Proposal of Combination System of Page-centric Communication and Search

Yuhki Shiraishi
Kyoto Sangyo University
603-8555 Japan
yuhki@cc.kyoto-su.ac.jp

Jianwei Zhang
Tsukuba University of Technology
305-8520 Japan
zhangjw@a.tsukuba-tech.ac.jp

Yukiko Kawai
Kyoto Sangyo University
603-8555 Japan
kawai@cc.kyoto-su.ac.jp

Toyokazu Akiyama
Kyoto Sangyo University
603-8555 Japan
akiyama@cse.kyoto-su.ac.jp

ABSTRACT

We present a novel system that combines the advantages of social communication and Web search by simultaneously discovering important pages and users. First, the system provides a communication interface attached to pages, which allows users to talk with each other in real time while browsing the same page, i.e., page-centric communication. Then, we extend the communication function by enabling real-time communication over different Web pages, and thus more opportunity is provided for communication. The system can efficiently provide two ranking lists of pages and users based on a hybrid structure of hyper links (page-page relationship) and social links (page-user relationship and user-user relationship). Thus, users can efficiently search for important pages as well as important users related to their queries and browsing pages through the ranking function, and immediately obtain useful information or knowledge from not only pages themselves but also from other users through the communication function. Experimental results show that the system has a potential to efficiently provide a novel page-centric communication and search experience to the users.

Categories and Subject Descriptors

H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval—*retrieval models, search process*

Keywords

Communication, Search, Social networking, Ranking

1. INTRODUCTION

People often use two methods to obtain information or knowledge from the Web. One is a search engine (SE) and the other is a social networking service (SNS). While SE has the advantages in high speed search and high data coverage, it has also the disadvantage in the absence of flexibility. On the other hand, communication through a SNS can overcome the problem of the absence of flexibility because users can freely ask questions and receive help from other users. However, the response may need much longer time and may have lower knowledge coverage than a SE.

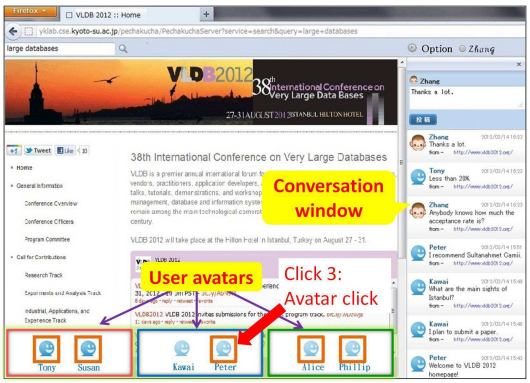

Figure 1: System snapshot: browsing and communication

Although several techniques on communication and search have been studied [1, 2, 3], the mentioned problems have been never solved. These studies have focused on either communication or search function. On the other hand, Twitter, Facebook, or other SNS are widely spreading on the Web. However, these services can be considered as a user-centric communication services, thus the users are often restricted to communicate with well-known or friendly users.

We have developed a system that combines the advantages of both social communication and Web search [4, 5]. As the developed system has a page-centric communication function among users who browse the same Web page, there is rare chance for the real-time communication when browsing unpopular Web pages. In this paper, we extend the communication function by enabling real-time communication over different Web pages, and thus more opportunity is provided for communication (Fig. 1). In this system, we

WOSS'12 in conjunction with VLDB'12, August 31, 2012, Istanbul, Turkey.
ACM .

propose a new ranking model, PURank (Page-user ranking model), that consists of two ranking lists of pages and users based on hyperlinks (page-page relationship) and social links (page-user relationship and user-user relationship) by advancing the concept presented in [4, 5]. Therefore, users can have a chance to communicate with the top n users ranked by the system depending on a query. The proposed system not only provides a communication function attached to all pages which allows users to talk with other users (Fig. 1) but also a ranking function for simultaneously evaluating pages and users (Fig. 2).

The system has two main features as shown in Fig. 1: 1) avatars are shown on each page, on which these avatars are represented as browsing users, passed users, and similar users; and 2) a conversation window is attached to each page so that users can communicate with each other. Whenever users have a problem, they can ask other users immediately and receive response from them in real time. The conversation window also has logs of previous communication, which may help users to obtain the answer from similar questions. Furthermore, the user can communicate with not only the users browsing the same page but also the users on other pages that have been browsed by him or her before.

