Hybrid Web Search with Social Communication Service

Yuya Matui, Yukiko Kawai, Jianwei Zhang

***Abstract*—When retrieving information, a user generally uses search engines or social communication services on the web. The advantages of search engines such as Google and Yahoo! are high speed search and high coverage. However, search results do not satisfy all users because they may have different needs and different levels of knowledge. On the other hand, the advantage of communication services such as SNS is to provide high quality information by means of user communication. However, it takes longer time than search engines. We develop a search system combining the merits of searching and social communication. This system includes a page ranking algorithm based on the analysis of a hyperlink structure and a social link structure, and a communication interface attached to a page which allows real-time users to communicate with each other. By our system, users can quickly search not only for popular web pages but also for other users currently accessing them.**

***Index Terms*—web search, social communication, real-time processing, link analysis**

I. Introduction

SEARCH engines and social networking services are commonly used in daily life. Both types of services have some advantages as well as several problems.

Search engines such as Google and Yahoo! can search for and almost instantaneously find web pages from among the huge volume of web pages, but long complex queries can result in bad results, important information on the web may not be immediately obvious, and the results may not be suitable for the user's knowledge level. A long complex query, such as a sentence, generally does not result in useful search results. Therefore, the user has to scan the entire search engine results page (SERP) to find results of interest. If the user comes across unknown keywords on a page and he or she is not an expert on the topic, more searches must be conducted and more, possibly irrelevant, pages must be read in order to understand the meanings of the words.

In contrast, submission of a long complex query to a social networking service based on human communication such as Facebook, Mixi, and Twitter can result in more useful results. Communication of human knowledge through such services can overcome the problems described above. However, the response time is much longer than from a search engine because when to be able to get a reply depends on other users' behaviors. Furthermore, the data coverage is much lower than from a search engine, which can access billions of pages.

We have developed a system that combines the merits of search and social communication. It can be used to quickly search not only for web pages but also for other users currently accessing those pages ("real-time users"). Conventional web searches (Fig.1.(a)) retrieve web pages and rank them based on hyperlinks. Social networking services (SNS) (Fig.1.(b)) gather information about the users, and users make use of this information to find experts. Our hybrid search with real-time user communication (Fig.1.(c)) gathers both web pages and user information, and connects the users through the accessed pages.

Our system offers the following advantages:

1) Our ranking method based on both hyperlinks and social links can provide a list of pages which are important and attractive.
2) Each hyperlink on a page (a SERP or a common page) is annotated by the number of users currently accessing its corresponding linked page.
3) Our communication interface attached to a page enables users to talk with other users in real time, or browse the past communication logs asynchronously.
4) Users are allowed to highlight and share important information of a page, so that other users can immediately find the key portion of the page.
5) In summary, our hybrid search system with social communication service can provide users with the information of high quality and high coverage.

In this paper, we describe our hybrid search system and its implementation. Section II presents related work and Section III gives an overview of the system. Section IV explains the page ranking algorithm and Section V presents the real-time communication function. Section VI describes the implementation of our system. Section VII presents the experimental evaluation. We conclude with a summary of the key points and a mention of future work in Section VIII.

II. Related work

There have been many studies on search and social communication [1], [2], [3]. For example, a recently proposed approach for combining search and social network [1] searches not only for web pages but also for experts. The user can send queries using a communication tool such as email. Another approach [2] identifies popular pages on the basis of users' browsing histories. The result is a SERP with links to websites frequently visited by other users with similar queries. Although these approaches may be able to locate expert users for each type of content, they do not solve the fundamental problems described in Section I.

Other approaches are to use community-based recommendations [4], [5], [6]. For example, [4] and [5] proposes a small community of searchers to do the collaborative filtering. These searchers search for target information or web pages by collaboratively improving the queries. However, a purpose-built interface is needed for real-time processing, and the searchers cannot use people outside their community in the collaboration process.

Kyoto Sangyo University, i1054056@cc.kyoto-su.ac.jp
Kyoto Sangyo University, kawai@cc.kyoto-su.ac.jp
Kyoto Sangyo University, zjw@cc.kyoto-su.ac.jp

