Improving Search and Information Credibility Analysis from Interaction between Web1.0 and Web2.0 Content

Katsumi Tanaka, Satoshi Nakamura, Hiroaki Ohshima,
Yusuke Yamamoto, Yusuke Yanbe, Makoto Kato
Graduate School of Informatics, Kyoto University, Kyoto, Japan
Email: {tanaka, nakamura, ohshima, yamamoto, yanbe, kato}@dl.kuis.kyoto-u.ac.jp

Abstract— **We describe a new concept for improving Web search performance and/or increasing the information credibility of search results using Web1.0 and Web2.0 content in a complementary manner. Conventional Web search engines still suffer from a low precision/recall ratio, especially for searching multimedia content (images, videos etc.). The quality control of Web search is generally insufficient due to low publishing barriers. As a result, there is a large amount of mistaken and unreliable information on the Web that can have detrimental effects on users. This calls for technology that facilitates the judging of the trustworthiness or credibility of content and the accuracy of the information that users encounter on the Web. Such technology should be able to handle a wide range of tasks: extracting credible information related to a given topic, organizing this information, detecting its provenance, and clarifying background, facts, and other related opinions and their distribution. We propose and describe a concept of enhancing the search performance of conventional Web search engines and analyzing information credibility of Web information using the interaction between Web1.0 and Web2.0 content. We also overview our recent research activities on Web search and information credibility based on this concept.**

I. INTRODUCTION

As computers and computer networks become more common, a large amount of information, such as that found in Web content including multimedia (images, videos etc.), has been accumulated and circulated. Such information gives people a framework for organizing their private and professional lives.

However, the quality control of Web content is generally insufficient due to low publishing barriers. As a result, there is a large amount of mistaken and unreliable information on the Web that can have detrimental effects on users. This calls for technology that facilitates the judging the trustworthiness of content and the accuracy of the information that users encounter on the Web [1][2][3]. Such technology should be able to handle a wide range of tasks: extracting credible information related to a given topic, organizing this information, detecting its provenance, and clarifying background, facts, and other related opinions and their distribution.

Conventional Web search engines still suffer from low precision/recall ratio, especially for searching multimedia content (images, videos etc.).

We propose a concept of enhancing conventional Web search and analyzing Web information credibility using Web1.0 and Web2.0 content in a complementary manner. Based on this concept, we overview our research activities on Web search and information credibility.

Web2.0 content, such as social bookmarks, social tagging, Blogs, SNS, QA site content, and Wikipedia, is a valuable collection of human knowledge created by users in a collaborative manner. In other words, Web2.0 content is called consumer-generated media (CGM). For example, Wikipedia is a user-generated encyclopedia, which is called "collective knowledge" on the Web. Blogs contain a large volume of user-side evaluation based on their experiences and viewpoints. Several Web QA sites offer users a large amount of collective knowledge consisting of questions and answers generated by other users. Unfortunately, their information credibility is, however, not always guaranteed [4]. It should be noted that most Web2.0 content is generated in a way that it is isolated from Web1.0 content.

As for Web1.0 content search, conventional Web image search engines still suffer from low precision/recall ratio because of a lack of metadata for searching images. Also, conventional Web (text) search engines do not always accept or use subjective evaluation terms (such as "good" or "useful") as query keywords. On the other hand, for example, social tagging information in Web2.0 content is valuable annotation data from the user viewpoint. Such social tagging information may help conventional Web search engines improve their precision/recall ratio or to support a wider class of query terms.

