SERPWatcher: A Sophisticated Mining Tool utilizing Search Engine Results Pages (SERPs) for Social Change Discovery

Yoshifumi Masunaga, Kazunari Ito, Yoichi Miyama
School of Social Informatics
Aoyama Gakuin University
Kanagawa 252-5258, Japan
{masunaga, kaz}@si.aoyama.ac.jp

Naoko Oyama, Chiem Watanabe, Kaoru Tachi
Institute of Gender Studies, Dept. of Computer Science
Ochanomizu University
Tokyo 112-8610, Japan
{naoko.oyama, chiemi.watanabe, kaoru.tachi}@ocha.ac.jp

Abstract—As various individuals and organizations disseminate information on their Web pages, real-world social events and changes are considered to be reflected in Web trends. The billions of Web pages that now exist are retrieved by Web search engines which accept keywords and return a search engine results page (SERP). Since the SERP itself and the ranking order change with time reflecting the changes in society, it might be possible to accurately follow the movement of society by mining SERPs. This paper reports the design and implementation of a SERP mining tool named SERPWatcher. It provides sophisticated interfaces and functions for SERP miners in the field of social sciences to discover social changes. It could be a novel social survey method in that it totally differs from the traditional methods such as questionnaires and interviews. A research prototype of SERPWatcher is currently under operation, and its validation shows that it has a fail-safe nature in the sense that social changes are mirrored on the changes of the SERP ranking order. Also, the testers conducting gender studies have been expressing positive opinions on the use of SERPWatcher as a novel research methodology in the field of sociology.

Keywords—SERPWatcher, search engine results page (SERP), SERP mining, social change discovery, social survey method, Web, Web mining

I. INTRODUCTION

A. Historical Background

The Web has seen rapid and widespread acceptance primarily because of the underlying technology of modern society. The most attractive feature of the Web is that it is a hypertext system even without repeating the word of Berners-Lee [1]. Web pages written in HTML are linked to each other in a nonlinear form, which enables us to navigate the Web. Web pages may describe various subjects of the real world for information dissemination purposes. The Web, as a collection of billions of linked Web pages, constitutes a significant information space called cyberspace. Real-world events and societal changes are sure to be mirrored on the Web. Therefore, by mining or searching the Web, it could be possible not only to find something new but also to interpret the complex movement of society. Historically, such research has worked, particularly in the early days of Web mining. If we look at Web structure mining, we can enumerate a variety of interesting research results such as the development of Web community extraction tools on the basis of HITS [2]. For instance, Companion- has also been successfully used as a device for analyzing societal behavior [5].

Although Web (structure) mining could be a novel research methodology for the social sciences, the present authors have been paying special attention to “SERP mining,” potentially another novel research methodology for sociological studies. In brief, as is widely known, the billions of Web pages that now exist on the Web are retrieved by Web search engines according to user-specified search keywords and listed on a search engine results page (SERP). Therefore, it may be possible to accurately follow the movement of the society by mining SERPs. As described in the related works (Section I.D), considerable research and development concerning SERPs has been carried out from various viewpoints. This paper reports a SERP mining tool named SERPWatcher, which is a new social survey method and is different from other Web structure mining tools such as Companion- as well as the traditional methods such as questionnaires and interviews.

