of the graph patterns extracted and not extracted. The node v_a in (a) has less links to the nodes whose degrees are 1 than the edge capacities, conversely, the node v_b in (b) has much more links.

(a) The potential noises which cannot be excluded by the current community extraction approach: The graph pattern that is *always* extracted.: Let v_a be a node that is directly connecting to a seed node. Suppose the number of the links from v_a to the nodes whose degree are 1 is less than the value of the edge capacities (see the Figure 4). Since the possible total flow coming into v_a is bigger than the maximum total flow that can go out from v_a, v_a and all the nodes incident to v_a are included in the web community.

In the case of (a), the links from the web pages represented by v_a are accidentally added for some reasons and the web pages pointed by these links are the least likely related to the seed pages. The reason of this is that if the links are well selected and pointing to related pages, those pages are most likely pointed from other nodes in the vicinity graph. Thus, the pages in (a) become noise pages. Moreover, the larger the edge capacities are raised, the higher the risks of extracting the parts like (a) become.

(b) The potential noises which can be excluded by the current community extraction approach: Let v_b be a node that is directly connecting to a seed node. Suppose the number of the links from v_b to the nodes whose degree are 1 is much more than the value of the edge capacities (see the Figure 4). Since the possible total flow coming into v_b is much less than the maximum total flow that can go out from v_b, v_b and the nodes incident to only v_b are included in the web community.

In the case of (b), the web pages represented by v_b can be regarded as *hubs* which have many links pointing to various pages. The pages pointed from multiple pages in addition to v_b are the most likely extracted. Conversely, the pages pointed from only v_b are the least likely extracted. Imagine that a hub has a few hundreds of links. In that case, while most pages pointed from the hubs are not really related to the seed pages, some pages pointed from another pages as well as the hub are the most likely related to the seed pages. Thus, the pages in (b), which are expected to be noises, are excluded by their approach.

Whether the nodes except for such extremes are extracted or not is mostly depends on the assignment of edge capacities and the degrees of nodes. Therefore, if there are many hubs, which are having a moderate numbers of well selected links pointing to the pages on the topic of the seed pages, the web community in high quality can be obtained. If not, the quality becomes quite low.

4. Assignment of edge capacities using HITS scores

As stated in the previous section, that edge capacities assigned a same constant value causes some problems in the context of both quantity and quality. Raising the edge capacity cannot always be a solution to avoid extracting only a few nodes or conversely most nodes in the vicinity graph as member pages, and moreover if the edge capacity is raised larger and larger, the risks of introducing the parts like (a) in Figure 2 become high. As a solution of these problems, we consider a way of assigning edge capacities, which is not a certain common value but reasonably chosen different value for each edge.

An ideal assigning of edge capacities is; larger capacity should be assigned for important edges i.e., the edges that both of the end nodes are related to the topic of the seed nodes. Conversely, not-important edges are assigned smaller capacities. Distributing the flow according to the edge importance can naturally lead to extract the set of related nodes.

4.1 Related page algorithm: HITS

The edge importance can be measured by the scores obtained from related page algorithms. Two algorithms, PageRank [17] and HITS, have been most well-known and accepted for scoring web pages based on link. PageRank scores web pages by a measure of *prestige*. The prestige score of a page is roughly proportional to the sum of the prestige scores of pages linking to the page. One of the biggest differences between two algorithms is that HITS first chooses the subgraph depending on *query*, in contrast PageRank. In our case, query equals to given seed pages. Unlike PageRank algorithm, HITS can work effectively even in the small subgraph of the web graph. Since the method this paper describes identifies a web community in the subgraph within 2 links away from the given seed nodes, HITS algorithm is fit to use for our method. By the

hub and authority scores of the both end nodes of the edge, the importance of directed edges can be easily estimated.