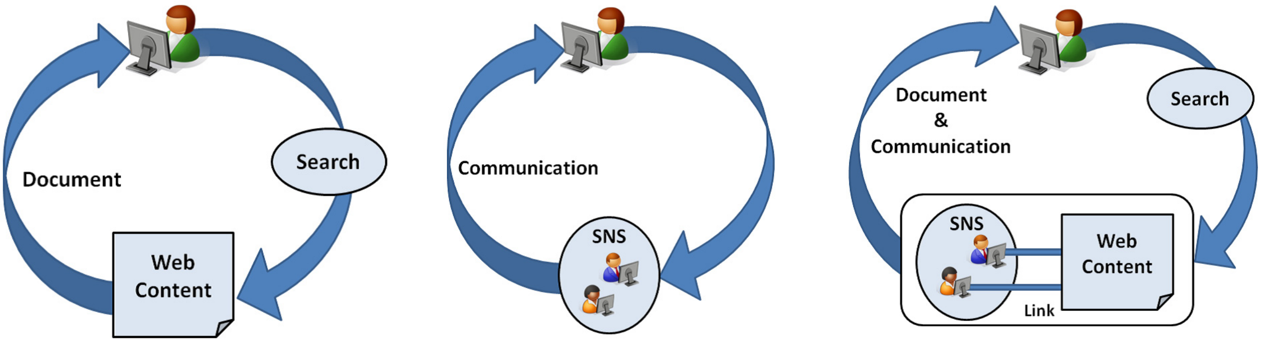


Fig. 1. (a) Conventional web search (b) Social networking service (c) Hybrid web search

The detection of user characteristics and their application to news and auction sites have also been studied [7], [8], [9]. The detection analysis uses the history of a user's behavior such as the click history. However, this analysis cannot detect an appropriate expert for a particular page or query.

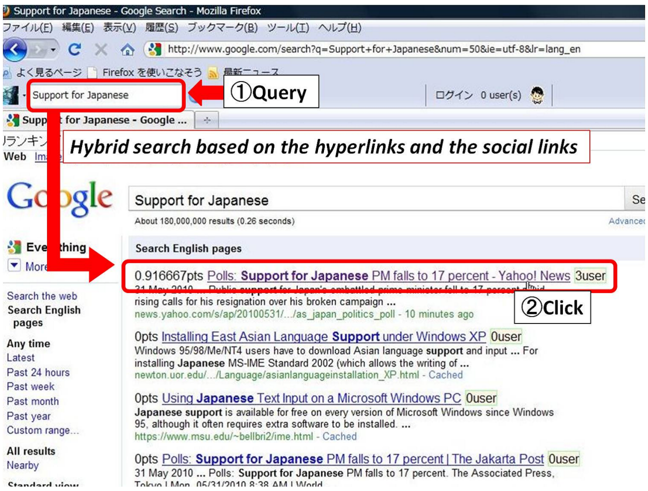
The use of chat for user communication is another kind of approaches [10], [11], [12]. These services support connections between users, and users accessing the same page can chat. However, because only communication service is provided, the users have to find interesting and popular pages for connecting with other users.

Our hybrid search system not only identifies other users accessing the same page but also provides high-quality information by analyzing the structure of hyperlinks and real-time user links. Furthermore, it shows the number of users currently accessing each page, and the users can communicate with each other.

III. System Overview

The flow of hybrid web search with social communication is as follows:

1) After a user submits a query, he or she receives a SERP ranked by the values of the hyperlinks and the social links (Fig. 2).
   The pages are first ranked based on the analysis of the hyperlink structure, and then re-ranked based on the analysis of the social links. The hyperlinks on the SERP are also annotated by the number of users currently accessing the corresponding page.
2) When the user clicks a hyperlink on the SERP, the corresponding page is shown with a communication window on the right for chat (Fig. 3).
   The users currently accessing this page are represented by avatars. The user can directly ask another real-time user a question by using the communication window. The communication window also has logs of previous communication, i.e., the history of communication between users, which can provide the answers to previous, possibly similar, questions. The important information can also be highlighted by a user and shared to other users, which can help them efficiently detect the key portion of the page.
3) Our system also provides each hyperlink on the accessed page with the number of users currently accessing its corresponding page, so that the user can recursively find and communicate with more users through the links (Fig. 4).
   Using our system, a user may reach a communication with a user with an appropriate level of knowledge by following the link structure (Fig. 5).


