Trustworthiness Analysis of Web Search Results

Satoshi Nakamura, Shinji Konishi, Adam Jatowt, Hiroaki Ohshima, Hiroyuki Kondo, Taro Tezuka, Satoshi Oyama, Katsumi Tanaka

Department of Social Informatics, Graduate School of Informatics, Kyoto University
Yoshida-Honmachi, Sakyo, Kyoto 606-8501, Japan
{nakamura, konishi, adam, ohshima, kondo, tezuka, oyama, tanaka}@dl.kuis.kyoto-u.ac.jp

Abstract. Increased usage of Web search engines in our daily lives means that the trustworthiness of searched results has become crucial. User studies on the usage of search engines and analysis of the factors used to determine trust that users have in search results are described in this paper. Based on the analysis, we developed a system to help users determine the trustworthiness of Web search results by computing and showing each returned page's topic majority, topic coverage, locality of supporting pages (i.e., pages linked to each search result) and other information. The measures proposed in the paper can be applied to the search of Web-based libraries or can be useful in the usage of digital library search systems.

Key words: Web search, trustworthiness, page locality, user study

1 Introduction

Web search engines have become indispensable tools for acquiring information over the Internet. Web search engines accept user queries consisting of a few keywords, retrieve relevant pages available from the Web, and rank the found results by using their own ranking systems. One of the most important problems with such a search process is that the search engine does not indicate the extent to which returned page is trustworthy for the user's request except for computing the rank of the page. That is, conventional search engines do not provide information concerning whether:

1. The significance of the content of each returned page is a majority or minority in the Web.
2. The extent to which each returned page contains typical query topics contained on the Web.
3. The extent to which each returned page is supported uniformly throughout the world.

If this information is displayed to users by search engines, users will be able to determine which page is trustworthy, and which page they should choose from the search results listed.

Fogg *et al.* analyzed factors with which the user determines the trustworthiness of Web pages [1, 2]. This work was performed by analyzing a questionnaire based on the Prominence-Interpretation theory [3] and means that the level of the user's trust depends on *prominence*, the strength of appeal of the page, and the user's *interpretation* of the page. Based on these results, they proposed guidelines to determine the credibility of the information about authors that is displayed on Web sites [4]. Zaihrayeu *et al.* attempted to calculate the trustworthiness of search results [5]. They computed the degree of trustworthiness by classifying search results based on IWTrust evaluation, which can be used to learn feature vectors created by linguistic analysis of browsed pages among search results. Yanbe *et al.* recently developed a new page reranking system using social bookmark information [6]. This system allows users to rerank Web search results based on a returned page's bookmark information. Yamamoto *et al.* also developed a system [7] that helps to determine the trustworthiness of sentences by searching and aggregating related Web pages.

We surveyed the search engine usage of users to understand the context in which users search, and which factors cause the user to trust the search results and to understand the requirements of the search system. In this paper, we describe these user studies and analyze the factors determining the trust that users have in search results. Based on this analytic work, we developed a system to help users determine the trustworthiness of Web search results independently from a conventional search engine's ranking mechanism. We computed and displayed measures including *topic majority*, *topic coverage*, *locality of supporting pages*, and others for each page. The topic majority measures the significance of the content of a returned page. The topic coverage measures how many topics concerned with a search query the returned page contains. The locality of supporting pages for a returned page denotes the localness of distribution of the supporting pages. These measures are useful for users to determine the trustworthiness of searched results.

We also describe our prototype system, in which those measures are displayed together with the standard search results. Additionally, we describe a two-dimensional display interface for the measures.

2 Survey

In this section, we describe the results of the user studies performed in order to gather information regarding the use of search engines. We especially focused on analyzing factors used to determine the trustworthiness of the search results by users. The objectives of this survey were to:

1. investigate the frequency of Web searches by users
2. determine the circumstances in which users search the Web
3. understand the motivation of users for searching in the Web, i.e. why users do search in general
4. analyze how many results do users check, i.e. what is the lowest ranked item that users view before they decide to modify the query and search again

5. estimate the number of times users access search results before they decide to modify query terms
6. investigate whether users are aware of the underlying mechanisms by which search engines determine ranks of pages
7. analyze how much users trust the ranking method used by search engines
8. examine the features of a page used by users to determine the trustworthiness of its contents
9. determine whether users had experienced obtaining information from search engines that was incorrect, obsolete, or untrue
10. analyze what additional information should be provided by search engines, such as URLs and page snippets, to improve search efficiency
11. understand what kind of search engines users would like to use in the future.