This system can construct the links between a user and his or her browsing pages, and also the links between each other. Thus, we build a hybrid structure of pages and users including hyperlinks between pages, social page links between users and pages, and social user links between users. These social page links are generated based on browsing history and relationship between users and pages. When a user browse a page, a bidirectional link between the user and the page is generated. On the other hand, social user links are generated by the relationship between users based on analyzing their browsing logs and their similarities. Since not only hyperlinks but also social links are integrated into a hybrid structure, the system can calculate each eigenvalue of pages and users simultaneously and immediately for ranking.

2. SYSTEM OVERVIEW

In the preprocessing, the system crawls Web pages using Apache Nutch [6] and indexes them using Apache Solr [7]. To use this system, users are first required to simply install a toolbar (a browser plug-in). Then, users can configure their options such as the nickname and the avatar icon, etc. After receiving query keywords from a user, the system returns two ranking lists of pages and users. The user can follow either a page link or a user link to further surf the Web. Once the user browses a page or chat at that page with other users, the system records the information into a database, which is used for future ranking. The relationship between communication and ranking function is shown in Fig. 3.

The flow of the system is as follows:

1. After a user submits a query, the system returns a page list and a user list, in which the important pages and users related to the query are ranked to the top place (Fig. 2).

2. When the user follows a page link on the search result page (Click 1 on Fig. 2), the corresponding page is shown (Fig. 1).

   The avatars at the bottom of the page represent the users who are currently accessing this page (opaque icons in the red frame), the users who have browsed this page before (transparent icons in the blue frame), and other important users related to the query (transparent icons in the green frame). The top n users ranked by the system are automatically marked by bold outline borders.

   In the right conversation window, users can communicate with each other in real time or retrieve previous communication logs. Users can also communicate with the users having browsed the same page before.

3. When the user follows a user link on the search result page (Click 2 on Fig. 2), he can jump to the page that the clicked user is accessing. Similarly, the related users and the conversation window are presented on the page.

   At all pages, the user can click a user avatar and go to the page that the clicked avatar is accessing. Our system allows users to surf the web not only by following the hyperlinks but also by following the user avatars.

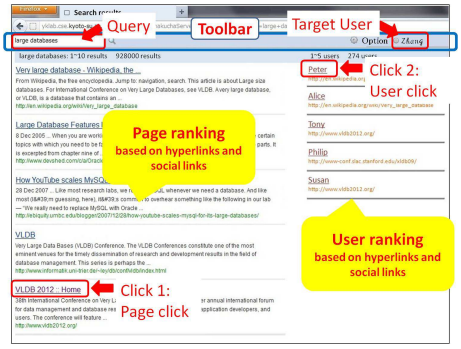

Figure 2: System snapshot: rankings of pages and users

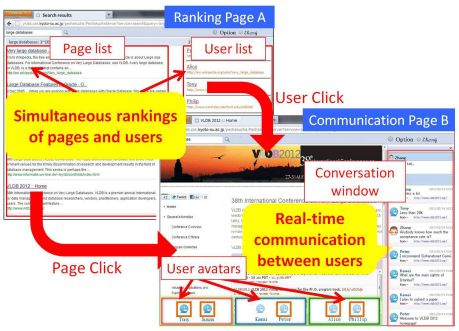

Figure 3: System snapshot: relationship between communication and ranking

3. PAGE-CENTRIC COMMUNICATION OVER DIFFERENT WEB PAGES

Users can communicate with other users in real time (Fig. 1). If the user has a question about something on the page, he or she can immediately ask other users by typing a sentence into the conversation window. Other users can see the

question in the conversation window and respond with an answer.

For some unpopular pages, the number of users accessing them may be small, even zero. In this case, few or no other users can be communicated with in real time. Our system also supports asynchronous communication, making the past conversation logs on a page available to the current users. A user may find the answer to a previous question similar to his or her current one from the conversation logs.