Support for Japanese - Google Search - Mozilla Firefox
ファイル(F) 編集(E) 表示(V) 履歴(S) ブックマーク(B) ツール(T) ヘルプ(H)
よく見るページ Firefox を使いこなそう 最新ニュース
①Query
ログイン 0 user(s)
Supp t for Japanese - Google ...
ランキン
Web Im
Hybrid search based on the hyperlinks and the social links
Se
Advance
Eve thing
More
②Click
Search the web
Search English pages
Any time
Latest
Past 24 hours
Past week
Past month
Past year
Custom range...
All results
Nearby
Standard view

http://www.google.com/search?q=Support+for+Japanese&num=50&ie=utf-8&lr=lang_en

Support for Japanese

About 180,000,000 results (0.26 seconds)

Search English pages

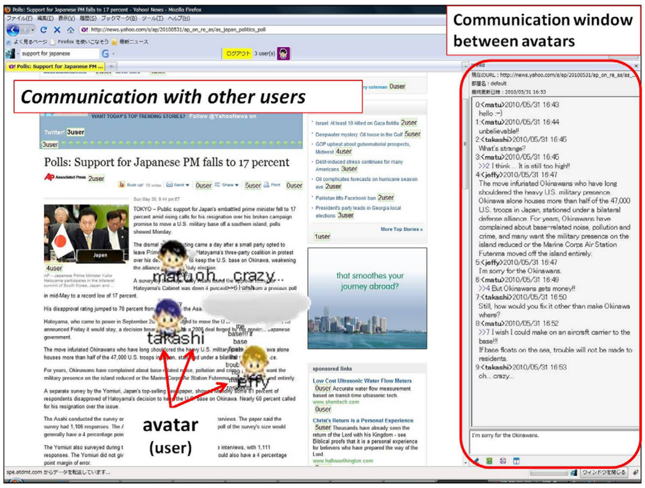
0.916667pts Polls: **Support for Japanese** PM falls to 17 percent - Yahoo! News 3user
31 May 2010 ... Public support for Japan's embattled prime minister fell to 17 percent amid rising calls for his resignation over his broken campaign ...
news.yahoo.com/s/ap/20100531/.../as_japan_politics_poll - 10 minutes ago

0pts Installing East Asian Language **Support** under Windows XP 0user
Windows 95/98/Me/NT4 users have to download Asian language **support** and input ... For installing **Japanese** MS-IME Standard 2002 (which allows the writing of ...
newton.uor.edu/.../Language/asianlanguageinstallation_XP.html - Cached

0pts Using **Japanese** Text Input on a Microsoft Windows PC 0user
Japanese support is available for free on every version of Microsoft Windows since Windows 95, although it often requires extra software to be installed. ...
https://www.msu.edu/~bellbri2/ime.html - Cached

0pts Polls: **Support for Japanese** PM falls to 17 percent | The Jakarta Post 0user
31 May 2010 ... Polls: **Support for Japanese** PM falls to 17 percent. The Associated Press, Tokyo | Mon, 05/31/2010 8:38 AM | World

Fig. 2. Search results with the number of access users


Communication window between avatars
Communication with other users
Polls: Support for Japanese PM falls to 17 percent
matuoh
crazy
takashi
jeffy
avatar (user)
3user
4user
2user
5user
0user
1user
that smoothes your journey abroad?
ウィンドウを閉じる

現在のURL：http://news.yahoo.com/s/ap/20100531/ap_on_re_as/as_...
部屋名：default
最終更新日時：2010/05/31 16:53

0:<matu>2010/05/31 16:43
hello :-)
1:<matu>2010/05/31 16:44
unbelievable!!
2:<takashi>2010/05/31 16:45
What's strange?
3:<matu>2010/05/31 16:45
>>2 I think ... It is still too high!!
4:<jeffy>2010/05/31 16:47
The move infuriated Okinawans who have long shouldered the heavy U.S. military presence. Okinawa alone houses more than half of the 47,000 U.S. troops in Japan, stationed under a bilateral defense alliance. For years, Okinawans have complained about base-related noise, pollution and crime, and many want the military presence on the island reduced or the Marine Corps Air Station Futenma moved off the island entirely.
5:<jeffy>2010/05/31 16:47
I'm sorry for the Okinawans.
6:<matu>2010/05/31 16:49
>>4 But Okinawans gets money!!
7:<takashi>2010/05/31 16:50
Still, how would you fix it other than make Okinawa where?
8:<matu>2010/05/31 16:52
>>7 I wish I could make on an aircraft carrier to the base!!!
If base floats on the sea, trouble will not be made to residents.
9:<takashi>2010/05/31 16:53
oh... crazy...

I'm sorry for the Okinawans.

