COMBATING SOCIAL BOOKMARKING POLLUTION

Hiroyuki Hisamatsu[1] Takahiro Hatanaka[2]

[1]Department of Computer Science
Osaka Electro-Communication University
hisamatu@isc.osakac.ac.jp
[2]Graduate School of Computer Science and Arts
Osaka Electro-Communication University
takahiro.hatanaka@olnr.org

ABSTRACT

Social Bookmarking (SBM) is one of the most widely used Web services. An SBM website displays and shares each user's bookmarks. The SBM service aggregates the number of users who bookmark a given Web page and provides useful information as a result of these aggregations. However, an increase in the popularity of the SBM service and in the number of the users of the SBM service results in an increase in the amount of SBM SPAM. In addition, the SBM service generates irrelevant information to many users because of the aggregation of a large number of bookmarks; we call this problem "SBM pollution."

In this paper, we propose a method for countering the problem of SBM pollution based on the degree of bookmark similarity. The proposed method creates blacklists that contain lists of users having a high degree of bookmark similarity. Based on the created blacklists, the number of bookmarks of the Web pages influenced by SBM pollution is reduced. From the results of the performance evaluation, we show that our method reduces the number of bookmarks of most Web pages influenced by the SBM pollution to a great extent.

KEYWORDS

Social Bookmarking (SBM), Consumer-Generated Media (CGM), Collective Knowledge, Social Bookmark SPAM

1. INTRODUCTION

In recent years, many new Web services [1-4] based on Web 2.0 [5] have been launched. Consumer-generated media are one of the features of Web services that are based on Web 2.0. In previous Web services, users could only read information on the Web and only a few users were able to upload information on the Web. However, even if users do not have valuable information to share, services such as blogs and social networks allow them to voice their opinion on the Web, rendering these services popular. Consequently, users actively exchange information over the Web. The focus is now on collective knowledge [6], wherein voluminous information that users publish get aggregated on the Web, and then, new useful information is generated from the aggregated information. Social bookmarking (SBM) [7] is one of the Web services that implements the use of collective knowledge.

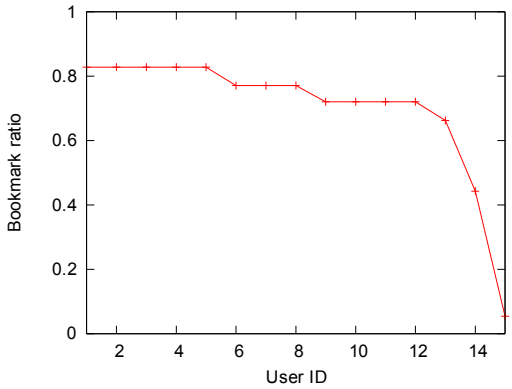
SBM is a Web service for displaying and sharing each user's bookmark and is offered by various corporations [8-11]. Furthermore, many SBM websites aggregate the number of users that

DOI : 10.5121/ijwsc.2012.3201

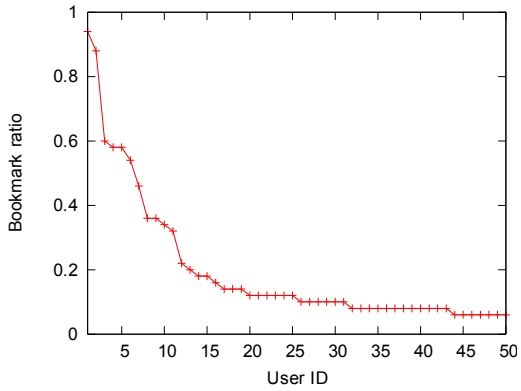
bookmark a Web page (hereafter, we will refer to the aggregated number of users as 'number of bookmarks') and displays the aggregated results by rank. The SBM service aims at providing useful information by using the obtained aggregated information. SBM can be considered a new method for retrieving useful information that is archived on the Internet. In recent times, SBM has been gaining popularity, and hence, the number of users using this service has been increasing. This has led to an increase in SBM SPAM, which abuses the SBM service for commercial purposes such as advertisements. Because of SBM SPAM, Web pages that are not desirable for many users are displayed on the ranking of SBM websites because of information aggregation.