We created an online questionnaire consisting of 26 questions that were answered by 1000 Internet users between 25th and 26th December 2006. Users were divided into four categories depending on their age: 20-29, 30-39, 40-49 and 50-59 years old. Each group consisted of 250 respondents; half males and half females. Respondents could choose several answers for some questions. The findings we obtained are discussed below based on the analysis of the survey results:

1. The analysis revealed that 68.7% of users usually use search engines less than 10 times a day. 27.5% of users search more than 11 but less than 30 times a day, and the rest search the Web more than 30 times per day.
2. Users decide to use search engines when they want to research particular information or browse the Web (Figure 1). It is also common for users to search without any particular reason. Two other common situations in which searches were performed are when watching TV and reading e-mails.
3. Users search the Web mostly because they require basic (46%) or detailed (36.8%) information about particular things (Figure 2). Another motivation for searching the Web is to do some comparison (7.4% of respondents selected it as a first reason). Few users chose other reasons for searching the Web. These results suggest that the depth and the coverage of topics in pages relevant to a query can improve the search experience.
4. More than 50% of users analyze only the top five search results. By this we mean that users read titles and snippets or pages that are provided by search engines. Only about 20% of users actually go further than the top five search results. These results indicate the need for creating more efficient search techniques.
5. On average, users visit between one to three pages before they decide to modify the search query or finish search in the Web (78.37% users). Relatively few users analyze more than 11 search results. An interesting result is that more than 20% of respondents do not actually access the pages but only read the returned snippets.
6. Users often believe that the more frequently visited a page is, the higher rank it has in search results. Another common belief is that the relevance of a page to the search query is a major factor when determining its rank in search results. A third belief is that the freshness level considerably influences search

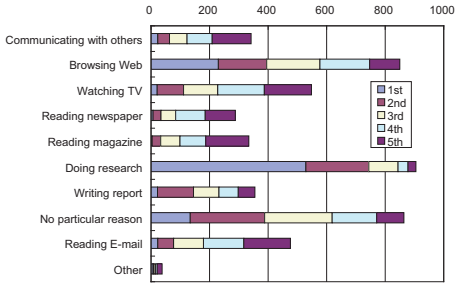

Fig. 1. Situations when users search Web

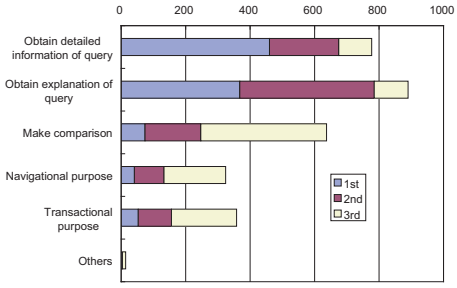

Fig. 2. Reasons for searching Web

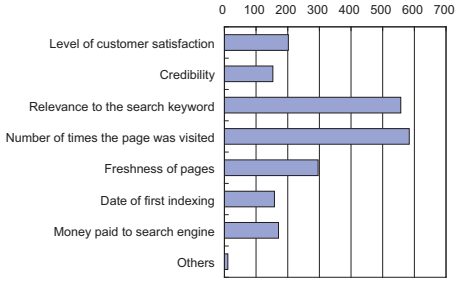

Fig. 3. User beliefs about ranking methods used by search engines

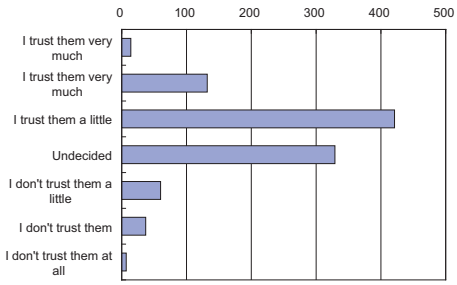

Fig. 4. User trust levels of search results

ranking. An interesting observation is that 17.1% of users actually think that the ranking depends on the amount of money paid by Web authors to search engine companies. Figure 3 shows the popularity of common beliefs among users about the ranking mechanisms used by current search engines.