Fig. 3. Web page with a communication window

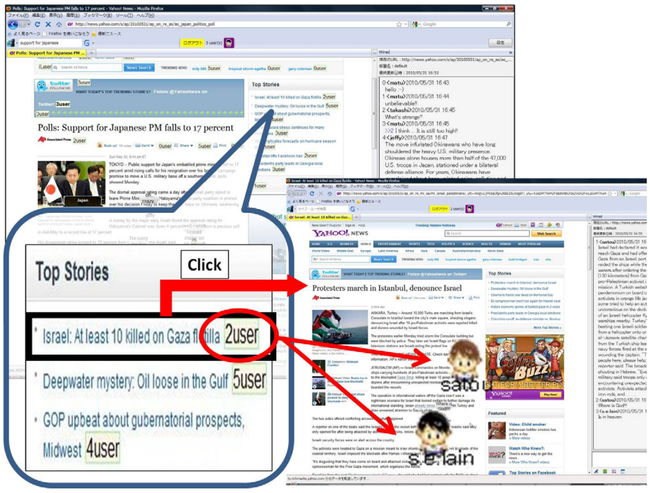

Top Stories
Click
Israel: At least 10 killed on Gaza flotilla 2user
Deepwater mystery: Oil loose in the Gulf 5user
GOP upbeat about gubernatorial prospects, Midwest 4user
Polls: Support for Japanese PM falls to 17 percent
Protesters march in Istanbul, denounce Israel
sato
s.e.lain

Fig. 4. Recursive user communication through deeper links

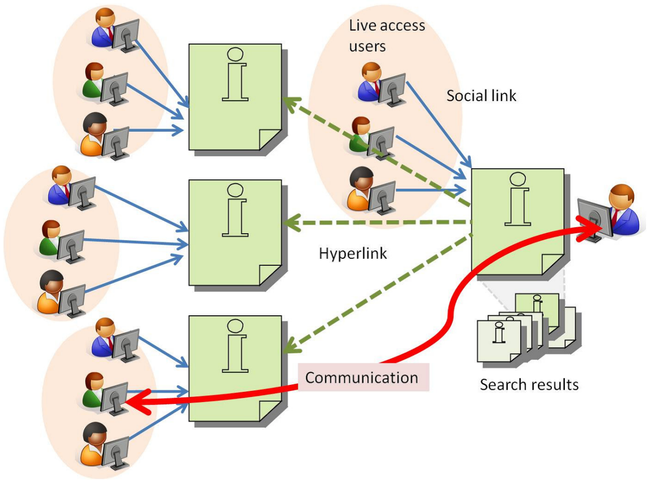


Fig. 5. Flow of finding a suitable expert

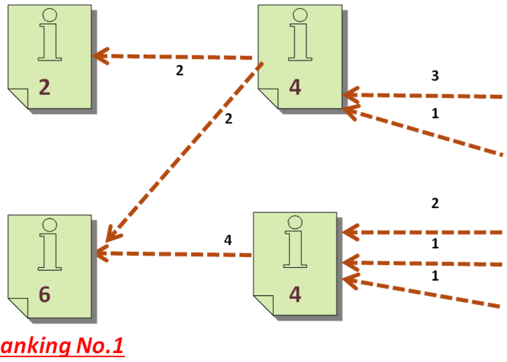


Fig. 6. Ranking based on the hyperlink structure

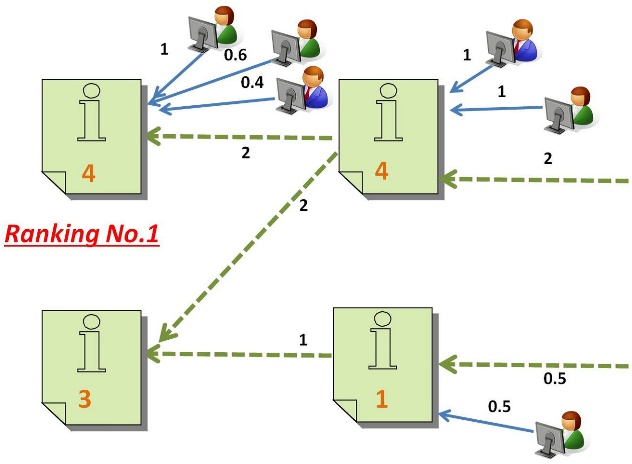


Fig. 7. Ranking based on the social link structure

IV. Ranking based on hyperlinks and social links

Our ranking method uses both hyperlinks and social links. Social links are the links from the real-time users to the pages. Page ranking has two major steps. First, the weight of each page is calculated considering the hyperlink structure. Then, each page is re-ranked based on the quality and quantity of the users currently accessing it. In this section, first we explain the method for calculating the weight of hyperlinks, and then describe the method for calculating the weight of social links.