SBM services face another problem besides SBM SPAM. For instance, in a lecture at the university, students performed an exercise; they created Web pages and bookmarked them to each other. Consequently, the created Web pages ranked high in the SBM website. Moreover, there were many animation videos in the "Attention videos" [12], a service for ranking Web pages of videos with a high number of bookmarks. Therefore, the ranking in Attention videos are not useful for users who are not interested in animation. Although it is not used for commercial purposes, the current SBM service has a drawback, in that it provides a high ranking to Web pages that may not be useful for many users. In this paper, we will hereafter name this drawback as SBM pollution, which occurs when Web pages that are not useful for many users appear in the ranking of SBM websites as a result of the aggregation of bookmarks. These Web pages can be considered as useful Web pages by implementing the concept of collective knowledge. Note that SBM SPAM is an example of social bookmarking pollution.

Much research has been conducted on SBM [13-18]. For instance, in [13], the authors proposed an algorithm for SBM analysis, and they improved the personalized recommendation using their algorithm. In [14], a system that recommends the user and/or the document was proposed using SBM data. Moreover, in [15-17], methods were proposed for improving the referencing accuracy of a search engine on the basis of information obtained using the SBM service. In [18], the authors proposed a ranking algorithm based on tags coincidence between users. However, because SBM pollution is a relatively new problem, sufficient research has not yet been conducted for addressing it. The present method for countering SBM pollution is one that involves denying access to users who spam SBM websites by using their IP address. SBM pollution, however, which is a greater issue than SBM SPAM, has not been examined. Therefore, in this paper, we propose a method for countering SBM pollution. First, we analyze the



(a) Web pages bookmarked in the exercise



(b) Web pages with “idolmaster” tag

Figure 1. User-bookmark ratio

characteristics of SBM pollution and show that the high degree of similarity in users’ bookmark choices results in SBM pollution. Then, we propose a method for countering SBM pollution. Our method reduces the number of bookmarks of a Web page based on the degree of bookmark similarity between users. The results of the performance evaluation of our method show that the proposed method does not reduce the number of bookmarks of Web pages that are not influenced by SBM pollution. We also show that our method reduces the number of bookmarks of most Web pages that are influenced by SBM pollution to a great extent.

The rest of this paper is as follows. First, in Section 2, we analyze the characteristics of SBM pollution. Next, in Section 3, we propose a method for countering SBM pollution. In Section 4, we examine the parameters of the proposed method. In Section 5, we evaluate our method and show its effectiveness in countering SBM pollution. Finally, in Section 6, we conclude this paper and discuss the possibility of a few future studies.

2. ANALYSIS OF SBM POLLUTION CHARACTERISTICS

In this section, we analyze the characteristics of SBM pollution. In particular, we studied the following two examples of SBM pollution in Hatena Bookmark [11], which is the most popular SBM service in Japan: (a) In a lecture in the university, an exercise was conducted wherein students created individual Web pages and bookmarked them. Consequently, only the Web pages that the students created were displayed temporarily in “the Attention Web pages” [19], which is a service for ranking Web pages with a high bookmark number. (b) Only videos with the tag “idolmaster” were displayed in "the Attention videos".

We investigated the bookmarks of the users who bookmarked Web pages or videos displayed in the Attention Web pages or the Attention videos. Fig. 1 shows the user bookmark ratios in decreasing order of their value. That is, the number of bookmarks of these Web pages was higher than the actual number of bookmarks because of SBM pollution. Fig. 1(a) shows that the ratio is over 0.7 for almost all the students; that is, almost all the students had bookmarked the Web pages created by other students in the same lecture. Moreover, Fig. 1(b) shows that several users had bookmarked all the videos with the “idolmaster” tag. Therefore, it turns out that several of these users had a huge impact on the Attention videos service.

These results show that the users who bring about SBM pollution bookmark a series of Web pages. If such users' bookmarks are compared with the bookmarks of users who do not bookmark series of Web pages, we find that the degree of similarity among the former users' bookmarks is high compared with that of the latter users'. When the degree of similarity of users' bookmarks is high, it is inferred that SBM pollution occurs. We can therefore counter SBM pollution by reducing the number of bookmarks of Web pages that are bookmarked by users who cause SBM pollution.

3. COUNTERING SBM POLLUTION

In this section, we propose a method for countering SBM pollution. The proposed method involves the creation of blacklists, which are lists of users having a high degree of similarity in their bookmarks. Next, based on the created blacklists, the number of bookmarks is reduced for Web pages that are influenced by SBM pollution.