As shown in Figure 1, our proposing concept leads to (1) improving Web1.0 search by knowledge extracted from Web2.0, and (2) evaluating Web2.0 information credibility by aggregating Web1.0 information. We overview our recent research on Web search and their information credibility based on this concept.

doi:10.4304/jsw.5.2.154-159

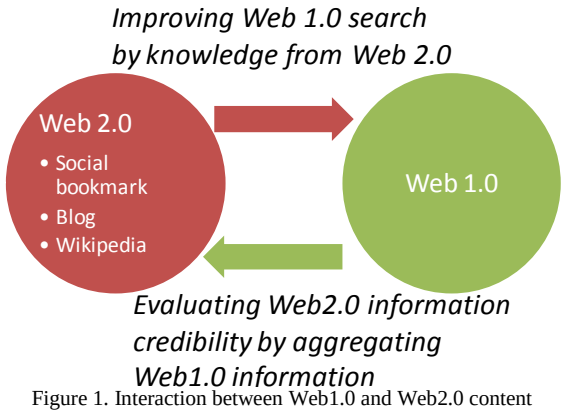

Figure 1. Interaction between Web1.0 and Web2.0 content to enhance each other

The knowledge extracted from Web2.0 information will improve Web1.0 search engines and help to determine information credibility of Web1.0 information. Also, the knowledge extracted and aggregated from Web1.0 information will help users to determine information credibility of Web2.0 information.

The remainder of this paper is organized as follows. In Section 2, we introduce our work concerned with the improvement of search and the credibility analysis of Web1.0 information using Web2.0 knowledge. In Section 3, we introduce our work related to the credibility analysis of Web2.0 information using the knowledge extracted and aggregated from Web1.0. Finally, we conclude in Section 4.

II. IMPROVEMNT IN WEB1.0 SEARCH USING WEB2.0 KNOWLEDGE

A. *Can social bookmark enhance search on the Web?*

Social bookmarking is an emerging Web service that helps users share, classify, and discover interesting resources. In our previous paper [5][6], we explored the concept of an enhanced search, in which data from social bookmarking systems is used for enhancing search on the Web. We proposed combining the widely used link-based ranking metric with the one derived using social bookmarking data.

First, this adds "freshness" and "user-interestingness" as ranking metrics, as well as the precision of a standard link-based search by incorporating popularity estimates from aggregated data of bookmarking users, to conventional Web search engines.

Second, this brings about a new kind of search: "allowing social bookmarkers' subjective evaluation as query keywords" into conventional search engines. Besides improved relevance, social tags allow for a more complex quality estimation of pages. This can be achieved using sentiment tags, user comments, and general global statistics derived from user behavior in relation to pages. For example, it is possible to search for pages that feature certain characteristics like being "`useful`" or "`funny`".

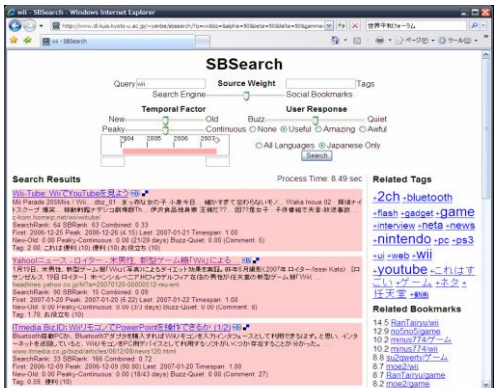

Figure 2. SBSearch: Reranking using Social Bookmarks

This type of sentiment analysis is not feasible using only page content or standard popularity rankings.

Additionally, social bookmarking systems allow for temporal search since bookmarks are usually time stamped. For example, it is possible to distinguish fresh pages from obsolete ones or to detect pages with certain popularity patterns.

Figure 2 shows a prototype system[5][6], called SBSearch, that implements our concept. Users can re-rank the search results by conventional search engines by using the number of social bookmarks, temporal distribution of social bookmarks, and social bookmarkers' subject evaluation terms.

B. *Can social tagging improve Web image search?*

Conventional Web image search engines return reasonably accurate results for queries containing concrete terms, but the results are less accurate for queries containing only abstract terms, such as "`spring`" or "`peace`." To improve the recall ratio without drastically degrading the precision ratio, we developed a method that replaces an abstract query term given by a user with a set of concrete terms and that uses these concrete terms in queries as input into conventional Web image search engines [7].