A. Ranking based on hyperlinks

Our ranking method is based on the idea of PageRank [13]. Each page has a pagerank deriving values equally divided by the number of outlinks from its parent pages (Fig. 6). The pagerank value is given by

$$PR_p = \sum_{q \in parent(p)} PR_q / N_q \tag{1}$$

where p is a target page, q is a page which links to p, $parent(p)$ represents the set of q, PR_p and PR_q represent the pagerank values of p and q respectively, and N_q is the number of outlinks from q.

B. Ranking based on social links

As described above, our ranking method uses the hyperlink and social link structure, and is based on the PageRank algorithm. Each page has two values: one is the weight of the hyperlinks calculated using Equation (1), and the other is the weight of the social links (Fig. 7). The weight of the social links includes the weights based on the quality and quantity of users currently accessing a page, and the weights derived from the page's parents. Each page has a UR (UserRank) weight value given by

$$UR_p = \sum_{u \in user(p)} W_u + \sum_{q \in parent(p)} UR_q / N_q \tag{2}$$

where u is a user currently accessing page p, $user(p)$ represents the set of those users accessing p, W_u is the quality of user u, q is a parent page of p, UR_p and UR_q are the UR weight values of p and q respectively, and N_q is the number of outlinks from q. W_u is initiated as 1. Considering that a user's effect decreases as the access time increases, we calculate W_u by the following equation:

$$W_u = \begin{cases} 1-(T_{now}-T_{access})/T & \text{if } (T_{now}-T_{access}) < T \\ 0 & \text{if } (T_{now}-T_{access}) \geqslant T \end{cases} \quad (3)$$

where T is the useful-life of users, T_{now} is now, and T_{access} is the time when the page was accessed.

V. Communication through a page

Many search algorithms have been proposed for precision improving. However, even if users can get a good SERP, they often find information on a linked page is confusing. This is because the content of a web page does not have sufficient information for all users. Our system solves this problem by enabling the user to communicate with other users currently accessing the page.

A. Real-time communication

Because the hyperlinks on each page are annotated with the number of users currently accessing them, the system enables a user to find even more real-time users. The user can thus connect with other users by accessing pages deeper from the SERP. The user can communicate in real time with other users accessing the same page. If the user has a question about something on the page, he or she can immediately ask other users about it by typing a query into the communication window. The other real-time users see the question in the communication window and can respond with an answer.

If a page is being accessed by a large number of users and its communication window has a large number of dialogues, it is difficult for users to find an expert with an appropriate level of knowledge. To solve this problem, we design the system by allowing users to freely create a "room" in which real-time users can talk with respect to a specific topic.

B. Asynchronous communication

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. To solve this problem, our system is also designed to support asynchronous communication. The server maintains a communication log for each page, and the log is made available to the current users of that page. The log is presented on the basis of content. A user searching this log may find the answer to a previous query similar to his or her current question.

C. Text highlighting

A user can highlight text that he or she considers important by selecting it and clicking a popup command, and other users can efficiently detect the important parts of the page. The highlighting remains on the page, and the system can automatically scroll down and present the highlighted text after the page is accessed. We also plan to make use of the highlighting for improving the ranking of pages. If there are many highlighted text in a page, the page is ranked higher.

VI. Implementation

We implemented our hybrid search system [14] as illustrated in Fig. 8. It consists of the server side and the client side.

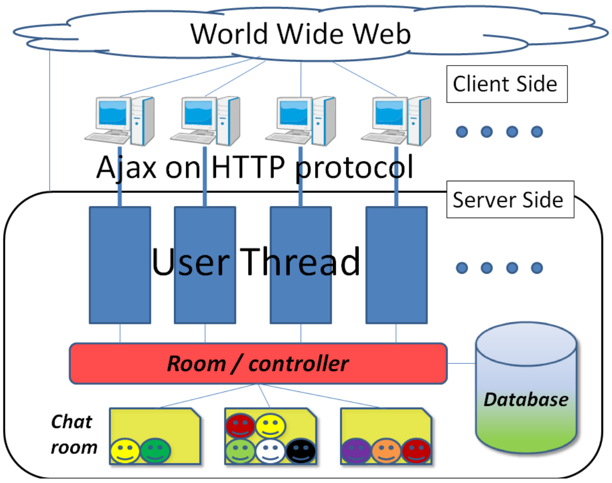


Fig. 8. System structure

A. Server side

The server side receives and processes the client requests. It was built using Apache Tomcat 6.0.18 and Java Servlet on JDK 1.6. The servlet can perform parallel processing for multiple requests because it makes a "user thread" for each request. The servlet consists of two parts: a search servlet and a communication servlet.