3.1 Creating Blacklists

We first calculate the degree of similarity in users' bookmarks, and then create blacklists of the users who contribute to SBM pollution. First, we acquire the bookmarks during period *T*. Then, we determine the similarity between the bookmarks for different users. Let $\mathcal{U}$ be a set of users of the SBM service. The degree of similarity of the bookmark s_{uv} of users u and v is given by

$$s_{uv} = \min\left(\frac{c_{u\to v}}{m_u}, \frac{c_{v\to u}}{m_v}\right), \quad (1)$$

where m_u is the number of Web pages bookmarked by user u during period *T*, and $c_{u\to v}$ is the number of Web pages bookmarked by user u that were already bookmarked by user v.

The degree of similarity in the bookmarks is calculated for every user, and users with a high degree of similarity are registered in a blacklist. Here, we focus on a user $u \in \mathcal{U}$.

Algorithm 1 Creation of blacklists

```
for each u ∈ U do
  if u has already been registered on the blacklist then
    next
  end if
  for each v ∈ U \ {u} do
    if bookmark similarity between u and v > γ then
      if v has already been registered on the blacklist l then
        if ∃x ∈ l, bookmark similarity between u and x > γ then
          register u on the blacklist l
        end if
      else
        create a new blacklist & register u and v on it
      end if
    end if
  end for
end for
```

If user u is already registered in the blacklist, we terminate the processing of user u. If user u is not registered in any blacklists, we calculate the degree of bookmark similarity with all other users. If we discover that the bookmarks of user v have a degree of similarity in those of user u that is greater than the threshold γ, there are then two possibilities, (1) if v is already registered in the

blacklist *l*, then we calculate the degree of similarity in the bookmarks of all users registered in blacklist *l*. If that degree of similarity is greater than threshold γ, we register user u in the blacklist *l*, and complete the processing of user u. (2) Conversely, if user v has not been registered in the blacklist, we create a new blacklist and register both users u and v in it. We then complete the processing of user u. Algorithm 1 shows the pseudocode used for creating blacklists.

3.2 Reduction of the Number of Bookmarks based on Blacklists

Next, the number of bookmarks of the Web page is reduced by using the created blacklists. Not all the bookmarks of the users belonging to blacklists contribute to SBM pollution. It is therefore not appropriate to remove all these users bookmarks. The more a Web page has the bookmarks of users registered on the blacklists, the more strongly the number of bookmarks of such a Web page reflects those users' intentions, and hence, the higher the probability of SBM pollution occurring. We therefore reduce the number of bookmarks of a Web page based on the ratio of the users belonging to the blacklist.

Let B_{orig} be the number of bookmarks of the Web page before countering SBM pollution; $\mathcal{L}$, the set of a blacklist; n_l, the number of users registered in the blacklist *l*; and m_l, the number of users in the blacklist *l* that bookmarked the Web page. The number of bookmarks after reduction B_{new} is given by

$$B_{new} = B_{orig} - \sum_{l \in \mathcal{L}} m_l \times \frac{m_l}{n_l} \quad (2)$$

We summarize the notations used in this section in Tab. 1.

Table 1. Definition of Notations

T	interval for acquiring bookmarks
γ	threshold for the degree of bookmark similarity
$\mathcal{L}$	set of blacklists
l	blacklist
$\mathcal{U}$	set of users
u, v	user
$c_{u \to v}$	number of Web pages bookmarked by user u that were already bookmarked by user v
m_u	number of Web pages bookmarked by user u
s_{uv}	degree of bookmark similarity of users u and v
B_{orig}	number of bookmarks before reduction
B_{new}	number of bookmarks after reduction

4. PARAMETER CONFIGURATION

In the proposed method, we need to configure two parameters, the interval T for acquiring users' bookmarks and the threshold γ for the degree of bookmark similarity. In this section, we configure these parameters.