Concrete terms are found for a given abstract term using social tagging information extracted from a social photo sharing system, such as Flickr. This information is rich in user impressions about objects in images. Extraction and replacement are done by

(1) collecting social tags that include the abstract term,
(2) clustering the tags in accordance with the term co-occurrence of the images,
(3) selecting concrete terms from the clusters using ontological knowledge about terms from such databases as WordNet,
(4) finding sets of concrete terms associated with the target abstract term using a technique for association rule mining (see Figure 3)..

Our experimental results showed that our method improves the recall ratio of Web image searches (see Figures 4, 5, and 6).

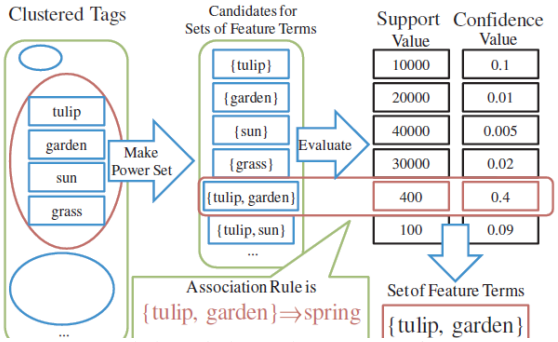

Figure 3. Association rule for "spring" extracted from Flickr tags

Figure 4. Search results for query "spring" using conventional Web image search engine

Figure 5. Search results for query "spring" using Flickr tags

Figure 6. Search results for query "cute" using Flickr tags

C. *Can Blog Photos Help Users to Evaluate Trustworthiness of Web Ad Photo?*

It is common to see many ads on the Web. An advertisement usually consists of a textual description and its corresponding images. The textual description points out the advantages and/or appealing points, and the corresponding images are used for evidence (see Figure 7).

Figure 7. Advertisement and photo gathered from blog of product

We proposed a method of analyzing the credibility of such text-image pairs on the Web[8]. Our method focuses on the consistency of the correspondence between a textual description and its corresponding image as one credibility criterion. Our basic idea is to estimate the image's credibility in a target text-image pair by gathering a set of images that are associated with similar text descriptions from Web2.0 content, such as blogs, and by analyzing the target image's "typicality" or "speciality" among the gathered images.

As shown in Figure 8, we proposed a method of calculating the "typicality" of a target image based on the VisualRank algorithm [9][10], which computes each image's dominance in a set of images based on their visual similarity.

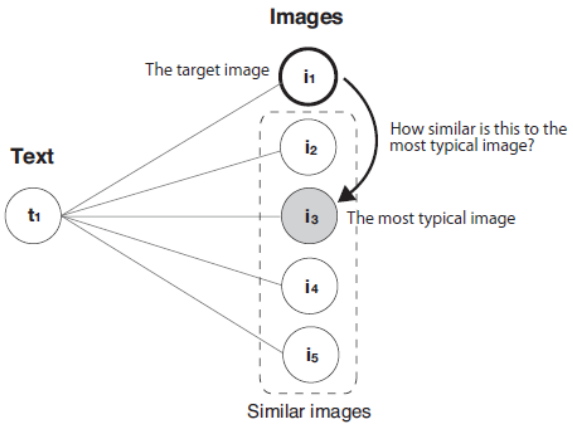

Figure 8. Method of calculating "typicality" of a target image

For example, if a beautiful image of a product in an ad is very different from actual images of the product gathered from blogs, we determine that the target image's credibility is low. On the other hand, if an ad photo image is used to demonstrate that a product is special compared with many other similar products, and the ad photo is regarded as credible if it is very different from other product images. We have developed a prototype system

called ImageAlert, which supports the assessment of a text-image pair's credibility when the user doubts this when browsing a Web page (see Figure 9).