Now, we review HITS scoring system. Let $V = \{v_1, v_2, ..., v_N\}$ be nodes in the graph. Denote $\mathbf{H} = (h_{v_1}, h_{v_2}, ..., h_{v_N})$ and $\mathbf{A} = (a_{v_1}, a_{v_2}, ..., a_{v_N})$ be a hub vector and an authority vector respectively. First, initialize all elements of the vectors $\mathbf{H}$ and $\mathbf{A}$ to 1, and iterate the following calculation until the vectors $\mathbf{H}$ and $\mathbf{A}$ have converged:

1. For all node $v \in V$, $a_{v_i} := \sum_{(v_j, v_i) \in E} h_{v_j}$.
2. For all node $v \in V$, $h_{v_i} := \sum_{(v_i, v_j) \in E} a_{v_j}$.
3. Normalize each vector $\mathbf{H}$ and $\mathbf{A}$, so that the sum of squares of each entry becomes 1.

That the iteration would converge soon was proved in [14]. In our empirical knowledge, about 30 iterations are enough to have the rankings of hub and authority scores stabilize.

4.2 Edge capacities based on HITS score

First of all, we define *adaptive hub score* h'_{v_i} and *adaptive authority score* a'_{v_i} as follows.

$$h'_{v_i} = r_h \; h_{v_i}, \qquad a'_{v_i} = r_a \; a_{v_i}. \tag{1}$$

Originally hub and authority scores are $0 \leq h_{v_i}, a_{v_i} \leq 1$ and not suitable for edge capacities; remember edge capacities have to be assigned a positive integer. Therefore, r_h and r_a are key factors in order to use original hub and authority scores well suited for edge capacities.

First, the maximum value of the adaptive hub and authority scores has to be determined. As can be seen in the previous section, there exists one *final quantum jump point* i.e., if the edge capacities exceed a certain value, all nodes in the graph turn out to be community members. For example, in the case of Figure3, the final quantum jump point is at edge capacities = 22. We use the value of the final quantum jump point for the maximum value.

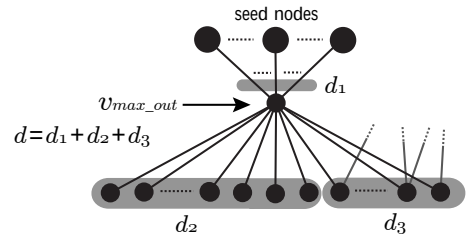

Fig. 5 Local condition around the node with the largest degree, v_{max_out}.

The value of final quantum jump point fq can be estimated as follows. Since the final quantum jump point means the edge capacities that are large enough to keep all edges unsaturated, in the case of the edge capacities $= fq - 1$, the total possible flow that can enter a certain node v_{max_out} is smaller than the total flow that has to go out from v_{max_out}. It is assumed that v_{max_out} is the nodes with the largest degree. Moreover since the vicinity graph is crawled within depth 2 from seed nodes, v_{max_out} is the most likely a node directly connected with the seed nodes. Let d and d_1 be the degree of v_{max_out} and number of edges directly connected with the seed nodes respectively, and also d_2 and d_3 be the numbers of edges connecting to the nodes with degree 1 and the others respectively (see Figure5). Note that $d = d_1 + d_2 + d_3$. The following inequality can be obtained.

$$(fq - 1)\; d_1 \leq d_2 + 1 \leq fq\; d_1$$

Note that $+1$ in the center represents that consider an edge to the virtual sink node. Consequently, fq can be obtained by the following equation.

$$fq = \lfloor \frac{d_2 + 1}{d_1} + 1 \rfloor$$

Empirically, there is often a great difference between top hub and top authority scores. In order to make both adaptive hub score and adaptive authority score ranged in a common scale, use r_h and r_a obtained by

$$r_h = \frac{max(\mathbf{A})}{max(\mathbf{H})} fq, \qquad r_a = fq.$$

Finally, let $c(v, u)$ denote the edge capacity for the edge from node v to u i.e., $v, u \in V$ and $(v, u) \in E$. We will assign edge capacity $c(v, u)$ by the following formula.

$$c(v, u) = \lfloor \frac{h'_v + a'_u}{2} \rfloor \tag{2}$$

The effects of using edge capacities obtained by the formula (2) will be discussed in the section 5.