7. Analysis revealed that 56.7% of respondents generally trust ranking methods used by search engines, and only 10.4% of users do not trust them (Figure 4). This observation indicates the necessity of providing trustworthy search mechanisms as Internet users often assume the correctness of information provided by search engines.
8. Users take into account information about the author or the owner of the page when deciding whether to trust the information. The second trust-invoking characteristic of pages is their relevance to the search query. Respondents tend not to trust pages if they contain spelling errors, grammatical mistakes, or biased information. Users also consider the page creation date as an important factor to determine the trust level of pages. Additionally, users do not trust information that is unique among different sources. The results of this analysis are shown in Figure 5.

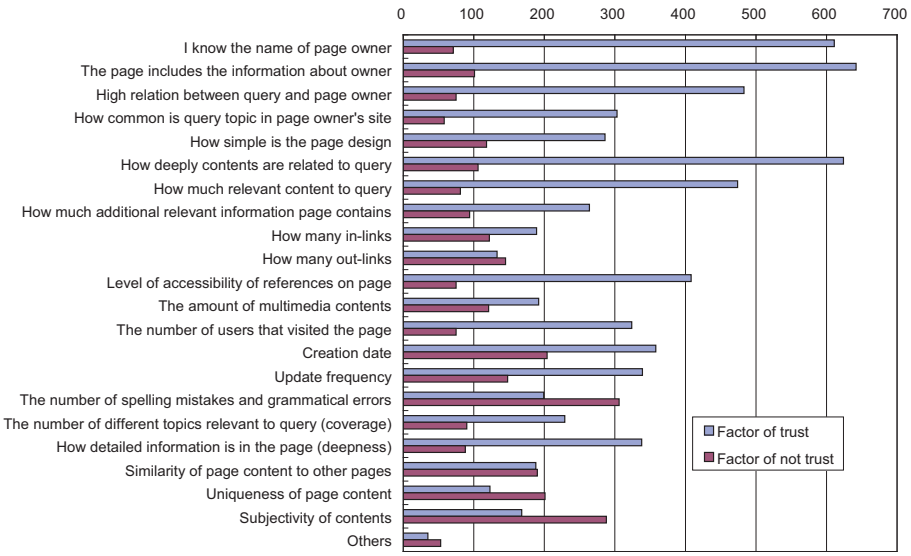

Fig. 5. Characteristics of pages that users trust

9. Some users (12.3%) experienced obtaining information which, after subsequent inspection, turned out to be erroneous, obsolete, or untrue, 3.5% of users accessed adult content, pages containing viruses, or phishing sites, and 5.2% of users detected untrue, obsolete or subjective information when using search engines.
10. Our study indicated that users would like search engines to provide the following types of information: publication date, related words, information about the page author or owner, scoring reflecting trustworthiness of pages, page type, thumbnail image of pages, and third party evaluations.
11. The main search engine characteristics that users wish to use in future are the capability to provide additional information about the results (48.08%) and domain-focused searching (45.7%) (Figure 6). Other common features are: automatic analysis of trust levels of pages, context-aware search and indication of the current popularity levels measured by the number of users visiting pages at query time. Respondents also wished search engines provided summaries of search results or performed result clustering.

3 Prototype System for Determining Trustworthiness of Web Search Results

Based on the survey results described in the preceding section, we designed a prototype system that helps a user to determine the trustworthiness of information