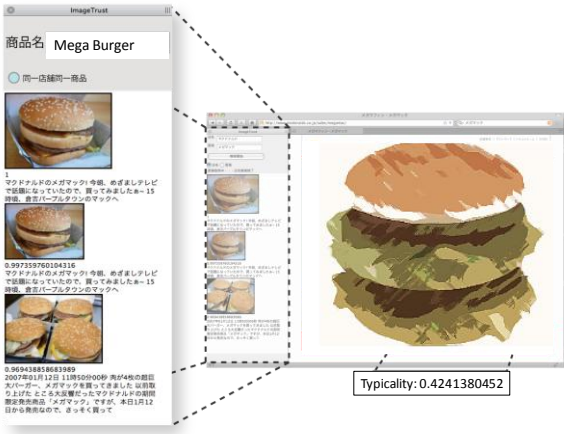

Figure 9. ImageAlert system

D. *Can Social Annotation Support Users to Evaluate Trustworthiness of Video Clips?*

Recently, video sharing Web sites, where users can upload and view video clips, have become extremely popular. YouTube is the most popular video sharing Web site in the world. On a video sharing Web site, there are two types of users (see Figure 10). One type is an uploader who uploads a video clip to a video sharing Web site. The other is a viewer who views and gives an annotation (comment) on the uploaded video clips.

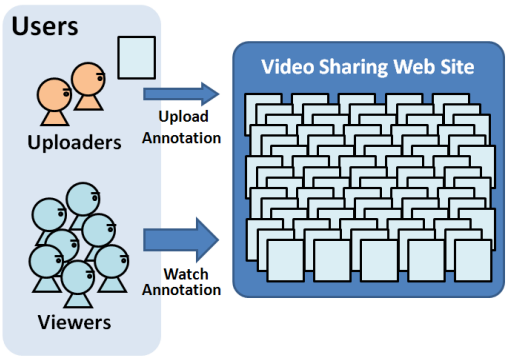

Figure 10. Image of video sharing Web site

We proposed a support system for evaluating the trustworthiness of video clips on a video sharing Web site [11]. Our system analyzes user-posted comments of a video clip and displays the temporal changes of sentiments of user comments. A viewer user judges if a video clip is trustworthy or not, by looking at the visualized sentiments of other viewers' comments. Our system shows the changes in the positive and negative levels of comments by generating two types of time-related graphs. One is related to playback time, and the other is related to the date of a posted comment. We implemented the proposed system and we developed two dictionaries for classifying comments into positive or negative and into two types of sentiments (happy and sad).

Video credibility is concerned with the consistency between the video's description (textual description such as video title and snippet) and the video. In general, whether viewers' comments to a video are positive or negative is independent from whether the video is credible or not. However, if the viewer is anxious about the consistency, it would be a sign that the video is not credible. If the viewer gives a positive comment to the video, then we regard the video as credible. The relationship between video credibility and viewers' comment sentiments (positive or negative) depends on the video type. For example, the credibility of "how-to" videos, such as how to make something or how to cook, can be measured by the viewers' comment sentiment (positive or negative).

