Enhancing Credibility Judgment of Web Search Results

Yusuke Yamamoto
JSPS Research Fellow and Kyoto University
Yoshida-Honmachi, Sakyo, Kyoto, Japan
yamamoto@dl.kuis.kyoto-u.ac.jp

Katsumi Tanaka
Kyoto University
Yoshida-Honmachi, Sakyo, Kyoto, Japan
tanaka@dl.kuis.kyoto-u.ac.jp

ABSTRACT
In this paper, we propose a system for helping users to judge the credibility of Web search results and to search for credible Web pages. Conventional Web search engines present only titles, snippets, and URLs for users, which give few clues to judge the credibility of Web search results. Moreover, ranking algorithms of the conventional Web search engines are often based on relevance and popularity of Web pages. Towards credibility-oriented Web search, our proposed system provides users with the following three functions: (1) calculation and visualization of several scores of Web search results on the main credibility aspects, (2) prediction of user's credibility judgment model through user's credibility feedback for Web search results, and (3) re-ranking of Web search results based on user's predicted credibility model. Experimental results suggest that our system enables users — in particular, users with knowledge about search topics — to find credible Web pages from a list of Web search results more efficiently than conventional Web search interfaces.

Author Keywords
Web search, credibility analysis, credibility feedback

ACM Classification Keywords
H.5.4 Hypertext/Hypermedia; H.3.3 Information Search and Retrieval: Search process.

General Terms
Design, Human Factors.

INTRODUCTION
Today, the Web is becoming important information source in our daily life. A large amount of information is uploaded on to it, varying from lightweight readings like product information to serious readings that affect our daily life like news, product information, medical information, and so on.

Users can freely obtain and publish information on the Web, but most information is not fact-checked before it is uploaded, unlike in other media. Therefore, Web information is not always accurate or correct. For example, Sillence et al. reported that there are more than 20,000 medical Web sites on the Web, but over half are not checked by medical experts [23]. Denning et al. warned about the credibility of Wikipedia, which is one of the most popular websites [4]. On the other hand, some researchers reported that a lot of Web users trust Web information to some extent [15]. In particular, for Web search engines, which are used as gates to vast Web information, many users perceive their search results to be somewhat credible, and some users think that the Web search engines rank Web pages on the basis of credibility [16]. If users obtain Web information without caring enough about its credibility, they can be easily misled by incorrect Web information. In the worst cases, users can suffer from actual harm in their daily lives. Therefore, it is important to measure the credibility of Web information and to help users judge the credibility of Web information.

There has been a lot of research into evaluating Web information in terms of information retrieval. However, most has focused on ranking a set of data on the basis of relevancy or popularity by analyzing similarity between query and documents, link structures [3], and user-behavior data [12]. Recently, researchers have started to discuss information quality, and some have argued that the importance of credibility is a factor in information quality [21, 20]. However, research into information quality has only just started, and few studies have measured Web information credibility.

In the research area of communication and social psychology, credibility researcher has been an important topic since 1950s [11, 9, 14]. However, in those research fields, few researchers have developed practical methods to measure information credibility. In library science, some researchers developed a guideline to evaluate information credibility [13]. However, despite the presence of ready guidelines, it is difficult for many users to appropriately assess the credibility of Web information. This is partly because in contrast to traditional publishing, Web pages often lack clues to judge their credibility such as the credentials, names of authors, references to information sources, and evidence information. Consequently, few users rigorously judge the credibility of obtained information. Therefore, automatic tools for helping users judge Web information credibility are becoming increasingly necessary.

The aim of this paper is to bridge the gap between credibility theory in social psychology and search technology in Web

Submitted for review to CHI 2011.

information retrieval. In this paper, we propose a system to help users judge the credibility of search results that Web search engines provide and find credible Web pages. Information credibility is a subjective quality that differs depending on the users obtaining the information [9, 24]. Therefore, search systems should not measure the credibility of Web pages and provide search results while ignoring users. Our system's goal is to help users obtain credible Web search results while users judge the credibility of Web search results on the basis of their own credibility criteria.