Furthermore, we extend the communication function by enabling real-time communication over different Web pages. The details are as follows: A target user can communicate with not only users browsing the same page that the target user is browsing now, but also

1. users having browsed the page that the target user is browsing now, and
2. users browsing the page that the target user browsed before.

To distinguish the chat of a user browsing different Web pages from that of a user browsing the same page, the former is displayed transparently and is not saved to the log.

In order to encourage users to talk more, if a user enters more sentences, his or her user avatar is enlarged.

3.1 Implementation of communication function

The communication function is achieved by the collaboration between the server side and the client side. The server side receives and processes the client requests. The client side sends users' requests to the server.

3.1.1 Server side

The server side has been built by using Apache Tomcat 6.0.32 and Java Servlet on JDK 1.7. The servlet can perform parallel processing for multiple requests to make a "user thread" for each request: the user thread 1) receives the requests from the client, 2) parses them, 3) performs the corresponding action, 3) returns information to the requesting user, and 4) sends information to other users. The browsing and communication logs are stored in the server database using PostgreSQL 9.1.1.

3.1.2 Client side

The client side is implemented as a toolbar (browser plug-in) as shown in Fig. 1 and Fig. 2. The browser interface is programmed using XUL (eXtensible User-interface Language) and the development is programmed using JavaScript. A user is connected to the server using an asynchronous communication program running in an Ajax script. After the user logs in to the server, the client program sends the user's requests to the server. The client program continuously polls the read buffer by using Comet. If the server responses immediately after receiving a request, the request with polling is used. This polling operation reduces network traffic because the client need not periodically send requests to confirm the server's status.

4. PURANK: PAGE-USER RANKING MODEL

After a user submits a query, the system first retrieves the pages including the query keywords from the crawled pages. Then the system specifies the users who are accessing and have browsed these pages from the database that records the browsing history. Next, a page-user graph, its adjacency matrix, and probability transition matrix are constructed. Finally, PURank scores of pages and users are calculated.

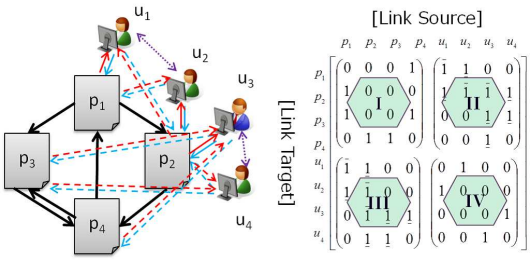

Figure 4: Page-user graph and adjacency matrix

4.1 Construction of page-user graph and adjacency matrix

A page-user graph is first constructed. The nodes of the graph consist of pages and users. The links of the graph consist of four parts: hyperlinks between pages, social page links from users to pages, social page links from pages to users, and social user links from users to users. Based on an example shown in Fig. 4, we explain four parts of this graph (left part of Fig. 4) and its corresponding adjacency matrix (right part of Fig. 4).

1. Hyperlinks between pages (Part I)

   If a page p_i has a hyperlink to another page p_j, the element of $matrix(p_j, p_i)$ is set to 1. For example, page p_1 links to page p_2, the element of $matrix(p_2, p_1)$ becomes 1.

2. Social page links from users to pages (Part II)

   If a user u_i is browsing a page p_i (blue real lines in the left graph of Fig. 4) and has browsed a page p_j before (blue dotted lines in the left graph of Fig. 4), the elements $matrix(p_i, u_i)$ and $matrix(p_j, u_i)$ are marked as $\overline{1}$ and $\underline{1}$. For example, if user u_1 is browsing page p_1 and has browsed page p_2 before, $matrix(p_1, u_1)$ and $matrix(p_2, u_1)$ become $\overline{1}$ and $\underline{1}$ respectively.

3. Social page links from pages to users (Part III)

   This part represents whether a page is being browsed or was browsed by a user. The links of Part III (red real lines and red dotted lines in the left graph of Fig. 4) have reverse directions to those of Part II. Part III of the matrix is the transposed matrix of Part II.