B. Motivation of This Research

To better elucidate the potential utility of SERP mining as a novel social survey method, let us introduce a lucid example: We began our research at the beginning of fiscal year 2007. We paid special attention to the very unstable political situation in Japan at that time and specified the search keyword “gender-equal” on Google. The collection started on August 5, 2007, and SERPs were collected each week. When we analyzed the collected SERPs, we found a very interesting phenomena regarding the change in the ranking order of the Web pages.
related to gender bashing and the change in the ranking order of the Web pages related to gender protection sects, which splendidly corresponded to the change in prime minister in Japan at that time; that is, Shinzo Abe resigned from the position of prime minister on September 12, 2007, and Yasuo Fukuda took charge as the new prime minister on September 26, 2007. Here, it might not be easy for the readers to understand how the search keyword “gender-equal” related to the change in prime ministers. Specialists in gender interdiscipline know that Shinzo Abe is a strong supporter of gender bashing and Yasuo Fukuda is very influential in the gender protection sects. This is the reason why the ranking order of the Web pages related to gender bashing fell and why the ranking order of the Web pages related to the gender protection sects rose considerably, corresponding to the change in prime ministers. Figure 1 shows a part of the ranking order change of Google’s Web pages at that time: After Abe resigned on September 12, it was observed that the ranking order of a Web page related to gender protection sect that held the 19th place on the base point (August 5, 2007) rose to the 11th place around September 16 when Fukuda became powerful as the prime minister for the next term. On the other hand, it was observed that the ranking order of the Web page related to bashing sectors that held the 14th place on the base point fell to the 25th place, accordingly.

(3) The system will notify the scientists when it detects certain changes in the ranking order of Web pages in SERPs.

In this sense, it might be interesting to say that a search keyword is regarded as a “probe” (medical examination needle) that detects a change in society. Therefore, it may be more accurate if we pay attention to SERPs collected by several search engines because the crawled Web pages differ from one search engine to another because of the different collection algorithms of the search engines.

To offer such a novel research environment based on Web technology for social scientists, SERPWatcher is available at our institute. The set of SERPs that SERPWatcher collected is called a “SERP Archive,” and it constitutes a “multidimensional database.” More precisely, the SERP Archive composes a 4-dimensional "cube" whose axes are the search keyword, the search engine, the collection date, and the Web page, and whose “cell” takes the value of the ranking order of the Web page. For the scientists to read the social movement from the SERP Archive easily, the SERPWatcher collects the title, snippet (explanation of the Web page of the search keyword), the search engine, the collection date, and the Web page. This is the reason why the ranking order of the Web pages related to gender protection sects, which are collected on the same day the SERP is collected, are also archived. Obviously, this is for helping the researchers to check the “news titles” on the day they are paying attention to.

D. Related Works

Considerable research and development concerning SERP has been carried out from various viewpoints. First, we should state the research of Brin and Page [6] that proposed PageRank. PageRank ordering, which is completely different from the traditional content-based approach, is taken as a typical example of the wisdom of the crowd [7]. There are many Web sites offering a Web application that returns the ranking order of the Web page on a SERP by specifying a Web page URL, the search keyword, and the search engine. Needless to say, organizations are very anxious about knowing how high their Web pages rank on a SERP with respect to the intended search keywords. The SEO technology and the anti-SEO technology have been developed for this purpose. SERP Checker [8], SERP Rank [9], Search Engine Position Checker [10], and many other similar systems have been developed for such a purpose. The SERP Watcher differs from those systems in that it aims to provide a novel social survey method in the field of social sciences.

Since the ranking order change of a SERP is able to detect the scent of a social change, it is desirable that the ranking order of a Web page that a SERP shows is close to its significance order in the real world. Therefore, the related work should also include research and development for more accurate ranking by search engines. The research for accuracy improvement is divided roughly into two categories: the development of meta-search engines [11], and the development
of plug-ins to increase the accuracy of search engines [12]. Meta-search engines try to bring up SERPs of two or more search engines into one SERP in order to increase retrieval accuracy.

Recently, Google began to offer Google Trends [13] that presents the history of the frequency of the usage of a search keyword. By this Web application, we can know the degree of interest of the society for that search keyword. Although the search keyword that shows “trends” is predefined by Google, it could be an auxiliary tool for analyzing the social change in addition to the Google News and Yahoo! News items that are archived in SERPWatcher. Further, with relation to the alert function of SERPWatcher, Google Alerts [14] provides e-mail updates of the latest relevant Google results (Web, news, etc.) on the basis of the user’s choice of query or topic. Note that SERPWatcher and Google Alerts are different in that the former sends alerts to notify the detection of significant ranking order changes in the latest SERP when it is compared with the ranking order of the adjacent SERP with respect to a keyword specified by the user (see section III.D) while the latter sends alerts only if new content matches the user-selected search terms.