Currently, Web search results contain only the information representing content of Web pages like titles, snippets, and URLs. These do not give enough clues to enable the credibility of Web pages to be judged. Therefore if users are not careful about the credibility of Web information, and even if Web pages in a list of Web search results are not credible, the non-credible Web pages are difficult for users to detect. On the other hand, even if users are careful about the credibility, it takes too long for them to judge the credibility of Web search results. This is because there are few clues to judge credibility of Web pages, and users need to compare them with other Web pages or to search for information to judge credibility. For example, when for credibility criteria users want to know whether a target Web page describes content similar to those of many other Web pages, users need to check a large amount of Web pages and compare them with the target Web page. Also, opaque ranking algorithms of Web search engines are a big problem when users judge the credibility of Web search results. Ranking algorithms of Web search engines do not always match users' criteria for credibility judgment. If users' credibility criteria do not match ranking algorithms of search engines, it is difficult to efficiently search for credible Web pages. For example, when users wish to obtain credible Web pages that are often updated and not biased by their authors, but conventional Web search engines provide more popular, but less credible Web pages.

If a system provides additional information for credibility judgment of Web search results and ranks Web search results considering the factors users think to be important in judging credibility, users can obtain more credible Web search results. To achieve such a system, we propose the following three functions as an extension system for Web search engines:

- Analysis and visualization of scores of Web search results on the main credibility factors.
- Prediction of users' credibility judgment model through users' credibility feedback for Web search results.
- Re-ranking of Web search results based on a predicted users' credibility model.

In the field of Web searches, a lot of researchers have focused on searching for Web pages relevant to users' information needs, but few researchers have discussed Web searches from the viewpoint of Web page credibility. Our proposed system enables users to do credibility-oriented Web searches.

In summary, the contributions of this paper are:

- We address the new concept of credibility-oriented Web searches and necessary requirements for achieving systems to help users do credibility-oriented Web searches.
- We introduce and implement three functions to help users judge the credibility of Web search results and to search for credible Web pages.
- We examine how our system affects users' credibility judgment of Web search results. We show that our proposed system enables users to find credible Web pages from a list of Web search results more efficiently than conventional Web search interface.

RELATED WORK

Credibility has been studied in communication and social psychology since 1950s [11]. In these fields, credibility is defined as a perceived quality and acceptance by receivers of messages or senders of the messages. Generally, people often think of the credibility of information as an objective quality like authenticity or accuracy. However, researchers in the above fields indicate that information credibility is a subjective quality, and its interpretation depends on the receivers and types of information [9, 24, 14]. In communication and social psychology, a main research topic of credibility is to reveal the mechanisms of or factors that affect credibility judgment. Many factors concerned with credibility are grouped into two key components: expertise and trustworthiness [9]. Some researchers focus on Web information credibility. Fogg et al. conducted a large-scale survey to investigate what factors affect users' perception of the credibility of Web sites [8]. Metzger et al. surveyed how college students check the credibility of Web information [15].

In the field of information science, some researchers developed methods to analyze the credibility of specific Web contents (such as Wikipedia, Q&A contents, and people on social networks) from specific credibility aspects. Vuong et al. developed a method to find controversial Wikipedia articles [28]. Adler et al. considered sentences that remain unedited as credible, and they developed an algorithm to evaluate the credibility of sentences in Wikipedia articles by analyzing edit histories [1]. Research aimed at evaluating the credibility of Q&A contents like those on Yahoo! Answers[1] is becoming more popular. Suryanto et al. focused on expertise of answerers on Q&A sites to evaluate answers posted on them [26]. Agichtein et al. evaluated authority and hub scores of users by using a graph structure representing the interactions between the users and then evaluated the quality of answers on the basis of those users' scores [2]. Guha et al. developed a framework to predict trust between two people on a large trust network through link analysis [10]. These studies focus on specific aspects to measure the credibility, but as mentioned before, information credibility varies in accordance with users and types of information. Therefore, credibility aspects for measuring the credibility of information must change depending on users and information types.