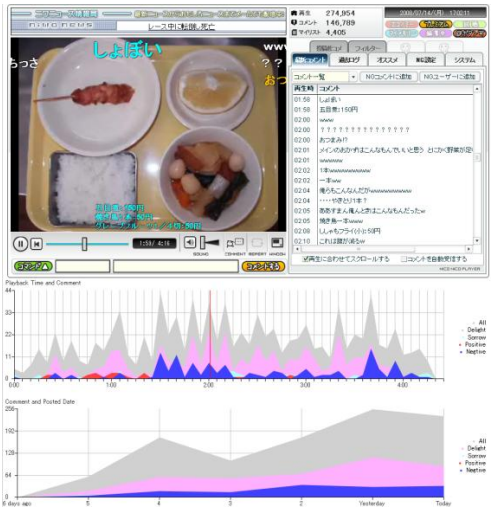

Figure 11. Visualizing sentiments of viewers' comments

III. CREDIBILITY ANALYSIS OF WEB2.0 CONTENT USING WEB1.0 KNOWLEDGE

A. *Supporting Credibility Judgment of QA Site Content by Aggregating Web1.0 Content*

When users see answers for a question on a QA site on the Web, it is often difficult to judge the credibility of each answer. While there are many factors influencing credibility perception of QA results, we introduced several criteria to predict the credibility of the answer: the degree of its similarity with other answers, the degree of "expertise" ("topic coverage"), and writing style. We developed a system that allows users to search QA site data and retrieves a question and corresponding answers from the QA site. The system shows each answer with its degrees of similarity, expertise and writing style when users select one of the searched answers. Figure 12 shows a screen shot of the system. The upper part shows a question, and the lower part shows the corresponding answers. The degrees of the topic coverage and expertise and writing style for each answer are shown on the right. If the answer is similar to other answers (a high degree of similarity), it contains many "typical" topics about the question (a high degree of expertise), and is written in a polite writing style, users can basically determine that the answer is credible.

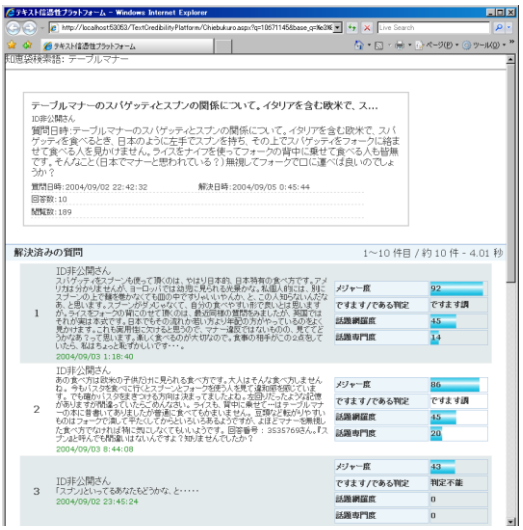

Figure 12. FAQ data analysis

Answer "majority" indicates to what extent the answer is similar to the majority of other answers. First, similarities between the target answer and other answers are calculated by computing the similarities of feature vectors of sentences in the answers. Feature vectors of answers are calculated using Term Frequency - Inverse Document Frequency (TF-IDF) weighting. The majority becomes high when the similarities between the answer and the other answers are high. The following formulas show how to calculate the majority:

$$Major_i = \frac{1}{k-1}\sum_{j,j\neq i} MajorWeight(i,j)$$

$$MajorWeight(i,j) = \begin{cases} 1 & (0.04 \le Similarity(i,j)) \\ \dfrac{Similarity(i,j)}{0.04} & (0.01 \le Similarity(i,j) < 0.04) \\ 0 & (Similarity(i,j) < 0.01) \end{cases}$$

Here, *k* is the number of answers and *Similarity(i,j)* is the cosine similarity between answers.

The degree of expertise ("topic coverage") of an answer indicates how much the answer covers "typical" topics of the question. It is calculated by counting the number of "typical" topic terms an answer contains. We developed a method for obtaining "typical" topic terms when a subject term is given. The search query terms are regarded as the subject terms, and "typical" topic terms are obtained from accessing a conventional Web search engine. An assumption in obtaining typical topic terms is that they frequently appear in the following linguistic pattern:

"<topic> of <subject>",

where <topic> is a topic term and <subject> is a given subject term. Another linguistic pattern is

"about <topic>" ,

where the sentence contains the subject term. Text resources that contain these patterns can be easily obtained by accessing a conventional Web search engine.

The degree of expertise of an answer indicates how much an answer is mentioned from a technical perspective. The number of technical terms contained in an answer is calculated. Technical terms concerned with a query term are obtained using a method we recently developed [12].

Writing style is whether an answer is written politely or not. This is judged using natural language processing functions.

B. Supporting Judgment of Fact Trustworthiness considering Temporal and Sentimental Aspects

We have developed a system for helping users determine the trustworthiness of uncertain facts based on sentiment and temporal viewpoints by aggregating information from the Web [13][14]. Our goal is not to determine whether uncertain facts are true or false, but to provide users with additional data on which the trustworthiness of the information can be judged. The system shows with what sentiment and in what context facts are mentioned on the Web and displays any temporal change in the fact's popularity. Furthermore, the system extracts counter facts and analyzes them in the same way.