4.3 Procedure of maximum flow based method

Now we summarize the procedure of applying the maximum flow algorithm by using HITS score based edge capacity.

1. Input $S = \{v_{s_1}, v_{s_2}, ..., v_{s_l}\}$ as a set of seed nodes.
2. Extract a subgraph within depth 2 around each $v_{s_i} \in S$.
3. Calculate hub and authority vectors $\mathbf{H}$ and $\mathbf{A}$.
4. Construct a vicinity graph $G(V, E)$ similarly as the previously proposed approach (reviewed in the Section 2.2).
5. Set edge capacity $c(u, v)$ to the original edge $(u, v) \in E$ by using the formula (2). If $(v, u) \notin E$, add an edge (v, u) to E with edge capacity $c(v, u) = c(u, v)$.
6. Perform $s - t$ maximum flow algorithm.

7. Obtain a set of the nodes still connected with seeds. Let $C = \{v_{c_1}, v_{c_2}, ..., v_{c_m}\}$, and score the nodes $v_{c_i} \in C$ (we will present the scoring procedure later).
8. Add a few highest ranked nodes to S and repeat the procedure until the nodes in C are stabilized.
9. Output the nodes in C in order of the scores.

Empirically, the nodes in C are usually stabilized very soon. The reason of this is that the vicinity graph is usually not (or just a little) expanded with added new seed nodes.

Our method scores the member nodes by the sum of the number of in and out links weighted by authority and hub scores respectively as follows. Let $v_{c_i}(In)$ and $v_{c_i}(Out)$ be a number of links from $v_{c_j} \in C$ to v_{c_i} and a number of links to v_{c_i} from $v_{c_k} \in C$ respectively. Then, let $Sc(v_{c_i})$ denote the score of v_{c_i}. $Sc(v_{c_i})$ is obtained by

$$Sc(v_{c_i}) = a_{v_{c_i}} v_{c_i}(In) + h_{v_{c_i}} v_{c_i}(Out).$$

The method in [7], [8] scored the member nodes the number of in/out links from/to the nodes in the community. Since a few or more highest ranked nodes have sometimes a same number of links, the score was not enough information for chosing new seed nodes. Our scoring method avoids this situation.†

5. Evaluation

We examined the effects of the assignment of edge capacities that we proposed in the previous section by using real web pages.

5.1 Experimental evaluation

Data set: An archive of Japanese web pages is used. The archive includes approximately 45 million pages in the 'jp' domain, or ones in other domains but written in Japanese characters, that we collected in February 2002. We built a connectivity database i.e., a *web graph database*. The web graph database enables us to find outgoing and incoming links of a given page. The web graph database holds approximately 84 million urls and 375 million hyperlinks.

Seed sets: 20 web pages related to each distinct topic are chosen as seed pages. The topics are determined so as to be quite specific. The Table1 shows those seed pages and the short descriptions about the pages i.e., the short terms, appeared in the page titles or the like, express well the corresponding topic.

Pruning: Pruning is done at the creation of vicinity graphs. First, we exclude the pages having more than 5000 in-links or out-links, such as Yahoo! and Google, from the subgraph (crawled from the web graph database). Those pages are generally "famous" in a sense, and so it is likely that we can find easily those pages without using any special strategies. Then prune the pages having URL that include the key words; %, ?, bbs, cgi-bin, diary, news, because those pages are usually blended with multiple topics and difficult to be categorized to a certain specific topic. Moreover, merge mirrors. The mirrors are identified as the pages having more than 10 out links of which more than 90% are common.