[1] http://answers.yahoo.com/

There have been some studies focusing on helping users judge credibility. Ennals et al. developed DISPUTE FINDER, the system that searches the Web for counter sentences when users specify questionable sentences on browsed Web pages [6]. Suh et al. developed WIKIDASHBOARD, a system to visualize edit histories on Wikipedia articles [25]. Pirolli et al. used WikiDashboard to study how the system affects users' credibility judgments on Wikipedia articles [19]. For support systems to judge the credibility of Web search results, Nakamura et al. developed a prototype system [16], but they did not evaluate its effectiveness for judging the credibility of Web search results.

SYSTEM DESIGN

Our goal is to realize a system to help users judge the credibility of Web search results like those on Google and Yahoo! so that users can find more credible Web pages more efficiently by themselves. In this section, we first summarize system requirements and then explain the approach for system implementation.

Requirements

As already mentioned, information credibility is a subjective quality, and its criteria depend on the users and types of information. Search results should contain clues to judge the credibility from various viewpoints, so that users can judge the credibility of Web search results. However, conventional Web search results contain only titles, snippets, URLs, and ranking orders. Therefore, users need to manually collect clues to judge the credibility of Web search results by browsing or comparing other Web pages, but it takes too long to make a solid credibility judgment. For these reasons, systems have to automatically collect and aggregate the information necessary to judge a Web page's credibility from various viewpoints, so that users can easily know how credible it is.

The goal of credibility-oriented Web searches is to obtain more credible Web pages from the large amount of Web pages relevant to a given query. To efficiently obtain credible Web pages, Web pages must be ranked in accordance with viewpoints of their credibility. Credibility criteria depend on the users and types of information. For example, as for Web pages about Apple products, some users think that unbiased Web pages are credible, and other users think that Web pages which Apple enthusiasts appreciate are credible. Therefore systems should not rank Web search results by using a pre-defined credibility-oriented ranking function. The important thing is to rank Web pages on the basis of users' credibility judgment model. To achieve this, systems need to predict users' credibility judgment model in some way.

To achieve the above requirements, in this paper, we propose the following three functions on our system:

- Analysis and visualization of scores of Web search results on main aspects to judge the credibility.
- Prediction of users' credibility judgment model through users' credibility feedback for Web search results.

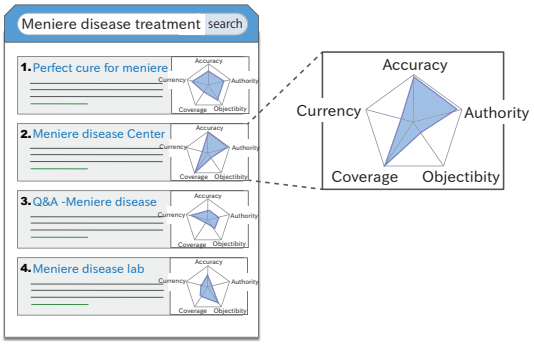

Figure 1. Visualization of scores for each of the main credibility aspects.

- Re-ranking of Web search results based on predicted users' credibility model.

Analysis and visualization of scores of Web search results for each of the main credibility aspects

In communication and social psychology, researchers have developed various taxonomies of credibility factors [5, 14]. In this paper, we selected the following five criteria, which many researchers mentioned, as the main factors to judge the credibility of Web search results [13]: accuracy, objectivity, authority, currency, and coverage. Accuracy means how accurate the Web information is, and it is measured, for example, by checking the number of errors, accuracy of links, the existence of sources, and so on. Objectivity means how biased the Web information is or is not. Authority is a criterion about the author of the Web page, and it is measured by reputation for the author and ability of the author to create contents, and so on. Currency is measured by checking when Web pages were last updated or whether Web pages are kept up to date. Coverage means how complete and comprehensive the Web information is.