The rest of the paper is organized as follows: The design principles of SERPWatcher are described in Section II, and the implementation of SERPWatcher is described in Section III. Section IV presents the conclusions and future work.

II. DESIGN PRINCIPLES OF SERPWATCHER

A. An Overview of SERPWatcher Design

SERPWatcher is a tool that observes the change in the ranking order of Web pages on a SERP, hereafter referred to as the “SERP ranking order,” in a comprehensive manner and provides sophisticated interfaces and functions to expected users, i.e., scientists from the field of social sciences. As described in the previous sections, the SERP ranking order changes adequately with the passage of time in response to the changes in society. SERPWatcher constructs the SERP Archive by regularly collecting the Web pages and their ranking order on a SERP when a certain search keyword is specified to a search engine. It is expected that a user can obtain many hints necessary to read and solve the events and the changes in the society by multilaterally analyzing the SERP Archive. Therefore, SERPWatcher should have the following functions: user specification, search engine specification, search keyword specification, (periodical) SERP collection, construction of the SERP Archive, multidimensional analysis of the SERP Archive, and related data collection such as backlinks of a Web page, and the collection of Google News and Yahoo! News at every SERP collection date.

The followings of this section describe the principles of how to collect SERPs, how to organize a SERP archive, and how to provide a graphical user interface for sophisticated analysis of the SERP Archive.

B. SERP Collection

1) Collection of Multiple SERPs: SERPWatcher collects the SERPs of several search engines registered beforehand so that it can construct the SERP Archive when the search keyword and the SERP collection date cycle (default: one week) are specified by a user. In the SERPWatcher current version, SERPs are collected up to the ranking order of 500 for each search engine. The collection method of a SERP and the cleaning of collected data are described in the following sections. The reason why SERPWatcher collects the SERPs of several search engines was mentioned in section I.C. However, we need to discuss which search sites (search engines) should be chosen as the target sites of the SERP collection. As is known, not all search sites have their own search robots. The latest (i.e., December 2008) correlation of the robot-type search engines available in Japan is reported by [15]. Although the correlation changes drastically with time because of the absorption and amalgamation of search sites in this business, we decided on seven search sites, i.e., Google, Yahoo! Search, Bing (formerly Live Search and MSN Search), Baidu, goo (has a tie-up with Google), infoseek.rakuten (has a tie-up with Yahoo! Search), and excite (has a tie-up with Yahoo! Search) as the search sites for the SERP Archive construction. Notice that Google, Yahoo! Search, Bing, and Baidu have their own search engines, while goo, infoseek.rakuten, and excite do not have their own search engines. In principle, it is not necessary to nominate a tie-up search site as a search site of SERPWatcher because it may return the same SERP as the original does. However, it is revealed that this is not true, as reported in our previous research on the trustworthiness of search sites [16]: We discovered a case where Google returned a different SERP to its tie-up sites in an arbitrary manner.