4. Social user links from users to users (Part IV)

   If two users have similar interests, a bidirectional social link is generated between them (purple dotted lines in the left graph of Fig. 4). Our implementation is based on comparing the overlap of pages that two users have browsed. That is to say, the elements of Part IV of the matrix are calculated based on the elements of Part II as follows:

   $$(u_i, u_j) = (u_j, u_i) = \begin{cases} 1 & \text{if } \sum_k (p_k u_i \wedge p_k u_j) / \sum_k (p_k u_i \vee p_k u_j) > \tau \\ 0 & \text{else} \end{cases}$$

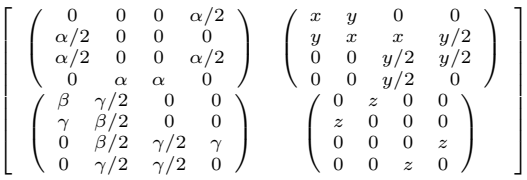

Figure 5: Probability transition matrix

where $p_k u_i$ and $p_k u_j$ are the information about whether u_i and u_j have browsed p_k or not, which can be obtained from Part II of the matrix.

If the overlap of the pages that two users have browsed is larger than a threshold τ, the corresponding element of Part IV is set to 1. Supposing $\tau = 0.5$ in the example shown in Fig. 4, the elements of (u_1, u_2), (u_2, u_1), (u_3, u_4), (u_4, u_3) become 1, because $(u_1, u_2) = (u_2, u_1) = 2/2 > 0.5$, $(u_3, u_4) = (u_4, u_3) = 2/3 > 0.5$.

4.2 Construction of probability transition matrix

We next describe the construction of probability transition matrix from the adjacency matrix. For the example in Fig. 4, it can be transformed to a probability transition matrix as shown in Fig. 5. α, β and γ are the weights assigned to the hyperlinks between pages, the social links from pages to users who are browsing them, and the social links from pages to users who have browsed them before. x, y and z are the weights assigned to the social links from users to pages that they are browsing, the social links from users to pages that they have browsed before, and the social links from users to similar users. Notice that the conditions of $\alpha + \beta + \gamma = 1$ and $x + y + z = 1$ should be satisfied.

The weights are assigned by users according to their requirements. For example, if a user attaches more importance to the links representing the current browsing relationship (i.e., gives higher values to β and x), the pages that are currently being browsed by many users are highly evaluated.

4.3 PURank calculation

We next calculate the scores of pages and users called PURank. It can be calculated by the following equation.

$$\boldsymbol{r} = d\mathbf{A}\boldsymbol{r} + \frac{(1-d)}{n}\boldsymbol{e} \tag{1}$$

where n is the total number of pages and users in the page-user graph, d is a damping factor given by a user, $\mathbf{A}$ denotes the probability transition matrix, and $\boldsymbol{e}$ denotes an n-dimension column vector with all elements set to 1. All elements of $\boldsymbol{r}$ are initialized to 1 and the calculation terminates when $\boldsymbol{r}$ converges.

Although the calculation equation is similar to PageRank [8] and ObjectRank [9], the sense of our PURank is different from them in that PURank is calculated based on the hybrid space of both pages and users, instead of only pages [8] or only database entries [9]. By simultaneously providing the rankings of both pages and users, the system can help users discover not only important pages but also important users so that they can find appropriate users to further communicate with. Important pages are the ones that are hyperlinked by many other important pages and are being browsed or have been browsed by many important users. Important users are the ones who are browsing or have browsed many important pages and are linked by many other important users.

In the implementation, the PURank scores of pages and users can be obtained by calculating the eigenvector of the probability transition matrix corresponding to the largest eigenvalue. For the acceleration of eigenvector calculation, we use a library SLEPc that can support parallel computing. For a matrix with 100 thousand * 100 thousand elements, the calculation time can be controlled to tens of seconds by using 8 processors.

5. EXPERIMENTS

In this section, we evaluate the developed system and verify the PURank.

First, we have evaluated the preliminary version of the developed system[1] for the basic evaluation especially focusing on the communication function. The system provides both the communication function over different Web pages and the ranking function based on the number of accessing Web pages, but provides neither user rankings results nor page rankings results based on hyperlinks and social links. This limitation is exposed because the first evaluation of the system should focus on the communication function among users who use the system for the first time for about ten minutes, and who have similar interests. In this case, it is difficult to evaluate the user ranking results because of the short period of the usage of the system. Therefore, the verification of the PURank is carried out by analyzing the ranking results by example.