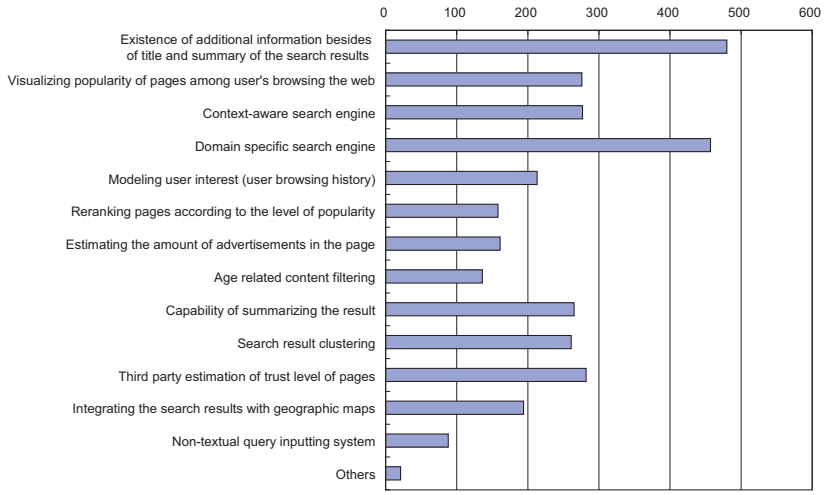

Fig. 6. Search engine characteristics that users would like to use in future

on the Web. The purpose of our system is not to determine the trustworthiness of content by itself but to provide the user with supplementary information to help determine the trustworthiness. We did not largely change the user interface of a current search engine that is familiar to the user but added supplementary information as *add-ins* beside the ranked results returned by the search engine. This enables the system to raise the user's awareness of the trustworthiness of the search results without unnecessarily disturbing the user.

3.1 Information Presented in Prototype System

There are many kinds of information available to assist users to determine the trustworthiness of search results. The major information presented with standard search results are as follows:

Topic majority Nearly half the respondents (43.4%) paid attention to how many similar pages to the search result exist when determining the trustworthiness of the search results. Topic majority is the number of similar pages to the search result that exist in the Web or in the set of pages related to the query. We calculated this by analyzing the number of pages related to the query and the number of pages similar to or containing the same topics.

Topic coverage More than half the respondents (63.2%) tended to trust search results that contain many topics about the search query when searching something they have little or no knowledge about. Topic coverage is how many topics about the query the search result contains. We calculate this number by analyzing the number of topics about the query that the search result contains.

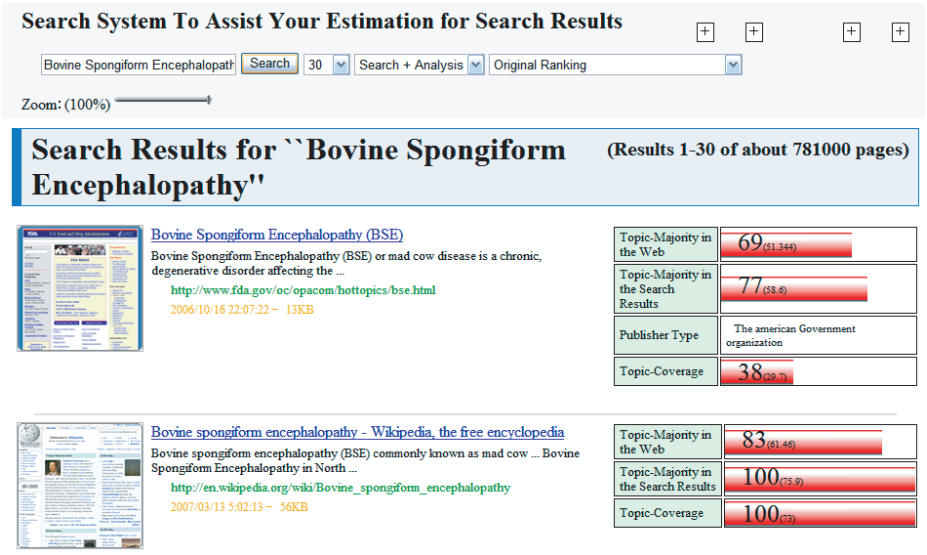

Fig. 7. User interface of prototype system

Locality of link sources Spatial information can also play an important role in estimating the trustworthiness of Web contents. For example, if a certain page is only linked by pages in a limited area, the user can consider that the page has only a limited local support. On the other hand, if the page is linked from pages distributed over a large area (e.g. many countries), the user may think that the page has higher reliability. To support such judgements, our system visualizes the geographic distribution of link sources and illustrates how uniformly they are distributed. In related work, Ma and Tanaka described "localness degrees" as a ranking measure of Web pages [8]. Zhang et al. proposed LocalRank, based on a graph structure of semantic and geographic relationships [9]. Such works are different from ours in that they analyze the content itself, whereas we focus on link sources.