Given a query and Web search results for it, our proposed system calculates scores of Web search results for each of the five aspects above. Each score is normalized by considering the distribution of the scores of all Web search results. This normalization of scores makes it easier to understand relative order of each Web search result from a specific aspect, and to compare the score of a Web page on one aspect with that on another aspect. To make users intuitively understand the scores of Web search results for each of the five credibility aspects, the system visualizes the scores as radar charts like in Figure 1. Visualization of such radar charts makes users aware of the credibility of Web search results. Furthermore, it enables users to intuitively judge the credibility of Web search results from various credibility aspects.

Prediction of user's credibility judgment model

To re-rank Web search results on the basis of users' credibility judgment criteria, we propose that the system predicts users' credibility judgment model through users' credibility feedback for Web search results. In this paper, we define the user's credibility judgment model as representation of how important the user think each of main credibility aspects to

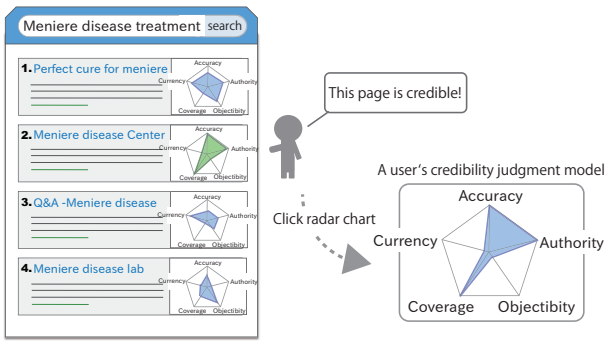

Figure 2. Prediction of user's credibility judgment model through credibility feedback for Web search results.

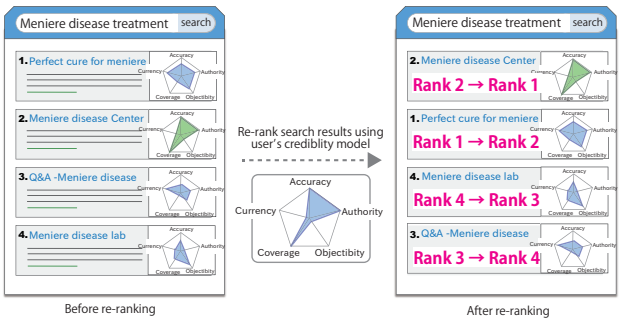

Figure 3. Re-ranking of Web search results using predicted user's credibility judgment model. Web search results are ordered in terms of accuracy, authority, and coverage.

judge the credibility of Web search results is. Users' credibility judgment model is represented by weights for main credibility factors. Users check a radar chart indicating credibility scores of a Web search result, and then they send feedback to the system by specifying the Web search result if they judge the Web search result to be credible, as shown in Figure 2. The system predicts the aspects that users consider important for judging credibility of Web search results, analyzing the scores of Web search results that users have fed back on each of the main credibility aspects. A predicted users' credibility model is shown as a radar chart on a Web browser in Figure 2, and the radar chart helps users check the predicted credibility model. In the case of Figure 2, the system predicts that the user think accuracy, authority, and coverage are the most important credibility aspects because the search result that the user specified has very high scores for these three aspects. Users can give credibility feedback to the system several times. Every time they do so, the system modifies the users' credibility judgment model that has been already predicted.

Re-ranking of Web search results using user's credibility judgment model

Our system re-ranks Web search results using a predicted users' credibility judgment model. The system analyzes credibility scores of Web search results in accordance with users' credibility judgment model and re-ranks Web search results. To calculate credibility scores of Web search results, the system uses scores of each Web search result for each of five credibility aspects and the predicted users' credibility judgment model. Every time users give credibility feedback to the system, the system re-ranks Web search results. Figure 3 illustrates an example in which the system re-ranks Web search results for the query "meniere disease treatment" by using the user's model predicted in Figure 2.