2) SERP Collection Method: There are two methods of acquiring SERPs: one uses the search engine API (Application Programming Interface), and the other uses the WUI (Web User Interface). As is pointed out in [17], the API results are not older, but they are probably smaller than the WUI results for Google and Yahoo!. This means that the recall ratio is probably worth recording in the API case. Hence, we adopted WUI. Concerning the number of search results to be retrieved from a search engine at a time, we chose 10 because of the following reasons: In Google, this number is arbitrarily set from 1 to 100 by setting the “num” attribute value of the request URL. However, in Google’s SERP, when more than one Web page with the same domain is retrieved by a single request, the Web pages with a lower ranking order are placed together with the highest ranking order. For instance, when ten Web pages are acquired, and the Web pages in the 2nd place and the 9th place have the same domain, the page in the 9th place is shown under the Web page in the 2nd place. (This page is ranked 3rd in our SERP collection program.) Therefore, when we set a large value for the number of search results to be retrieved, a Web page with originally a low rank tends to be ordered higher. In order to avoid this negative effect, small numbers are preferable. However, when a fairly large number of user requests are sent to Google by a single user, Google tends to stop returning the query results for a certain period of time to turn down the requests by robots. This is another trade-
off problem to be solved, and we set the number of search results to be retrieved as “10” although SERPWatcher needs to issue queries 50 times to obtain 500 results with respect to a single search keyword. Under the above-mentioned scheme, SERPs are collected each week from day one to now. A problem is that the total number of keywords registered by users is limited up to around 35 in the current implementation because of the robot exclusion. Collection statistics are as follows: It takes approximately 1 to 1.5 second to acquire 10 Web pages result of a single query, and its data size is approximately 30kB.

3) SERP Data Cleaning: Data that SERPWatcher collected compose the SERP Archive. Needless to say, the Web page data stored in the SERP Archive should be noise-less and highly credible. We inspected the SERP Archive and found the following problems: (a) Only the title of a Web page is acquired when the Web page is an image retrieval result or an animation retrieval result. (URLs and snippets cannot be acquired directly.) (b) There are overlapping Web pages in different SERPs with respect to the same search engine, the same search keyword, and the same collection date, except that the requested ranking order ranges are different. The reason for problem (a) is clear; the problem is due to the discrepancy between the strategy of our SERP collection algorithm and the Google page rank algorithm that displays only the title as an anchor to the real URL of an image retrieval result or an animation retrieval result. Problem (b) is investigated as follows: For example, when we examined the SERP Archive data collected on June 14, 2009, by selecting “Google” as the search site and “gender” as the search keyword, seven consecutive Web pages overlapped between the set of Web pages from the 481st place to the 490th place, and a different set of Web pages from the 491st place to the 500th place. Such phenomena were seldom observed in the subordinate position although they were not frequently seen in the high ranking order. To understand such phenomena, we have observed the ranking order changes of Web pages in the subordinate position and found that they were changing within a few minutes. However, we cannot know the real reason accurately, and this is an open problem. In the current implementation, whenever such duplication is found, it is removed manually and the ranking order of Web pages is re-numbered.

C. SERP Archive: Multidimensional Organization of SERP Collection

The SERP Archive that accumulates SERPs regularly collected on the basis of various search engines and the search keywords has two database features (Details of this database design are described in section III.B):

(a) The SERP Archive is not updated.
(b) The SERP Archive composes a multidimensional database.

Therefore, the SERP Archive is suitable for OLAP (Online Analytical Processing) [18], i.e., a multidimensional analysis of the SERP Archive. As was already mentioned in section I.C, the SERP ranking order is characterized by a function of the 4th order with the following variables: search keyword, search engine, collection date, and Web page URL. Further, note that a search keyword is specified by a keyword specifier (user), and SERP includes all the information about a retrieved Web page, i.e., ranking order, title, snippet, URL, backlinks, and the total number of backlinks of a retrieved Web page. Therefore, the SERP Archive constitutes a 4-dimensional cube shown in Figure 2, where “D_” represents dimension.

D. Multidimensional Analysis of SERP Archive

When the SERPWatcher current version was developed, we examined methods of analyzing the SERP Archive as a multidimensional database with the social scientists who were the intended users. As a result, it turned out that they wanted to specify a search keyword first (in other words, they want to fix the search keyword first), and then carry out the following three types of search carefully:

(a) {web page URL, collection date} → {ranking order of the web page}
(b) {web page URL, search engine} → {ranking order of the web page}
(c) {search engine, collection date} → {ranking order of the web page}