Other information Other types of information exist that are often requested by users and are provided by our prototype system. One is topic details because nearly three quarters of the respondents (72.6%) tended to trust pages describing specific topics about the query. Also, our system provides publisher information (because 85.1% of the respondents paid attention to the publisher's details), the social bookmark number for each returned page (because 38.3% of the respondents paid attention to how many users browsed the search result), and the last-modified date (since 61.4% of the respondents paid attention to when the page was last modified or created.)

3.2 Calculating Topic Majority and Topic Coverage of Pages Using Query Topic Terms

To calculate topic majority and topic coverage, we need to find *representative topics* for user-specified query words.

Wikipedia and Web search results returned by a search engine are used to identify topic terms for a query. First, the system sends the user-specified query terms to Wikipedia to retrieve entries that match the query term. If such an entry is found, the system chooses terms as topic terms whose frequencies in the entry page are larger than a given threshold. If no entries match the query, the system sends the query term to a Web search engine and retrieves top ranked pages. The system then chooses terms whose frequencies in the result pages are higher than the given threshold as topic terms. The system optionally applies statistical tests proposed by Oyama and Tanaka [10] to improve the accuracy of identifying topic terms.

Let q be the user-specified query terms and t be a potential topic term extracted from a Wikipedia page. We compare the values of the following two formulas:

$$p(t \mid q) = \frac{\mathrm{DF}(q \wedge t)}{\mathrm{DF}(q)}$$
$$p(t \mid \mathrm{intitle}(q)) = \frac{\mathrm{DF}(\mathrm{intitle}(q) \wedge t)}{\mathrm{DF}(\mathrm{intitle}(q))} \ ,$$

where DF is the number of results returned by the search engine for a query in the argument, and intitle(x) is the number of pages containing term x in their title. If $p(t \mid \mathrm{intitle}(q))$ is larger than $p(t \mid q)$, we determine that t is a topic term of q.

Let $T = \{t_1, \ldots, t_n\}$ be the set of identified topic terms for q. The system calculates topic majority and topic coverage as follows:

Topic majority(in the Web) This is the number of Web pages that have similar topics to the topic of the page being evaluated. Let P be the set of terms appearing in the page. We calculate Topic majority (in the Web) as

$$\mathrm{TopicMajority(in\ the\ Web)} = \mathrm{DF}(q \wedge s_1 \wedge \ldots \wedge s_m)$$

where $s_i \in T \cap P$ and i is up to three.

The higher this indicator, the more the page includes topics considered significant to the search query. This indicator depends on the search query.

Topic majority(in the search results) This is the number of search results similar to the search results that are to be evaluated. Let p_k be the page to be evaluated, $\boldsymbol{v}(p_k)$ be the feature vector of page p_k, $R(p)$ be the set of the search results for q, $\|\boldsymbol{v}\|$ be the norm for the vector $\boldsymbol{v}$, and θ be a threshold for the similarity between two vectors. We calculated Topic majority (in the

search results) as follows.

$$\text{TopicMajority(in the search results)} = \left| \{p_i \mid p_i \in R(q), \frac{\boldsymbol{v}(p_k) \cdot \boldsymbol{v}(p_i)}{\|\boldsymbol{v}(p_k)\|\|\boldsymbol{v}(p_i)\|} > \theta\} \right|$$

The topic majority (in the search results) indicator can be used to determine whether the search terms are significant in the search results. This indicator depends on the search query and the number and size of the search results.

Topic coverage Topic coverage is the rate of topic terms appearing in the page to be evaluated and is calculated as follows:

$$\text{TopicCoverage} = \frac{|T \cap P|}{|T|} \quad .$$

In the formula, no weight is assigned to topic terms to reflect topics, which are minor in the Web. Considering this indicator and other information, i.e. topic details, users can determine the bias of a page's contents.