5.1 Evaluation of developed system

In these experiments, according to the former usability experiments [11], we verify especially the communication function of the system by analyzing the questionnaire results for ten participants, after the realization of pseudo environment in which the system could improve the user experience of searching tasks. Thus, the followings three case studies are analyzed for two groups: each group includes five participants, at least one beginner participant with searching experience less than three years and at least one experienced participant with searching experience more than ten years.

1. Task becoming enjoyable or useful by communicating with each other
2. Difficult task obtaining the useful information by non-professional user
3. Task becoming efficient by collaboration with each other

In these cases, we suppose that a few users who utilize the system face the similar searching task simultaneously; this situation would be realized after the developed system spreads to enough users. Then, these users have a chance to communicate with each other by using the communication function of the system.

In Case 1, we suppose that users search something enjoyable or useful information, e.g., a plan for trip, play, etc.

[1]This system is now available in [10]. Firefox 13 or newer version is required to try this version.

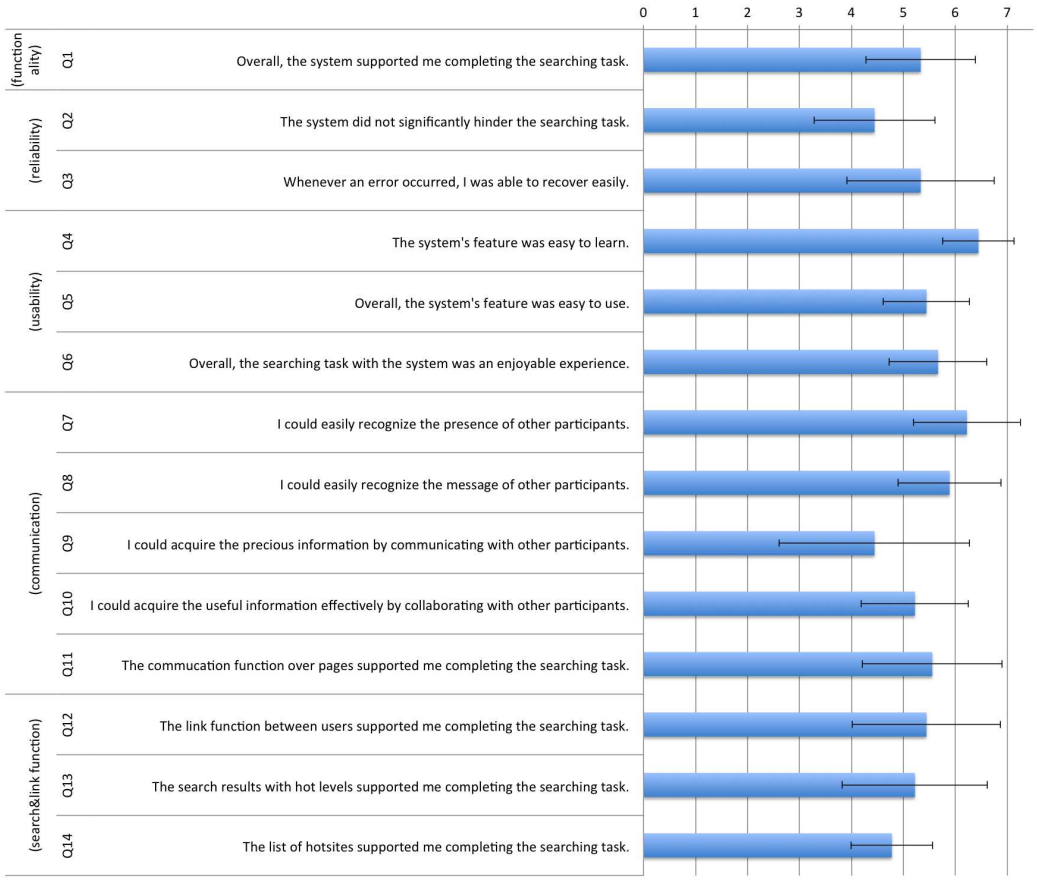

Figure 6: Questionnaire and evaluation results: 1 (strongly disagree) $\sim$ 4 (neither agree nor disagree) $\sim$ 7 (strongly agree), N=9, the mean μ and the standard deviation $\mu - \sigma \sim \mu + \sigma$ are shown.