In information retrieval, a lot of researchers have developed methods for relevance feedback [22] or personalized searches [27]. These methods focus on optimizing Web search ranking for users just as our approach does. However, relevance feedback and personalized search focus on relevance of search objects for users' information needed to enhance search ranking. Also, in the two approaches, only the data that search objects contain in themselves is used to enhance relevance-oriented searches. The goal of these studies is different from that of our research because our research focuses on credibility-oriented search ranking. Our proposed system enables users to re-rank Web search results in accordance with users' credibility model and to efficiently search for credible Web pages, by analyzing content data and meta-data of Web pages.

CREDIBILITY ANALYSIS OF WEB SEARCH RESULTS FOR EACH OF THE MAIN CREDIBILITY FACTORS

In this section, we describe a method to measure accuracy, authority, objectivity, coverage, and currency of Web search results. There can be various factors related to these five aspects, but we can practically measure only a portion of them. In this work, we focus on the following possible six factors for measuring the main five credibility aspects:

- Accuracy: referential importance of Web page
- Authority: social reputation of Web page
- Objectivity: content typicality of Web page
- Coverage: coverage of technical topics on Web page
- Currency: freshness and update frequency of Web page

Our system provides users with each score for each of the above factors as referential importance, social reputation, content typicality, topic coverage, freshness, and update frequency, respectively.

Referential importance

We assume that accurate Web pages are often linked to by other Web pages as references, and we define referential importance as one of criteria to measure accuracy of Web pages. In many cases, referential importance of Web pages is often measured by PageRank algorithm, HITS algorithm, and so on. In our implementation of our prototype, we used scores of PageRank provided by Google[2].

Social reputation

We assume that if a lot of people believe or highly regard certain Web pages, the Web pages are authoritative, and we

[2]http://djangosnippets.org/snippets/221/

define social reputation as one criterion to measure authority of Web pages. We use numbers of social bookmarks for Web pages as scores of social reputation. Our prototype system uses Hatena Bookmark service[3] to obtain the number of social bookmarks for Web pages.

Content typicality

We assume that if a Web page is similar to many other Web pages about a given query, the Web page is objective. Thus we define content typicality about a given query as one criterion to measure objectivity of Web pages. We use the LexRank algorithm to calculate content typicality [7]. The LexRank algorithm summarizes text contents. By using the algorithm, a graph is created from text contents where text contents are nodes and textual similarity between text nodes is the weight of the edge, and centrality of text nodes is calculated by using the graph. In our study, the system measures content typicality of Web pages by creating textual feature vectors of Web pages and applying the LexRank algorithm to a set of Web pages about a given query.

Topic coverage

To measure how Web pages in a list of search results cover technical topics about a given query, we use the method to search Wikipedia for technical terms that Nakatani et al. developed [17]. This method focuses on bias of the links among Wikipedia articles, and extracts the terms that Wikipedia articles only in specific categories link to as technical terms. Our system measures topic coverage of a Web page to calculate the number of technical terms on the Web page.

Freshness and update frequency

To measure freshness and update frequency of Web pages, we obtain a list of updated dates of Web pages using the Wayback Machine of Internet Archive[4]. The Wayback Machine archives versions of Web pages across time. Next, we denote a set of dates when a Web page was updated, by $T = \{t_1, t_2, \ldots, t_n\}$ $(t_1 < \cdots < t_n)$. Our system calculates freshness and update frequency of the Web page as $C - t_n$ and $|T|$, respectively (C is a constant).

RE-RANKING OF WEB SEARCH RESULTS THROUGH CREDIBILITY FEEDBACK

Scores of Web search results on each of six credibility factors are visualized by radar charts. When users check credibility radar charts of Web search results, if users judge some Web search results to be credible, users can feed them back to our system by double-clicking the radar charts of the credible Web search results. The system predicts users' credibility judgment model through users' feedbacks, and then the system re-ranks Web search results in accordance with the predicted model. In this paper, we use the six kinds of scores mentioned in the previous section, to represent the features of Web search results and users' credibility judgment model. Here we denote the feature vector $\boldsymbol{v}(p)$ of Web page p by $\boldsymbol{v}(p) = (s_1(p), s_2(p), \ldots, s_6(p))^{\mathrm{T}}$, where $s_i(p)$ means a normalized score of p on credibility factor i. When the user judges Web page p as credible and sends a credibility feedback to our system, the system predicts the user's credibility judgment model $\boldsymbol{u} = (u_1, u_2, \ldots, u_6)^{\mathrm{T}}$ using $\boldsymbol{v}(p)$. Model $\boldsymbol{u}$ represents how important the user thinks each of the six credibility factors to be.