1) search servlet: The search servlet extracts the query from line of request. It then retrieves web pages by using the query. Several search engines (Google, Yahoo!, Bing, etc.) are supported in this implementation. After the servlet gets the search results' URLs, it retrieves information about the URLs: the quality and quantity of current users from the database. Finally, it sorts the search results by PR values, extracts the top ones, and re-ranks them based on UP values.

2) communication servlet: The user thread on the communication servlet receives the requests from the clients and parses them. Table I shows the requests supported by the user thread, the corresponding action, and the information sent to the other clients and to the requesting client. The user thread sends commands to the database objects and room operation objects in accordance with Table I. Most user requests are stored as chat room objects. The chat room data includes the URL, the room name, the communication log, the user information (avatar name, etc.), and the server response. The user thread operates the chat rooms by using the room operation class. Asynchronous communication was implemented by unblocking the I/O using Commet technology. With this setup, the server can send the server status to the clients in real time. The communication logs are stored in the server database.

B. Client Side

To use this search and communication system, a user needs to simply install a plug-in. The plug-in was developed using the expanded functions of Firefox 3.0, a cross-platform browser. The browser interface was programmed using XUL (extensible user-interface language), which is an expanded version of XML (extensible markup language), and the development was programmed using JavaScript.

TABLE I
REQUESTS THAT THE SERVER CAN PROCESS

request	action	info sent to other users	info sent to requesting users
Log in	add user data to room and DB	notification	other users' data and chat log
Log out	delete user data from room and DB	notification	–
Chat write	store query in room log	message	–
Chat read	store reply in room log	–	–
Get room info	provide room data	–	room data
Change room	delete user data from room and DB	notification	new room's user data and chat log
Change user name	change user name	notification	–
Share text	–	highlighted text	–

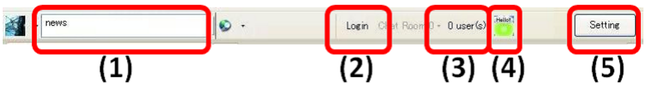


Fig. 9. toolbar interface

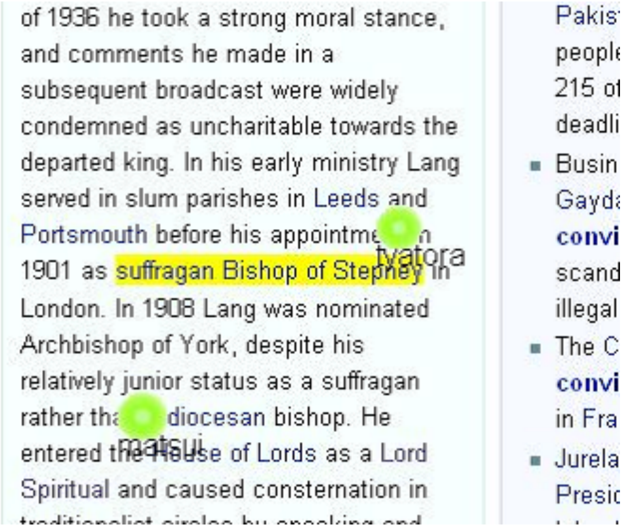
TABLE II
REQUESTS THAT THE CLIENTS CAN SUBMIT

request (command format)	info sent
roomsInfo	room name and number
	of users accessing the same URL
usersInfo	data on users in the same room
logs	room chat log
enter	data on newly entered users
exit	data on users recently logged out
message	message in new comment,
	data on user who wrote it
newRoom	data on newly created room
changeName	modified user data
shareText	locations of highlighted text

1) basic operation: A user connects 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. Table II shows the supported client requests. The client program continuously polls the read buffer. If the server is to reply immediately after receiving a request, request with polling is used. This polling operation reduces network traffic because the client does not have to periodically send requests to confirm the server's up-to-date. We implemented a toolbar (Fig. 9) supporting five basic operations. The user can (1) perform a hybrid search, (2) login and logout for communicating with other users, (3) get the information of the number of users accessing the current page, (4) set the avatar display on or off, and (5) configure his or her properties such as the name and the color of input text, etc.

2) highlighting of important information: As described above, the highlight function can be used to both indicate important information and draw the attention of other users. The user can highlight the target text by, for example, left-clicking, dragging and right-clicking to select the interested parts of the page (Fig. 10. This highlighting helps subsequent users to identify the important information because it remains on the page, and the hybrid search system can automatically scroll down and present the highlighted text after the page is accessed.