4.1 Intervals for acquiring bookmarks

When the degree of bookmark similarity is calculated from only a few bookmarks, it may not be useful. We therefore want to acquire as many bookmarks as possible from each user. However,

acquiring a large number of bookmarks may increase the processing time for calculating the degree of similarity. Therefore, it is necessary to acquire an appropriate number of bookmarks. In order to determine appropriate intervals for acquiring bookmarks, we investigated the number of Web pages that one user bookmarks in a single day. We randomly selected 2,000 users of the Hatena Bookmark service, and acquired their bookmarks over the past 10, 30, and 90 days. Fig. 2 shows the distribution of the number of average Web pages that one user bookmarks in a single day in decreasing order of their value. From this figure, we find that the average number of Web pages that a user bookmarks in a single day over a period of 10, 30 or 90 days is uniform. Tab. 2 shows the percentile of the number of Web pages that one user bookmarks; as the percentile increases, the average number of bookmarks decreases.

Users who bring about SBM pollution bookmark Web pages actively. We found that such users account for the top 25% of all users, in terms of the number of Web pages that one user bookmarks in a single day. Moreover, we think that at least 40 Web pages are required for acquiring the appropriate number of bookmarks for calculating the degree of bookmark similarity. Therefore, the interval T for acquiring users' bookmarks is set at 30 days (a user who is in the percentile range of 25% or below bookmarks 39 Web pages in 30 days).

Table 2. Percentile of Number of Web Pages that One User Bookmarks

q percentile	10 days	30 days	90 days
1%	11.40	12.80	11.93
10%	2.80	3.16	2.93
25%	1.20	1.30	1.25
50%	0.40	0.53	0.48
75%	0.20	0.16	0.16
90%	0.10	0.06	0.06
99%	0.00	0.00	0.01

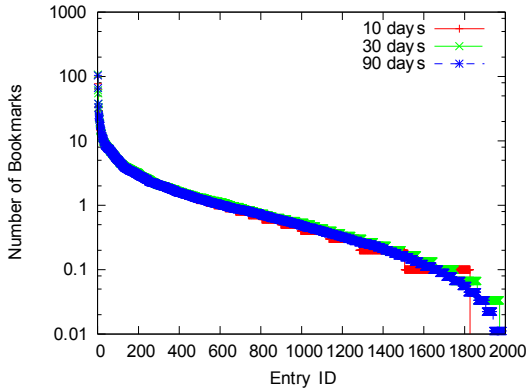


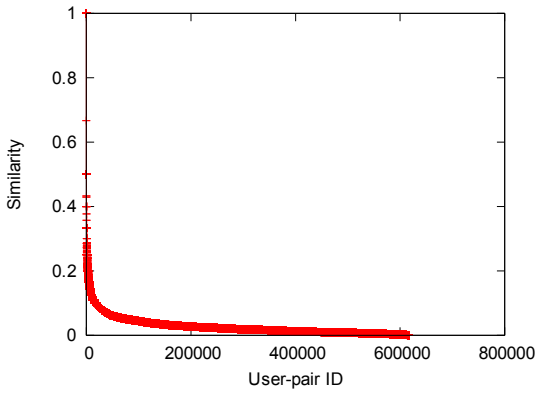
Figure 2. Distribution of number of Web pages that a single user bookmarks

4.2 Threshold for degree of bookmark similarity

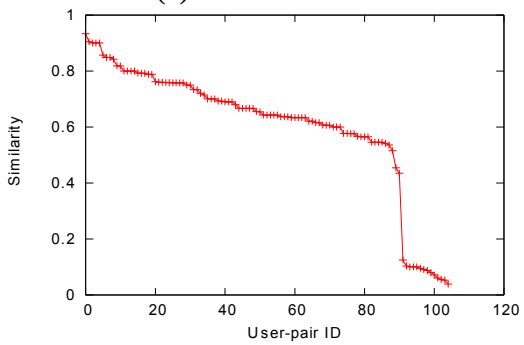
When the threshold γ for the degree of bookmark similarity is low, even a general user who does not contribute to SBM pollution may be registered in a blacklist. As a result, the number of bookmarks of Web pages that are influenced and not influenced by SBM pollution is reduced. Consequently, the number of bookmarks of all the Web pages in the SBM service is reduced. Therefore, the threshold γ should be configured appropriately.

In order to configure the threshold γ, we investigated the degree of bookmark similarity between users of the Hatena Bookmark service. Fig. 3 shows the degree of bookmark similarity of (a) 2,000 random users and (b) the users who contribute to SBM pollution. The degree of bookmark similarity was calculated by acquiring the bookmarks of each user over a span of 30 days. We would add that the users in (b) were the 19 students who participated in the university exercise, as explained in Section 2.