3.3 Calculating Locality of Supporting Pages Using Link Structure

As described in the previous section, spatial factors can help the user determine trustworthiness of Web content. We define Locality of Supporting Pages (L) of a Web page as follows.

$$L(p) = \frac{n}{\sum_{i=1}^{n} ln(d(p, p_i) + 1)} \tag{1}$$

In the formula, p and p_i are the coordinates of the target Web page and pages that link to it, respectively. $d(p, p_i)$ indicates the distance between p and p_i. n is the number of pages that link to the target page.

The system obtains the URLs of pages that link to the target page using the "link" operator of a regular search engine. The system then converts these URLs to IP addresses using DNS. Finally, it obtains geographical coordinates corresponding to these IP addresses using GeoLite City by MaxMind [11]. At this moment, our system can only support judgments on the trustworthiness of the Web page. It can not help the user in judging the trustworthiness of pieces of information on it. Providing finer granularity is a part of our future work.

Figure 8 illustrates that the locality of supporting pages is only weakly correlated with the number of links toward the target Web pages. It shows that the locality of supporting pages can provide a ranking different from conventional search engines, since they are basically based on the amount of links coming in. Figure 9 illustrates the system's visual interface. In this example, it shows the spatial distribution of pages that link to the government of South Africa [1]. The

[1] http://www.gov.za

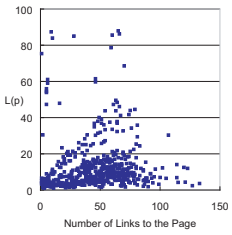

Fig. 8. Locality and Amount of Supporting Pages

Fig. 9. Visual Presentation of Locality Support

locality of supporting pages was $L = 2.427$. This is close to that of Google [2] ($L = 2.939$) and the government of Australia [3] ($L = 2.792$) whereas far from that of a locally targeted page, such as Alachua County Today [4], a local news site in Florida ($L = 42.240$).

3.4 User Interface

We show a screenshot of the prototype system's user interface in Figure 7. This interface has an input field for a search query, the number of results the user wants, and a result order combo box in the upper section of the interface. The results are shown in the lower section of the interface. The search results are displayed in the result section in the order that the user selected using the order combo box. For each search result, the system displays the title, the snippet, the URL, the thumbnail, the date when the page was last updated, and the page size on the left. Additional information the system has analyzed for the search result is displayed on the right. Moreover, a bar is displayed for each item to indicate the relative value, (the max value is changed to 100 and the min value is changed to zero). We also implemented a toggle display function for each additional piece of information.

We implemented a two-dimensional allocation display mode as shown in Figure 10 (the horizontal axis is the topic majority (in search results) and the vertical axis is the topic coverage). This mode will enable users to better understand the relationship among search results.

3.5 Evaluation

To evaluate processing time, we used snippets in analysis and obtained site information, topic majority, topic coverage, topic details, and publisher information. We submitted 5 queries: Measles, Metabolic Syndrome, National Referendum

[2] http://www.google.com
[3] http://www.australia.gov.au
[4] http://www.alachuatoday.com

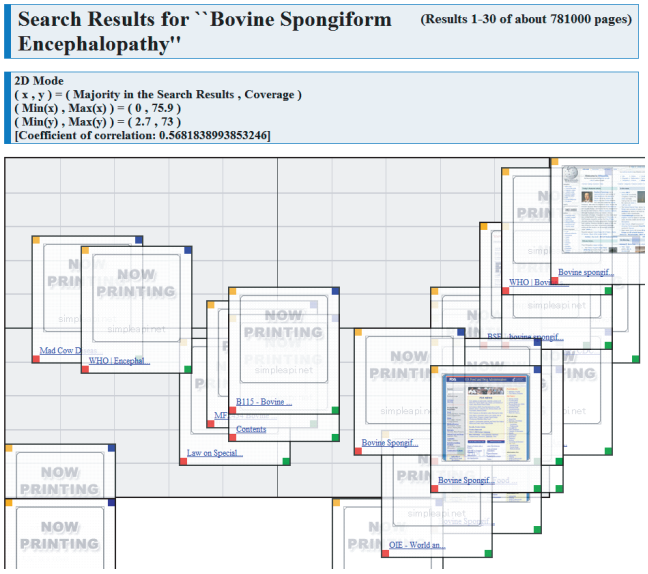

Fig. 10. Two-dimensional allocation display mode

Table 1. Time Analysis for Locality Support

Steps	**Avg. time (sec)**
Mapping of URLs	0.416
Link analysis	7.619
Retrieval of link sources (for 50 links)	1.088
Locating of link sources (for 50 links)	3.899
Graph consruction	0.090
Calculation of locality support	0.000
Rendering	0.004
Storing of cache	0.015
Total time	8.150