3) ranking by the number of current accesses: When a user gives a query, our system ranks the search results


of 1936 he took a strong moral stance, and comments he made in a subsequent broadcast were widely condemned as uncharitable towards the departed king. In his early ministry Lang served in slum parishes in Leeds and Portsmouth before his appointme n 1901 as suffragan Bishop of Stepney in London. In 1908 Lang was nominated Archbishop of York, despite his relatively junior status as a suffragan rather tha diocesan bishop. He entered the House of Lords as a Lord Spiritual and caused consternation in

Pakis
peopl
215 o
deadl
- Busin
Gayd
convi
scand
illegal
- The C
convi
in Fra
- Jurela
Presic

Fig. 10. Notable information shared by highlighting

related to the query. When no query is given, our system also provides a function for ranking all the pages in the access log. An example of this ranking is shown in Fig. 11. The URL at the top has the most users currently accessing it. The pages with higher ranks are considered popular pages because they are more frequently accessed by users. By accessing such ranking page, a user can find the popular topics and connect to more people.

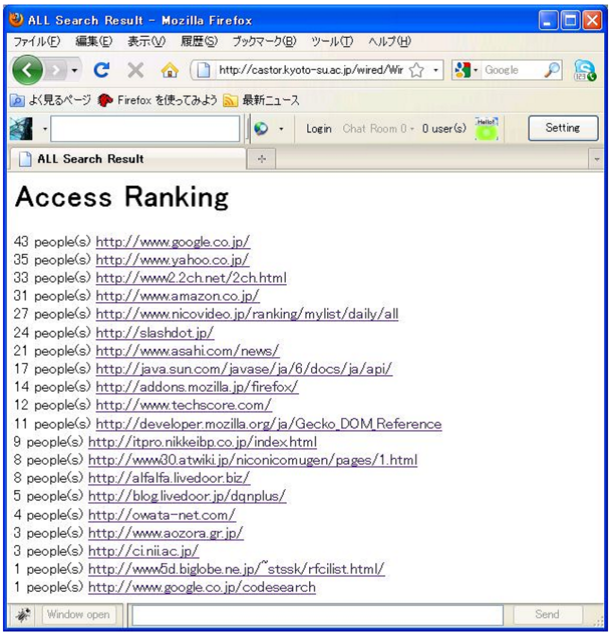
VII. EVALUATION

We compared the ranking given by a conventional search engine and the one given by our system. During the period from April 1 2010 to April 30 2010, there were 3 million access users and 16 million queries.

We selected eight popular queries from the stored data and calculated spearman correlation of the top 20 ranking between the conventional search engine and our system. The result is shown in Table III. A value close to 1 means a high correlation, and 0 means a low correlation. As we can, the relatively long term topics have high correlation and the temporarily hot topics have low correlation.

Figure 12 shows the ranking given by the conventional search engine and the ranking by our system are different.

For viewability, we only plot the top 10 ranking for three queries. For example, for the query “Sportsman A”, the page with rank 7 given by the conventional search engine is ranked as rank 2 by our system. This indicates that ranking pages based on the information of access users is necessary and useful.


ALL Search Result - Mozilla Firefox
ファイル(F) 編集(E) 表示(V) 履歴(S) ブックマーク(B) ツール(T) ヘルプ(H)
http://castor.kyoto-su.ac.jp/wired/Wir
Google
よく見るページ Firefox を使ってみよう 最新ニュース
Login Chat Room 0 0 user(s)
Setting
ALL Search Result
Window open
Send

Access Ranking

43 people(s) http://www.google.co.jp/
35 people(s) http://www.yahoo.co.jp/
33 people(s) http://www2.2ch.net/2ch.html
31 people(s) http://www.amazon.co.jp/
27 people(s) http://www.nicovideo.jp/ranking/mylist/daily/all
24 people(s) http://slashdot.jp/
21 people(s) http://www.asahi.com/news/
17 people(s) http://java.sun.com/javase/ja/6/docs/ja/api/
14 people(s) http://addons.mozilla.jp/firefox/
12 people(s) http://www.techscore.com/
11 people(s) http://developer.mozilla.org/ja/Gecko_DOM_Reference
9 people(s) http://itpro.nikkeibp.co.jp/index.html
8 people(s) http://www30.atwiki.jp/niconicomugen/pages/1.html
8 people(s) http://alfalfa.livedoor.biz/
5 people(s) http://blog.livedoor.jp/dqnplus/
4 people(s) http://owata-net.com/
3 people(s) http://www.aozora.gr.jp/
3 people(s) http://ci.nii.ac.jp/
1 people(s) http://www5d.biglobe.ne.jp/~stssk/rfcilist.html/
1 people(s) http://www.google.co.jp/codesearch