In other words, when a search keyword is fixed, it turns out obviously that the SERP Archive becomes a three-dimensional data cube with the Web page axis, the search engine axis, and the collection date axis. Therefore, the above-mentioned three demands (a), (b), and (c) correspond to the typical cubic operations named (a) search engine fixation view, (b) collection date fixation view, and (c) Web page fixation view, respectively. Figure 3 illustrates these demands. In the current version implementation of SERPWatcher, such peculiar slice and dice operations of OLAP are implemented by applying the relational database operation directly to the relations that compose the database representation of the SERP Archive.
E. Color Shading Algorithm for Ranking Order Changes

In this section the display of the above-mentioned analysis result is described. Because the principle of displaying the analysis result is the same for the three above-mentioned cases, case (a) is taken as the typical example of explanation: In order for a user to perform an analysis freely through the search engine fixation view, if a collection date is specified as the reference date by a user, the display order of the Web pages at other dates are re-arranged according to the ranking order of the reference date. In other words, the Web pages on the Web page axis in the search engine fixation view (case (a)) are ranked according to the ranking order of the Web pages of the reference date.

To tell the user the difference in the ranking order plainly, the cell of the search engine fixation view is colored according to the following idea: Roughly speaking, a Web page of a SERP collected at the date that is different from the reference date is colored “blue” when the ranking order is lower than the ranking order of that of the reference date and is colored “red” when the ranking order is higher. Moreover, the color becomes deeper according to the ranking order difference. Therefore, it becomes easy for a user to distinguish between a Web page whose rank increases gradually, the page whose rank has gradually fallen, and the page whose rank increases and decreases rapidly. The color shading algorithm is shown below:

The 8-bit RBG color model is used for color shading: For red color, the shading is determined by the function \((R, G, B) = (255, x, x)\) were \(x\) represents the brightness of color and its integer value ranges from 0 to 255. For blue color, the shading function is defined as \((R, G, B) = (x, 255, x)\). The value of the variable \(x\) is determined by a function shown in Figure 4: If the absolute value of the difference between the ranking order of a Web page at the date of “gaze” and that of the reference date is 0 then \(x = 255\), if 1 then 250, and so on up to 50, and if the difference exceeds 50 then \(x = 0\). By the way, when the ranking order difference is 0, the \((R, G, B) = (255, 255, 255)\), which represents white. On the other hand, if the order at the gaze date is higher than that of the reference date by more than 50, then \((R, G, B) = (255, 0, 0)\), which represents pure red. If it is lower than or equal to 50, then \((R, G, B) = (0, 0, 255)\), which represents pure blue. We asked the testers from the field of sociology to evaluate this shading algorithm. The result was that it excelled in the visibility, and it could be a strong user-friendly analytical tool for the researchers in this research field.

This is because of the observation that the changing places of the display order tend to be related to interesting changes in the society with respect to the social sciences. A typical screenshot of the search engine fixation view (case (a)) following to this color shading algorithm is shown in section III.C (Figure 7).

F. Collection of Auxiliary Data for SERP Analysis

The news from news sites on the SERP collection date are also collected to help users to perform a social analysis of the day when the SERP was collected. In the current implementation, the top page of Google News (http://news.google.co.jp) and Yahoo! News (http://headlines.yahoo.co.jp/hl) are archived. By accessing the news archive, users can confirm the titles, i.e. headlines of the top news and the news on society, international, business, politics, entertainment, sports, and technology.

G. Alert Function of SERPWatcher

SERPWatcher is designed to have a function to notify the researchers of the changes in the SERP ranking order immediately after they are detected. This is called the “Alert function.” The alert function informs the users of the change if it detects any one of the following three events when the latest SERP is collected:

(a) The Web page that appeared in the adjacent SERP disappeared from the latest SERP. (Recall that top 500 Web pages are collected by SERPWatcher at each collection date.)
(b) The Web page that did not exist in the adjacent SERP appeared in the latest SERP.
(c) The order of the Web pages that exist in the adjacent SERP has changed greatly in the latest SERP.