We have developed an extended search system, called Honto? Search, to help users more accurately determine the trustworthiness of facts on the Web. The system has three key factors: counter example extraction, sentiment distribution analysis, and popularity evolution analysis of facts. Honto? Search proposes counter examples to the input fact and provides a framework for their temporal and sentimental analysis. Sentiment analysis is used to categorize Web pages containing information about a doubtful fact as positive or negative and to present final sentiment distribution. This approach is augmented with the prior construction of a large-scale sentiment term dictionary from the Web. A temporal approach is also applied to analyze changes in popularity of facts over time and to display them to users.

For example, if a user inputs "Tulips are native to the Netherlands" as an uncertain fact and "the Netherlands" as a verification target into Honto? Search, our system returns several facts (see Figure 13). The basic idea is to extract patterns which match

"tulips are native to *(wildcard)"

in snippets returned from a Web search engine. From the result ranking, the user can find that the most popular fact is "tulips are native to Turkey (this is the correct answer)".

In addition to calculating simple fact popularity, our system estimates the sentiment behind the facts by analyzing context in aggregated Web pages. We use the Naive Bayes Classifier to categorize content as "positive" or "negative". The system shows the ratio of positive to negative sentiments to users. Figure 13 illustrates the overall organization of our system.

Our Honto? Search system can be applied to gathering information from the Web for analyzing the credibility of sentences that appear in encyclopedic sites such as Wikipedia, or QA sites. This also corresponds to analyzing information credibility of Web2.0 content using knowledge extracted from Web1.0.

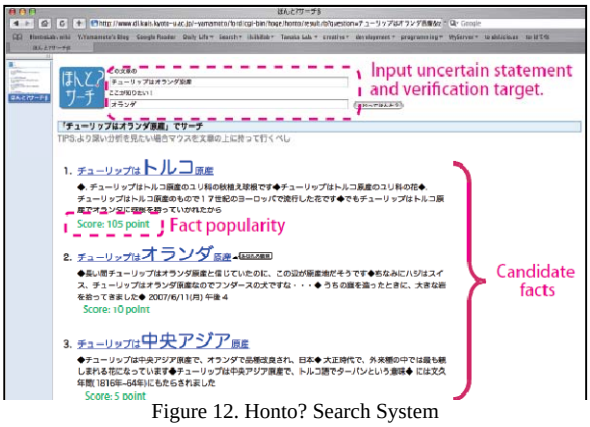

Figure 12. Honto? Search System

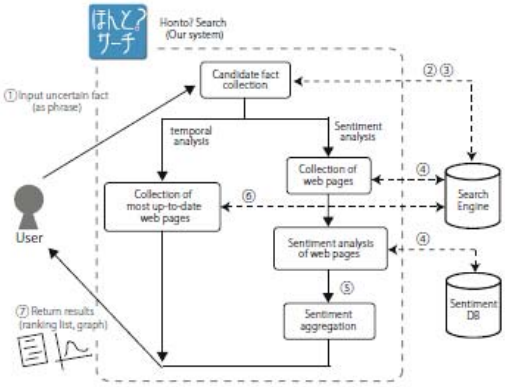

Figure 13. Overall organization of Honto? Search

IV. CONCLUSION

We explained a concept of useful interaction between Web 1.0 and Web 2.0 content. The knowledge extracted from Web2.0 content will improve Web1.0 search engines and help in judging the credibility of Web1.0 content. Also, the knowledge extracted and aggregated from Web1.0 information will help users judge the credibility of Web2.0 information. We overviewed our concept and corresponding research concerned with this concept.

ACKNOWLEDGMENTS

This work was supported in part by the MEXT Grant-in-Aid for Scientific Research in Priority Areas entitled: "Content Fusion and Seamless Search for Information Explosion" (Grant#: 18049041), the MEXT Global COE Program entitled "Informatics Education and Research Center for Knowledge-Circulating Society", and the National Institute of Information and Communications Technology (Information Credibility project).