5.2 Experimental results

Let $C1$ and $C2$ be the web communities obtained by using the assignment of edge capacities that we proposed in the previous section and Flake *et al.* proposed respectively. $C1$ is obtained along the procedure stated in the previous section. The brief procedure for obtaining $C2$ is as follows; 1) Input a seed, 2) Set all the edge capacities to 3, and 3) Raise the edge capacities until the size of $C2$ becomes bigger than 10. The is because we will compare the relevancy of the highest ranked 10 nodes in $C1$ and $C2$.

Various kinds of $s-t$ maximum flow algorithms have been proposed [1], [10]. Our implementation uses one of the traditional algorithms; the shortest augmentation path algorithm [6].

1. Find a shortest path p from s to t which consists of unsaturated edges.
2. Augment the flow on p until an edge of p is saturated.
3. Repeat until p cannot be found.

This algorithm uses breadth-first search to find the shortest path p, and its computation time takes $O(VE^2)$.

5.2.1 Extraction of web communities

Table 1 shows the size of $C1$ and $C2$ for each seed node. The mark "*" in the row $C2$ of No.11, 12 and 17 mean that these cases are failed to obtain the reasonable sized web community. In other words, no matter what value is assigned for edge capacities, a few nodes less than 10 or all nodes in the vicinity graph are extracted as member nodes. So the numbers in the row indicate the maximum size less than 10 nodes.

†As far as observing our experimental results, the scoring method is not as effective as changing the members of the highest ranked 10 nodes, though the ranking of those nodes can be different.

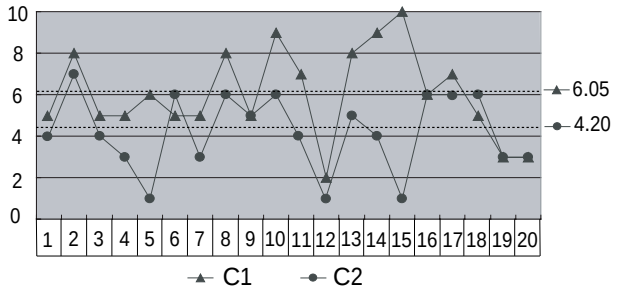

Fig. 6 # of relevant pages as member pages among the highest ranked 10 pages. The numbers in below correspond to the topic numbers (see Table 1). The average numbers for $C1$ and $C2$ are 6.25 and 4.30 respectively.

5.2.2 Quality evaluation of extracted communities

We observed whether the highest ranked 10 pages obtained by both approaches are relevant as member pages of corresponding web communities or not. The graph chart in Figure 6 shows the results. The average numbers of relevant pages in $C1$ and $C2$ are 6.25 (ranges from 3 to 10) and 4.30 (ranges from 1 to 7) respectively. As can be seen in the graph chart, in 14 among 20 cases, $C1$ is better than $C2$, 4 cases are in the same level, and in 4 cases, $C1$ is worse than $C2$. Most relevant pages included in $C2$ are also included in $C1$.

Table 2 shows that the URLs of the highest 10 ranked member pages of $C1$ and $C2$ about the Topic No.15. "+" and "-" in the Tables mean the page is relevant as a member page and not relevant respectively. The seed page of Topic No.15 is about the blind marathon and ultra-marathon. While all pages in $C1$ are related to the topic "Running", all member pages except for seed page in $C2$ are on the topic of "Sports" such as water polo, soccer, riding horse, and so on. The highest ranked page of $C2$ is a hub having 23 links to the pages about various kinds of sports including the seed page. When the edge capacities are assigned to 24, this hub is extracted as well as such noise pages. Actually 8 pages that ranked in the second to tenth are pointed from this hub. By using the assigning method, the edge capacity of the edge from this hub to the seed page set to 16 (the adaptive authority score of the seed page is 27.2 and the adaptive hub score of the hub is 5.7). Therefore, this hub including the pages pointed from this hub was not extracted.