Bill, Tokyo Midtown, and French President. The average processing time of top 10 pages for each query is 7.2 seconds and that of top 50 pages is 28 seconds.

The calculation time for locality support was obtained as average values of 9 pages. The result shows that most of the time comes from link analysis, which is dependent on the response time of the Web search engine that returns URLs of link sources (Table 1).

We tested our system on a computer equipped with Windows Vista, processor 1.83GHz, RAM 2GB.

4 Conclusion

We developed a way to help search engine users to determine the trustworthiness of Web search results by computing and showing several different types of

information concerned with the search results. We first surveyed users to understand the way they search the Web, how they determine the trustworthiness of search results, and user expectations of search engines. The supporting information that our system provides must be computed in real-time when users execute queries on search engines. Because of limited computation time, we restricted the supporting information to that which could be computed efficiently by accessing search engines. The future problems are how to extract valuable supporting information in a more efficient manner from Web search engines and the Internet.

Acknowledgments

This research was supported by MEXT Grant-in-Aid for Scientific Research in Priority Areas entitled: "Content Fusion and Seamless Search for Information Explosion" (Project Leader: Katsumi Tanaka, A01-00-02, Grant#: 18049041) and "Design and Development of Advanced IT Research Platform for Information" (Project Leader: Jun Adachi, Y00-01, Grant#: 18049073), by MEXT Grant-in-Aid for "Development of Fundamental Software Technologies for Digital Archives", Software Technologies for Search and Integration across Heterogeneous-Media Archives (Project Leader: Katsumi Tanaka), and by MEXT Grant-in-Aid for Young Scientists (B) (Grant#: 18700086, 18700111, 18700129).

References

1. Fogg, B.J., Marshall, J., Laraki, O., Osipovich, A., Varma, C., Fang, N., Paul, J., Rangnekar, A., Shon, J., Swani, P., Treinen, M.: Web credibility research: A method for online experiments and some early study results. CHI 2001 (2001) 295–296
2. Fogg, B.J., Kameda, T., Boyd, J., Marshall, J., Sethi, R., Sockol, M.: Stanford-Makovsky web credibility study 2002: Investigating what makes web sites credible today. Report from the Stanford Persuasive Technology Lab, Stanford University (2002)
3. Fogg, B.J.: Prominence-interpretation theory: Explaining how people assess credibility online. CHI2003 (2003) 722–723
4. Stanford Guidelines for Web Credibility: http://www.webcredibility.org/guidelines/index.html.
5. Zaihrayeu, I., da Silva, P.P., McGuinness, D.L.: IWTrust: Improving user trust in answers from the web. iTrust2005 (2005) 384–392
6. Yanbe, Y., Jatowt, A., Nakamura, S., Tanaka, K.: Can social bookmarking enhance search in the web? JCDL2007 (2007) (to appear).
7. Yamamoto, Y., Tezuka, T., Jatowt, A., Tanaka, K.: Honto? search: Estimating trustworthiness of web information by search results aggregation and temporal analysis. APWeb/WAIM2007 (2007) (to appear).
8. Ma, Q., Tanaka, K.: Retrieving regional information from web by contents localness and user location. AIRS 2004 (2005) 301–312
9. Zhang, J., Ishikawa, Y., Kurokawa, S., Kitagawa, H.: LocalRank: Ranking web pages considering geographical locality by integrating web and databases. DEXA2005 (2005) 145–155
10. Oyama, S., Tanaka, K.: Query modification by discovering topics from web page structures. APWeb2004 (2004) 553–564
11. MaxMind: http://www.maxmind.com/.