In Case 2, we suppose that it is difficult for users to find the appropriate queries for useful information due to the lack of the professional knowledge, e.g., in case of searching the function of unknown programming language, etc.

In Case 3, we suppose that users need to search some many things concerning the same topic, e.g., the list of computer languages, countries, etc.

5.1.1 *Experimental method*

At first, all participants are explained the way to use the system for about ten minutes. Next, five participants of each group are asked to solve each task within ten minutes. In this experiment, two tasks for Case 1, one task for Case 2, and one task for Case 3 are presented sequentially.

The communication with each other is allowed only by using the system. The searching method is also restricted by using the toolbar (Fig. 2), by which the system shows the top ten results with the hot levels (which represent how may users access the Web page according to eleven levels). The system also has a "hotsites" button, which shows the list of Web pages that have highest hot levels. In this version of the system, the avatars display on the center of Web page so that the beginner user can find out the avatars easily.

After the experiments, the questionnaire shown in Fig. 6 are completed by the participants with seven-level Likert scale [12] with the reason. The contents of questionnaire are determined based on the ISO/IEC TR 9126 standard for product quality [13] and the collaboration catalog [14] to verify the effectiveness of the system.

5.1.2 *Results and discussion*

Fig. 6 shows the evaluation results with the mean μ and the standard deviation σ for nine valid completions of questionnaire. These results show the effectiveness of the system because $\mu >= 4$ for all questionnaire and $\mu - \sigma >= 4$ for all but Q2 and Q9.

The reason of low score of Q2 is that avatars are shown in the center of the page in case of changing the browsing Web site. Thus, we will restrict the display area of avatars to avoid the hindrance of the browsing experience in the next version as explained in section 2.

The reasons of low score of Q9 are as follows: (1) other participants do not give me any useful information, and (2) only top ten sites are shown as the search results. As for (1), some participants state that they can acquire useful information by other participants, thus we consider that the system works effectively in the case study 2. On the other hand, (2) is cased by the experimental conditions, then we will get rid of this restriction in the future work.

As the evaluations of Q6 and Q10 are positive, the system works functionally in the case study 1 and 3. Moreover, as

Table 1: Ranking results

Case	Rank of pages	Rank of users
(1)	1:p_4, 2:p_3, 3:p_1, 4:p_2	
(2)	1:p_2, 2:p_1, 3:p_4, 4:p_3	1:u_1, 2:u_2, 3:u_3, 4:u_4

the evaluations of Q11 and Q12 are positive, the communication function over different Web pages works well.

5.2 Verification of PURank

We verify the PURank by the basic experiments as an example of a page-user relationship as stated below.

First, for the acceleration of eigenvector calculation, we investigate the possibility of using a library SLEPc that supports parallel computing. As a result, for a matrix with 100 thousand * 100 thousand elements, the calculation time can be controlled to tens of seconds by using eight processors.

Next, as for an example used in section 4, we compare the two ranking results as follows:

1. Considering only hyperlinks; In Fig. 5, $\alpha = 1, \beta = \gamma = x = y = z = 0$.
2. Considering also social links; In Fig. 5, we set $\alpha = 0.2, \beta = 0.6, \gamma = 0.x = 0.4, y = 0.2, z = 0.4$ in this example.

Case 2 means that the weights assigned to the social links from pages to users who are browsing them (β) and the social links from users to pages that they are browsing (x) are higher than those assigned to the social links from pages to users who have browsed them before (γ) and the social links from users to pages that they have browsed before (y).