In the case of Topic No.1, the forth ranked page in $C2$ has two links to the seed page and the sixth ranked page in $C2$. The forth ranked page is a top page of a private site that the author is writing about his/her some hobbies. This page has only two links to the pages extracted in the vicinity graph. One is the seed page and the other is the sixth ranked page (a home page of an N.G.O group "Foster Plan"). Since $C2$ was extracted by using the edge capacities = 6, those pages are included in $C2$. Those noise pages were not included in $C1$, because our assignment set the edge capacity between the seed node and the forth ranked page to 2.

Table 4 The highest ranked 10 authorities for No.15.

1	village.infoweb.ne.jp/~suzukik	+
2	www2s.biglobe.ne.jp/~BBA	-
3	www.normanet.ne.jp/	-
4	www.rehab.go.jp/	-
5	kyoyohin.org/	-
6	www.twcu.ac.jp/~k-oda/VIRN	-
7	www.shigapref-sb.ed.jp/	-
8	www2d.biglobe.ne.jp/~tenyaku	-
9	rel.chubu-gu.ac.jp/soumokuji	-
10	www.tcp-ip.or.jp/~chubu	-

5.3 Discussions

We examine the differences between the maximum flow based method and the set of highest ranked authorities.

Table 4 shows the highest ranked 10 authorities for the Topic No.15. These pages are; Rehabilitation center for the disabled, Support association for the blind's reading, Information network of persons with disabled, a school for the blind, and so on. The topic of the authorities were drifted to the disabled including the blind. Even though our assignment of edge capacities is based on HITS scores, influence of the topic drift by HITS is rarely found. This is because the high ranked hubs pointing to these authorities are on the shortest paths from the seed to these authorities and those hubs have many links enough to saturate easily the edges between the seed node and these hubs.

Table 3 show that the URLs of the highest 10 ranked member pages of $C1$ and the pages of the highest 10 authority scores about Topic No.11. The seed page of Topic No.11 is "The Kurashiki Information Box" that is a portal site for finding information including the hotels, the sight seeing spots and so on. The numbers of relevant pages of $C1$ and the highest ranked authority scores are same i.e., 6 pages each, in which only one page is overlapping. In the case of No.12, while most of member pages in $C1$ are irrelevant, all the highest ranked authorities except for one page are relevant. While the relevant member pages of $C1$ are the home pages of hotels located in Kurashiki, the authorities are the official pages of Kurashiki city and Kurashiki park, and portal sites for sightseeing in Kurashiki and so on. There exist a several hubs that have a great amount of links to the pages about Okayama prefecture where Kurashiki is located in. Since the links in these hubs are well categorized into the cities, sight seeing places, those authority pages are ranked high. The home pages of hotels, which are member pages of $C1$, provide the information about Kurashiki city and the sight seeing places around their hotels as well as information about

Table 1 Seed nodes used for experiments. Topics are based on the page titles, frequently referred terms etc. $|V|$, $|C1|$ and $|C2|$ indicate the number of nodes in the vicinity graph, the number of nodes of the web community obtained by using the HITS score based edge capacities and the edge capacities of a constant value respectively.

No.	Seed URLs	Topics	$\|V\|$	$\|C1\|$	$\|C2\|$
1	member.nifty.ne.jp/andes/	Andes civilization	822	19	18
2	www.alive-net.net/	Animal protection	898	15	10
3	www.t3.rim.or.jp/~star	Stars and astronomy	3727	21	12
4	www.ask.ne.jp/~bungy	Bungy jump	675	14	12
5	coffee.ajca.or.jp/	Coffee	5663	47	20
6	well-mannered.org/	Dogs and pet	2587	58	13
7	www.zipangu.com/Gagaku	Gagaku	2978	20	13
8	www.gender.go.jp/	Gender	1043	34	17
9	jazzfusion.com/	Jaz fusion	2746	31	11
10	www.joa.or.jp/	Orthopaedics	1554	48	13
11	www.kurashiki.or.jp/	Kurashiki city	2196	13	6*
12	www.lurefishing.net/	Lure fishing	1685	14	2*
13	www.railfan.ne.jp/	Reil fan	2670	42	10
14	www1.ocn.ne.jp/~tatsujin/ropework	Rope work	1275	34	10
15	village.infoweb.ne.jp/~suzukik	Running	1021	14	28
16	www.salsa.org/	Salsa dance	1395	17	17
17	www.hi-ho.ne.jp/nua-nua/nua.html	Ukulele	1098	10	8*
18	www.eu-ki.com/	Organic foods	380	11	17
19	gnl.cplaza.ne.jp/walking	Walking	883	25	17
20	www.e-wedding.ne.jp/	Wedding	771	15	11