From Fig. 3(a), it is determined that the degree of bookmark similarity between general users is very low. On the other hand, from Fig. 3(b), it is determined that the degree of similarity between users who contribute to SBM pollution is very high. Therefore, it is presumed that by selecting a sufficiently high threshold value, the accidental registration of a general user in the blacklist can be prevented. The threshold γ is set to 0.6 in the proposed method.



(a) General users



(b) Users that contribute to SBM pollution

Figure 3. Degree of bookmark similarity

5. Performance Evaluation

In this section, we show the effectiveness of the proposed method by applying it in to the Hatena Bookmark.

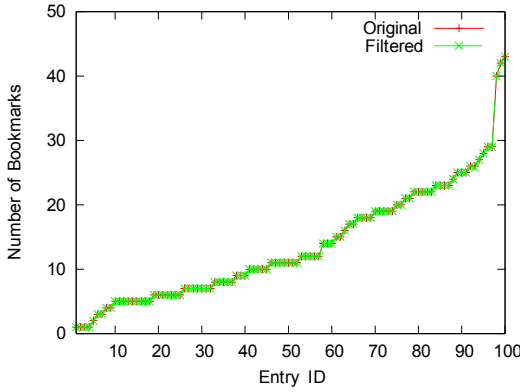
5.1 Evaluation Environment

First, we acquired 10,000 Web pages that appear in “New arrival Web pages” [20] in the Hatena Bookmark website for creating blacklists. We randomly selected 2,000 users, as well as the users

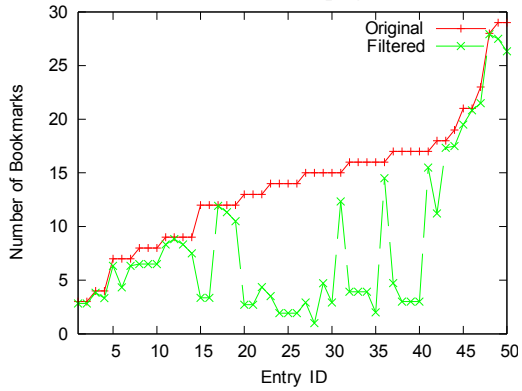
who contributed to SBM pollution, as explained in Section 2. The blacklist was created based on these users' degree of bookmark similarity. Hundred Web pages were randomly selected from the Attention Web pages from the Hatena Bookmark website, as well as 50 Web pages influenced by SBM pollution, as explained in Section 2; these Web pages were selected to test the proposed method. Based on the obtained blacklist, Eq. (2) was applied to the number of bookmarks of these 150 Web pages.

5.2 Evaluation Results

Fig. 4 shows the number of bookmarks on the Web pages before applying the proposed method, in increasing order of their value. In Fig. 4, we also plot the number of bookmarks on the Web pages after the proposed method was applied, in the same order as the number of bookmarks before the method was applied, in order to compare the number of bookmarks before and after the application of our method. From Fig. 4(a), it turns out that the number of bookmarks before applying the proposed method is almost the same even after the application of the proposed method to general Web pages. On the other hand, from Fig. 4(b) we find that the number of bookmarks of most Web pages that are influenced by SBM pollution is reduced greatly. However, there are Web pages that show hardly any reduction in the number of bookmarks. This is because these Web pages are bookmarked by many users who are not registered in any blacklists.



(a) General Web pages



(b) Web pages that are influenced by SBM pollution

Figure 4. Number of Bookmarks

The proposed method does not set the number of bookmarks that is not registered in a blacklist as the target number of reduction. Therefore, most of those Web pages do not show a reduction in the number of bookmarks.

From these observations, we found that the proposed method reduces the number of bookmarks of most Web pages influenced by SBM pollution to a great extent. However, it turned out that there are Web pages that hardly have any reduction in the number of bookmarks among the Web pages influenced by SBM pollution. Furthermore, we confirmed that there is no reduction in the number of bookmarks of the Web pages that are not influenced by SBM pollution.

6. Conclusion And Future Work

In this paper, we proposed and evaluated a method for countering SBM pollution. First, the characteristics of SBM pollution were examined. Our examination showed that users who bookmarked Web pages that are influenced by SBM pollution had an extremely high degree of bookmark similarity. Moreover, we proposed the method for countering SBM pollution based on the result of the examination. The proposed method creates blacklists of users who have an extremely high degree of bookmark similarity. By implementing the created blacklists, the number of bookmarks of the Web pages influenced by SBM pollution was reduced.