REFERENCES

[1] B. J. Fogg, J. Marshall, O. Laraki, A. Osipovich, C. Varma, N. Fang, J. Paul, A. Rangnekar, J. Shon, P. Swani, M. Treinen: What makes Web sites credible?: a report on a large quantitative study. Proceedings of the SIGCHI conference on Human factors in computing systems , CHI2001, pp.61–68, 2001.

[2] K. Tanaka, T. Matsuyama, E.-P. Lim, A. Jatowt Eds.: Proceedings of the 2nd ACM Workshop on Information Credibility on the Web, WICOW 2008, Napa Valley, California, USA, October 30, 2008.

[3] S. Nakamura, S. Konishi, A. Jatowt, H. Ohshima, H.Kondo, T. Tezuka, S. Oyama, and K. Tanaka: Trustworthiness Analysis of Web Search Results, Proc. of the 11th European Conference on Research and Advanced Technology for Digital Libraries (ECDL 2007), LNCS, Vol. 4675, pp. 38-49, Budapest, Hungary, September 2007.

[4] N. L. Waters: Why you can't cite Wikipedia in my class, Communications of the ACM, Volume 50 , Issue 9, pp. 15 - 17 , September 2007

[5] Y. Yanbe, A. Jatowt, S. Nakamura and K.Tanaka: Can Social Bookmarking Enhance Search in the Web?, Proc. of the 7th ACM/IEEE-CS Joint Conference on Digital Libraries (JCDL2007), pp. 107-116, Vancouver, Canada, June 2007.

[6] Y. Yanbe, A. Jatowt, S. Nakamura, and K. Tanaka: Towards Improving Web Search by Utilizing Social Bookmarks, Proc. of the 7th International Conference on Web Engineering (ICWE2007), LNCS, Vol. 4607, pp. 343-357, July 2007.

[7] M. Kato, H. Ohshima, S. Oyama and K. Tanaka: Can Social Tagging Improve Web Image Search?, Proceedings of the 9th International Conference on Web Information Systems Engineering (WISE 2008) Lecture Notes in Computer Science, Vol. 5175, pp. 235-249, September 2008.

[8] Y. Yamamoto, T. Yamamoto, S. Nakamura and K. Tanaka: Image Credibility Analysis based on Typicality and Speciality in Related Images (in Japanese), Proc. WebDB Forum 2008, December 2008.

[9] Y. Jing and S. Baluja. Pagerank for product image search. In Proc. of the 17th international conference on World Wide Web (WWW 2008), pp. 307–316, 2008.

[10] Y. Jing and S. Baluja. VisualRank: Applying PageRank to Large-Scale Image Search. IEEE Transactions on Pattern Analysis and Machine Intelligence, 30(11):1877–1890, 2008.

[11] S. Nakamura, M. Shimizu, and K. Tanaka: Can Social Annotation Support Users in Evaluating the Trustworthiness of Video Clips?, Proceedings of the 2nd ACM Workshop on Information Credibility on the Web (WICOW 2008), Napa Valley, California, USA, pp. 59-62, October 2008.

[12] M. Nakatani , A. Jatowt , H. Ohshima , K. Tanaka: Quality Evaluation of Search Results by Typicality and Speciality of Terms Extracted from Wikipedia, Proceedings of the 14th International Conference on Database Systems for Advanced Applications, April 2009. (to appear)

[13] Y. Yamamoto, T. Tezuka, A. Jatowt, and K.Tanaka: Honto? Search: Estimating Trustworthiness of Web Information by Search Results Aggregation and Temporal Analysis, Proc. of the Joint Conference of the 9th Asia-Pacific Web Conference and the 8th International Conference on Web-Age Information Management (APWeb/WAIM'07), LNCS Vol. 4505, pp. 253-264, June 2007.

[14] Y. Yamamoto, T. Tezuka, A. Jatowt and K. Tanaka: Supporting Judgement of Fact Trustworthiness Considering Temporal and Sentimental Aspects, Proceedings of the 9th International Conference on Web Information Systems Engineering (WISE 2008), Auckland, New Zealand, Springer LNCS, Vol. 5175, pp. 206-220, September 2008.