Fig. 11. Example of ranking based on current accesses

TABLE III
SPEARMAN CORRELATION

query	ρ
Sportsman A	0.04
Idol B	0.2
Ipad	0.28
Singer C	0.54
Idol D	0.55
Song E	0.62
Idol F	0.77
Bean Diet	0.93

VIII. CONCLUSION

We developed a hybrid search system with social communication service. The hyperlinks are used to obtain search results, and the social links are used to re-rank the results. A user can communicate with other users currently accessing the same page, and possibly find an expert with appropriate knowledge. Since the hyperlinks on each accessed page are annotated with the number of users accessing them, a user can connect to even more users by accessing deeper pages. Evaluation of our ranking method using eight queries showed its usefulness.

We intend to further evaluate users’ quality by analyzing the communication logs and the access history. A user study will be done for evaluating the performance of the communication service. The capacity of the system will also be expanded.

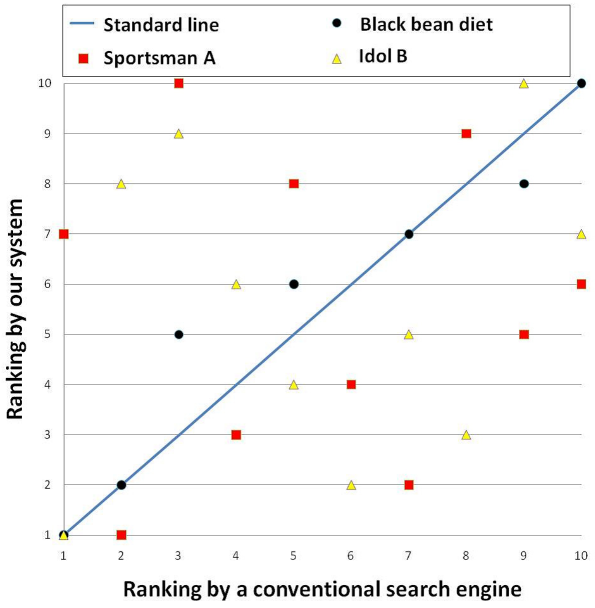


Fig. 12. Ranking by a conventional search engine vs. ranking by our system

ACKNOWLEDGMENT

This work was supported in part by the National Institute of Information and Communications Technology, Japan, and by SCOPE (Ministry of International Affairs and Communications, Japan).

REFERENCES

[1] E. Y. Chang. Confucius and “ Its ” Intelligent Disciples. In *Proc. CIKM 2009*, pp.3-3, 2009.

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

[4] J. Pickens, G. Golovchinsky, C. Shah, P. Qvarfordt, and M. Back. Algorithmic Mediation for Collaborative Exploratory Search. In *Proc. SIGIR 2008*, pp.315-322, 2008.

[5] M. R. Morris and E. Horvitz. SearchTogether: An Interface for Collaborative Web Search. In *Proc. UIST 2007*, pp.3-12, 2007.

[6] M. B. Twidale, D. M. Nichols and C. D. Paice. Browsing is A Collaborative Process. *Information Processing and Management, 33(6)*, pp. 761-783, 1997.

[7] A. Goel and K. Munagala. Hybrid Keyword Search Auctions. In *Proc. WWW 2009*, pp.221-230, 2009.

[8] M. Richardson, E. Dominowska, and R. Ragno. Predicting Clicks: Estimating the Click-through Rate for New Ads. In *Proc. WWW 2007*, pp.521-530, 2007.

[9] A. Jatowt, Y. Kawai, and K. Tanaka. What Can History Tell Us? Towards Different Models of Interaction with Document Histories. In *Proc. Hypertext 2008*, pp.5-14, 2008.

[10] F. B. Viegas, and J. S. Donath. Chat Circles. In *Proc. CHI 1999*, pp.9-16, 1999.

[11] S. Greenberg, and M. Rounding. The Notification Collage: Posting Information to Public and Personal Displays. In *Proc. CHI 2001*, pp.515-521, 2001.

[12] A. Ranganathan, R. H. Campbell, A. Ravi, and A. Mahajan. ConChat: A Context-Aware Chat Program. *IEEE Pervasive Computing*, Vol. 1, No. 3, pp.51-57, 2002.

[13] S. Brin, and L. Page. The Anatomy of A Large-Scale Hypertextual Web Search Engine. *Computer Networks 30(1-7)*, pp. 107-117, 1998.

[14] Hyper Web Search System with Social Communication Service: http://klab.kyoto-su.ac.jp/ mito/index.html