As a result, the rank values of each nodes are as follows:

1. $[p_1, p_2, p_3, p_4] = [0.2, 0.1, 0.3, 0.4]$
2. $[p_1, p_2, p_3, p_4, u_1, u_2, u_3, u_4]$
   $= [0.129, 0.196, 0.044, 0.066, 0.186, 0.176, 0.121, 0.080]$

Table 1 shows the comparison of ranking results. From these results, Case 1 shows the page p_4, which is most linked from other pages, is top ranked, but Case 2 shows the page p_2, which is most visited by users, is top ranked. This shows that the PURank can provide the important page based on the users' browsing history.

On the other hand, Case 2 shows the pages p_2 and p_3, which are visited by users currently, are ranked higher than the pages p_1 and p_4, which are visited by users previously, and also the users u_4, who visited the pages only previously, is lowest ranked. This shows that the weight values assigned to the social links affect the ranking results, which can be adjusted according to the requirements.

6. CONCLUSION AND FUTURE WORKS

In this paper, we proposed a novel system for simultaneous realization of page-centric communication and search. In this system, the PURank is calculated by using two ranking lists of pages and users by analyzing a hybrid structure of hyperlinks and social links, and a new page-centric communication function over different Web pages is realized. As a result, the system can provide users an efficient search function for important pages as well as important users related to their queries through PURank. Experimental results show the system has a potential to efficiently provide a novel page-centric communication and search experience to the users.

In the future works, we will improve the evaluated system based on the experimental results and verification experiments will be carried out for much more participants. Furthermore, we will carry out the experiments for evaluating the page ranking and users ranking based on the PURank by using the developed system.

7. ACKNOWLEDGMENTS

This work was partially supported by SCOPE (Ministry of International Affairs and Communications, Japan) and by JSPS KAKENHI Grant Number 24780248.

8. REFERENCES

[1] E. Agichtein, E. Brill, and S. Dumais. Improving Web Search Ranking by Incorporating User Behavior Information. In *Proc. SIGIR 2006*, pp.19-26, 2006.

[2] R. W. White, M. Bilenko, and S. Cucerzan. Studying the Use of Popular Destinations to Enhance Web Search Interaction. In *Proc. SIGIR 2007*, pp.159-166, 2007.

[3] X. Si, E. Y. Chang, Z. Gyongyi, and M. Sun. Confucius and Its Intelligent Disciples: Integrating Social with Search. In *Proc. VLDB 2010*, pp.1505-1516, 2010.

[4] Y. Matsui, Y. Kawai, and J. Zhang. Hybrid Web search with social communication service, In *Proc. the International Multi-conference of Engineers and Computer Scientists 2011 (IMECS 2011)*, pp.687-962, 2011.

[5] Y. Matsui, Y. Kawai, and J. Zhang. Page as a Meeting Place: Web Search Augmented with Social Communication, *Intelligent Control and Innovative Computing*, *Lecture Notes in Electrical Engineering*, vol.110, pp.303-317, 2012.

[6] Apache Nutch. `http://nutch.apache.org/`

[7] Apache Solr. `http://lucene.apache.org/solr/`

[8] S. Brin and L. Page. The Anatomy of a Large-scale Hypertexual Web Search Engine. *Computer Networks*, vol.30, no.1-7, pp.107-117, 1998.

[9] A. Balmin, V. Hristidis, and Y. Papakonstantinou. ObjectRank: Authority-Based Keyword Search in Databases. In *Proc. VLDB 2004*, pp.564-575, 2004.

[10] Combination System of Search and Communication. `http://klab.kyoto-su.ac.jp/~mito/index.html`

[11] M. Heinrich, F. Lehmann, T. Springer, and M. Gaedke, Exploiting Single-User Web Applications for Shared Editing - A Generic Transformation Approach. In *Proc. WWW 2012*, pp.517-526, 2012.

[12] R. Likert, A technique for the measurement of attitudes. *Archives of Psychology*, vol.22, no.140, pp.5-55, 1932.

[13] ISO/IEC, Software engineering - Product quality - Part 1: Quality model. *ISO/IEC 9126-1*, 2001.

[14] D. Pinelle, C. Gutwin, and S. Greenberg, Task analysis for groupware usability evaluation: Modeling shared-workspace tasks with the mechanics of collaboration. *ACM Trans. Comput.-Hum. Interact.*, vol.10, no.4, pp.281-311, 2003.