From the results of the evaluation on the Hatena Bookmark, we showed that the proposed method reduced the number of bookmarks for most of the Web pages influenced by SBM pollution. Moreover, we also showed that the proposed method does not reduce the number of bookmarks of those Web pages that are not influenced by SBM pollution.

As future work, it is important to develop a method for preventing SBM pollution of Web pages whose number of bookmarks are hardly reduced by the current method. Furthermore, it would be interesting to investigate a method for increasing the satisfaction of SBM users based on the degree of bookmark similarity.

Acknowledgment

This research was supported in part by Dayz Inc.

References

[1] “YouTube.” available at http://www.youtube.com/.
[2] “Gmail.” available at http://www.gmail.com/.
[3] “Flickr.” available at http://www.flickr.com/.
[4] “Facebook.” available at http://www.facebook.com/.
[5] T. O’Reilly, “What is Web 2.0: Design patterns and business models for the next generation of software.” available at http://www.oreillynet.com/pub/a/oreilly/tim/news/2005/09/30/what-is-web-20.html.
[6] J. Surowiecki, The Wisdom of Crowds: Why the Many Are Smarter Than the Few and How Collective Wisdom Shapes Business, Economies, Societies and Nations. Anchor Books, May 2004.
[7] R. Keller, S. Wolfe, J. Chen, J. Rabinowitz, and N. Mathe, “A bookmarking service for organizing and sharing URLs,” in Proceedings of the 6th International Conference on World Wide Web (WWW 1997), pp. 1103–1114, Apr. 1997.
[8] “delicious.” available at http://delicious.com.
[9] “digg.” available at http://digg.com.
[10] “reddit.” available at http://www.reddit.com.
[11] “Hatena bookmark.” available at http://b.hatena.ne.jp/.
[12] “Attention videos.” available at http://b.hatena.ne.jp/video.
[13] Y. Xu, L. Zhang, and W. Li, “Cubic analysis of social bookmarking for personalized recommendation,” in Proceedings of the 8th Asia-Pacific Web Conference (APWeb 2006), pp. 733–738, Jan. 2006.

[14] H. Wu, M. Zubair, and K. Maly, “Harvesting social knowledge from folksonomies,” in Proceedings of the 17th ACM Conference on Hypertext and Hypermedia (HYPERTEXT 2006), pp. 111–114, Aug. 2006.
[15] A. Hotho, R. Jaschke, C. Schmitz, and G. Stumme, “Information retrieval in folksonomies: Search and ranking,” in Proceedings of the 3rd European Semantic Web Conference (ESWC 2006), pp. 411–426, June 2006.
[16] Y. Yanbe, A. Jatowt, S. Nakamura, and K. Tanaka, “Can social bookmarking enhance search in the Web?,” in Proceedings of the 7th ACM/IEEE-CS Joint Conference on Digital Libraries (JCDL 2007), pp. 107–116, June 2007.
[17] M. Noll and C. Meinel, “Web search personalization via social bookmarking and tagging,” in Proceedings of the 6th International Semantic Web Conference and 2nd Asian Semantic Web Conference (ISWC/ASWC 2007), pp. 365–378, Nov. 2007.
[18] G. Koutrika, F. A. Effendi, Z. Gyöngyi, P. Heymann, and H. Garcia-Molina, “Combating spam in tagging systems: An evaluation,” ACM Trans. Web, vol. 2, pp. 22:1–22:34, Oct. 2008.
[19] “Attention Web pages.” available at http://b.hatena.ne.jp/hotentry.
[20] “New arrival Web pages.” available at http://b.hatena.ne.jp/entrylist.

Authors

Hiroyuki Hisamatsu received M.E. and Ph.D. degrees from Osaka University, Japan, in 2003 and 2006, respectively. He is currently an associate professor of Department of Computer Science, Osaka Electro-Communication University. His research work is in the area of performance evaluation of TCP/IP networks. He is a member of IEEE and IEICE.

Takahiro Hatanaka received B.E and M.E. degrees from Osaka Electro-Communication University, Japan, in 2009 and 2011, respectively. His research work is in the area of Web architecture.