Table 2 Topic No.15. Running; village.infoweb.ne.jp/~suzukik

	$C1$		$C2$	
1	village.infoweb.ne.jp/~suzukik	+	www.sportsnet.ne.jp/look/bk01.html	-
2	www.souji.co.jp/mall/yume012.html	+	village.infoweb.ne.jp/~suzukik	+
3	mm.amie.or.jp/scmt/1224/frame/kunio.htm	+	www.fsinet.or.jp/~kunren/flyball.html	-
4	www.runnet.co.jp/online/relay/2/relay.html	+	www.padi.co.jp/	-
5	www.mars.sannet.ne.jp/timers/link.html	+	www.waterpolo-japan.com/	-
6	www.hiroba.gr.jp/yours/yourssk3.htm	+	www.asahi-net.or.jp/~BY3S-FET/autumn.htm	-
7	www3.infoweb.or.jp/rikuren/	+	www.jouba.jrao.ne.jp/	-
8	www.kikimimi.ne.jp/www/yyyrun/	+	www.nippon12.com/nippon.html	-
9	www2s.biglobe.ne.jp/~tetsu-y	+	www.gochomuseum.net/kouyou	-
10	www.bbm-japan.com/jp/rikujyo/rikujyo.html	+	nakata.net/jp	-

Table 3 Topic No.11. Kurashiki; www.kurashiki.or.jp/

	$C1$		Authority ranking	
1	www.kurashiki.or.jp/	+	www.kurashiki.or.jp/	+
2	ww1.tiki.ne.jp/~tombow21/link/links.htm	+	www.tivoli.co.jp/	+
3	www.mmd.co.jp/tsurugata/annai.html	+	www.city.kurashiki.okayama.jp/	+
4	www.yadonet.ne.jp/junior/b-west/b-tyugoku1.htm	-	www.oka.urban.ne.jp/home/bao/	-
5	www.net626.co.jp/kurasiki/	+	www.kurashiki.co.jp/	+
6	home.kiui.ac.jp/~pt/gakkai/vol26.html	-	www.tiki.ne.jp/~shimoden	+
7	www.kurashiki.jp/	+	www.sqr.or.jp/usr/muni/	-
8	www.kurashiki-kokusai-hotel.co.jp/	+	www.spa-yunogo.or.jp/	-
9	www.ecolife.gr.jp/uomoto/	-	www.fmkurashiki.com/	+
10	www.jpan.org/presenter/station/station-s/	-	www.johshuya.co.jp/	-

the hotel themselves. Those pages turned out to be high quality small sized hubs. HITS algorithm finds hubs having many links pointing to other pages and authorities having many links pointed from other pages. Highest ranked hubs and authorities have usually many links. In other words, the small sized hubs (not having many links) are not ranked higher in the scoring of HITS. As the result, the page set that cannot be easily found by HITS was obtained as member pages of $C1$.

6. Conclusion

In this paper, we proposed a method for finding web community using the maximum flow algorithm. In particular, we propose an assignment of edge capacities for the best use of the maximum flow algorithm. First, we reviewed the previously proposed method by [7], [8] and examined the approach. Especially we examined the effects of the edge capacities assigned constant value. Then, we proposed a well-assignment of the edge capacities using hub and authority scores. In order to examine the effects, we performed experiments using a Japanese archive crawled in February 2002. Our experimental result showed that the highest ranked ten nodes in each web community, on the average, 1.44 times more relevant pages are extracted by using the well-assigned edge capacities. We also demonstrated the examples that the noise pages are excluded.