In each case, SERPWatcher sends an e-mail to the users to notify them of the change. The change information includes the following: the search keyword, the adjacent SERP collection date, the latest SERP collection date, the search site, the ranking order of the Web page in the adjacent SERP, and the ranking order of the Web page in the latest SERP. In case (c), some exploratory experiments have been carried out to determine how much the page order has to change in order to trigger a notification that would be of value to sociological researchers, the typical users of this system. As a result, it was decided that the notification would be sent when the ranking order changed by at least 40%. If we set a small value for the percentage, obviously the alert becomes meaningless. This is a
typical trade-off problem examined by system developers and users. An actual alert screenshot is shown in section III.C (Figure 8).

III. SERPWATCHER PROTOTYPING

A. System Configuration of SERPWatcher

As clarified in the foregoing sections, SERPWatcher should have the following functions:

(a) Periodic collection of SERPs
(b) Construction of SERP Archive
(c) Multidimensional analysis and display of SERP Archive
(d) Alerts
(e) Collection and display of news (Google News and Yahoo! News)

Figure 5 shows the system configuration of SERPWatcher. The SERP collection program, the backlink collection program, and the news collection program operate regularly as batch programs. Collected SERPs are recorded in a file whose records are ordered by search engine, collection date, and search keyword. Another batch program with this file as an input is run to store the file into a database. A relational database management system (MySQL, in our current implementation) is used for storing the collected SERPs. Details of the database design are described in the next section.

B. Construction of SERP Archive

This section describes how to construct a SERP archive as a multidimensional database. As described before, SERPWatcher collects SERP (up to the 500th position) with respect to the search keywords specified by users of all search engines at constant intervals (default: one week). Moreover, it uses a relational database system to manage the collected data uniformly. The data that should be managed in this database are as follows:

(a) User
(b) Search keyword
(c) Search engine
(d) SERP collection date
(e) SERP data

Figure 6 shows the database schema. Note that the news items from Google News and Yahoo! News are not included in this database because they are stored in a separate file.

C. Implementation of Three Views of SERPWatcher

The current version of SERPWatcher is developed by using the client server approach. Because the function required by clients is simply a Web browser such as Internet Explorer or Firefox, SERPWatcher is realized as a remote system with a thin client. The operating system of the server is Red Hat Enterprise Linux 5.1 (x86/x86_64). The development programming language is Ruby 1.8.5 (2006-08-25) [i386-linux] and Ruby on Rails 2.1. MySQL 5.0.45 is used as a relational database management system.

Figure 7 shows a snapshot of the SERPWatcher. This screenshot represents a “Search engine fixation view” (case (a)) described in section 2.D. More precisely, this represents the changes in the SERP ranking order where user selected the following specifications: “gender-equal” as a keyword, “Google” as a search site, “1st to 40th” as the ranking order to be displayed, and “from 2009/02/07 to 2009/06/14” as the display period. (By the way, day one of the SERP collection is July 29, 2007, and the data have been collected each week from that date.) The user can inspect any Web page to the 500th subordinate position place by scrolling the screen. Similarly, by scrolling the screen, the user can inspect the ranking order by going back to an arbitrary collection date. To improve the visibility, the ranking order cells of the Web pages are colored according to the color shading algorithm described in section II.E. In this example, June 14, 2009, is specified as the “reference” date. The coloring of the cells of this two-dimensional table obeys the color shading algorithm defined earlier. By selecting a value from the “search engine” pull-down menu, users can select one of the seven search sites: Google, Yahoo! Search, Bing, Baidu, goo, infoseek.rakuten, and excite. The search keyword can be selected from the pull-down menu in the “search keyword” window. (It is possible to update it at any time although only 37 search keywords are registered currently.) The users can specify a range of interested ranking order and an interesting range of the
collection dates by specifying their values in the “ranking order” window and the “date” window, respectively. The user attempts to capture the change in society by reading and solving the ranking order changes of Web pages displayed in this view.