We recognized that our proposed method is just a first trial for assigning variable edge capacities. We would like to examine more sophisticate assignment of edge capacity as the future work.

Acknowledgements

This research was partially supported by e-Society Leading Project and Grant-in-Aid for Scientific Research on Priority Areas C (No.13224014) of the Ministry of Education, Culture, Sports, Science and Technology.

References

[1] R. K. Ahuja, T. L. Magnanti, and J. B. Orlin. Network flows: Theory, algorithms, and applications. *Algorithms, and Applications*, 1993.

[2] A. Borodin, G. O. Roberts, J. S. Rosenthal, and P. Tsaparas. Finding authorities and hubs from link structures on the world wide web. In *Proceedings of the 10th International World Wide Web Conference*, pages 415–429, 2001.

[3] A. Broder, R. Kumar, F. Maghoul, P. Raghavan, S. Rajagopalan, R. Stata, A. Tomkins, and J. Wiener. Graph structure in the web: Experiments and models. In *Proceedings of the 7th International World Wide Web Conference*, pages 309–320, 2000.

[4] S. Chakrabarti, B. Dom, D. Gibson, J. Kleinberg, P. Raghavan, and S. Rajagopalan. Automatic resource list compilation by analyzing hyperlink structure and associated text. In *Proceedings of the 7th International World Wide Web Conference*, 1998.

[5] J. Dean and M. R. Henzinger. Finding related pages in the World Wide Web. In *Proceedings of the 8th International World Wide Web Conference*, pages 1467–1479, 1999.

[6] J. Edmonds and R. Karp. Theoretical improvements in algorithmic efficiency for network flow problems. *Journal of the ACM*, 19(2):248–264, 1972.

[7] G. Flake, S. Lawrence, and C. L. Giles. Efficient identification of web communities. In *6th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 150–160, 2000.

[8] G. W. Flake, S. Lawrence, C. L. Giles, and F. Coetzee. Self-organization of the web and identification of communities. *IEEE Computer*, 35(3):66–71, 2002.

[9] D. Gibson, J. M. Kleinberg, and P. Raghavan. Inferring web communities from link topology. In *UK Conference on Hypertext*, pages 225–234, 1998.

[10] A. V. Goldberg and R. E. Tarjan. A new approach to the maximal flow problem. *Journal of the ACM*, 35(4):921–940, 1988.

[11] N. Imafuji and M. Kitsuregawa. Effects of maximum flow algorithm on identifying web community. In *4th International Workshop on Web Information and Data Management*, pages 43–48, 2002.

[12] N. Imafuji and M. Kitsuregawa. Finding a web community by maximum flow algorithm with hits score based capacity. In *8th International Conference on Database Systems for Advanced Applications*, pages 101–106, 2003.

[13] L. Jr. and D.R.Fulkerson. Maximal flow through a network. *Canadian Journal of Mathematics*, 8:399–404, 1956.

[14] J. M. Kleinberg. Authoritative sources in a hyperlinked environment. *Journal of the ACM*, 46(5):604–632, 1999.

[15] R. Kumar, P. Raghavan, S. Rajagopalan, and A. Tomkins. Trawling the Web for emerging cyber-communities. In *Proceedings of the 8th International World Wide Web Conference*, pages 1481–1493, 1999.

[16] S. R. Kumar, P. Raghavan, S. Rajagopalan, and A. Tomkins. Extracting large-scale knowledge bases from the web. In *The VLDB Journal*, pages 639–650, 1999.

[17] L. Page, S. Brin, R. Motwani, and T. Winograd. The PageRank Citation Ranking: Bringing Order to the Web. Stanford Digital Libraries Working Paper, 1998.

[18] M. Toyoda and M. Kitsuregawa. Creating a web community chart for navigating related communities. In *12th ACM Hypertext*, pages 103–112, 2001.