Because of space limitations, the screenshots for case (b), i.e., the collection date fixation view, and case (c), i.e., the Web page fixation view are shown together in Figure 8. Note that three views, i.e., search engine fixation view, collection date fixation view, and Web page fixation view together provide a very sophisticated analytical environment to the users to discover social events and changes: Because a user can move back and forth among three views by clicking the buttons arranged on the vertical and horizontal axes of each view with additional information access such as viewing a snapshot of the Web page captured at the SERP collection date with respect to any combination of a search engine and a search keyword as well as viewing Google News and Yahoo! News on the collection date, the user can discover a social event or a social change happened around that date.

E. Validation of SERPWatcher Design

The purpose of SERPWatcher development is to provide a sophisticated user interface and a set of strong functions to scientists from the field of social sciences so that they can discover social events and changes by analyzing the movement of the SERP ranking order. In order to validate the design principles of SERPWatcher, we have conducted an experiment under the collaboration with scientists in sociology, particularly in gender studies: We first registered a certain number of search keywords (37 keywords in this experiment) related to social science matters such as Japanese politicians’ names or current topics on gender issues, and then watched the movement of the SERP ranking order concerning the keywords for more than 1.5 years. The result was that whenever something noteworthy happened in the real world with respect
to a keyword, it caused a SERP ranking order change with respect to that keyword. A typical example of this causality was seen in the Search Engine Fixation View with respect to Google as a search engine and Ichiro Ozawa (the present Secretary-General of the Democratic Party of Japan) as a search keyword. That is, whenever his political finance scandal is reported, the SERP ranking order of related Web pages changed significantly.

Formally, this causality means that SERPWatcher guarantees a user’s investigation will be conducted in fail-safe mode. Due to contraposition, if any SERP order change is not observed, then there is no event and change in the real world. However, the inverse is not always true. Therefore, if some SERP order change is observed with respect to a search engine and a search keyword, then there are two possibilities: (a) something happened in the real world, or (b) nothing happened in the real world with respect to the search engine and the keyword. Case (b) might happen if the search engine’s SERP strategy changes for some reason or if something happens in the real world which does not relate to the search keyword directly but relate to it indirectly. This kind of indirection could be illustrated using a Japanese proverb: “if the wind blows, a cooper makes a profit.” (Bliss often falls into the hands of an unexpected person.) This fail-safe nature of SERPWatcher is similar to the false dismissal error handling of information retrieval. The investigation of the fail-safe “rate” could be an interesting research topic in the future.

IV. CONCLUSIONS AND FUTURE WORK

This paper reported the design and implementation of SERPWatcher, which is a SERP mining tool. This development was based on our previous research on Web mining [4, 16], which gave us the insight that “SERP mining” could be a novel social study method that is completely different from the traditional social study methods such as questionnaires and interviews. Although SERPWatcher is a research prototype, the testers from the field of gender studies have expressed a very positive impression through the exploratory experimental use of SERPWatcher for more than 18 months because they have observed many changes in the SERP ranking order, which were closely related to social changes.

Some of the interesting research topics have been suggested as future work in the forgoing sections which might include the followings: Introduction of non-linear function to the color shading algorithm (section II.E) to make the shading fit more likely to the human sense of importance of social changes. Another future work is the enhancement of the alert functions which are currently under-investigation at our research group. Needless to say, the improvement of the collection method of SERPs to increase credibility is also an important future work. In addition, because we intend SERPWatcher to be public for every social scientist, a user management function would be implemented.

ACKNOWLEDGMENTS

The authors are thankful to Mr. Sumito Takeuchi of Aoyama Gakuin University for his comments and suggestions. This research is partly supported by a Grants-in-Aid for Scientific Research of MEXT of Japan in the Category of Scientific Research (B) (Grant number 19300025) on “Construction of an Integrated Web Mining Environment as a New Methodology for Social Science Studies” (FY 2007-FY 2009).

REFERENCES