Where a set of Web pages that the user judges to be credible is denoted by $P = \{p_1, p_2, \ldots, p_n\}$, we define the n-th degree of the user's credibility judgment model $\boldsymbol{u}$ as

$$u_n = \left(\frac{1}{n} \sum_{p \in P} \frac{s_n(p)}{|\boldsymbol{v}(p)|}\right)^3. \quad (1)$$

Here $|\boldsymbol{v}|$ is a norm of vector $\boldsymbol{v}$. The system cubes the arithmetic average of Web pages' scores for each credibility factor because we wish to emphasize scores for credibility factors that users think important/unimportant.

The system re-ranks Web search results in accordance with the predicted user's credibility judgment model $\boldsymbol{u}$. Web search result p is scored by the following ranking function:

$$\mathrm{rank}(p) = \boldsymbol{u}^{\mathrm{T}} \cdot \boldsymbol{v}(p) = \sum_{k=1}^{6} u_k s_k(p) \quad (2)$$

PROTOTYPE SYSTEM

We implemented our prototype system as an extension of Safari Web browser[5]. We used Javascript to implement the following three functions on Safari browsers: visualization of radar charts of credibility analysis, credibility feedback to the system, and re-ranking of Web search results. Also, we implemented the function to calculate scores of Web search results on each of the credibility factors on our server by using Python and Sqlite3.

When users run the system on Google's search engine result pages, (1) the system inserts radar charts that illustrate scores of Web search results on each of credibility factors into search results, and (2) users can re-rank the search results in accordance with their credibility judgment model by double-clicking radar charts of credible Web search results and sending feedbacks for the credible Web search results to the system. Immediately after running the system, Google's original ranking is directly presented. Figure 4 shows an example where a user uses our system for Web search results for "meniere disease treatment".

EXPERIMENT

To examine how using our system affects users' credibility judgments of Web search results, we conducted experiments using our prototype system.

Experiment #1: Online user study

We conducted an online user study (1) to investigate which factors are important for users to judge the credibility of Web search results about specific topics, and (2) to investigate how different the consistency of credibility judgment is depending on users' familiarity with search topics when

[3] http://b.hatena.ne.jp/help/api

[4] http://www.archive.org/web/web.php

[5] http://cowsearch.hontolab.org/ (Web service version)

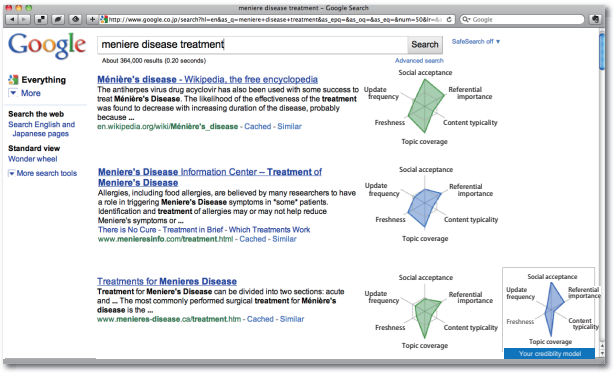

Figure 4. Screenshot of our system. In this example, a user uses our system for search engine results for "meniere disease treatment".

Table 1. Query set.

Topic	Query set A	Query set B
Disease	Meniere disease treatment	Parkinson disease treatment
Politics	Dual surnames problem	Child allowance problem
Science	Global warming cause	Earthquake cause
Health	Effective dieting	Reduce blood pressure
Sightseeing	Osaka famous place	Sendai famous place
Economics	Forex risk	Futures contract risk
Goods	Camera comparison	Video camera comparison
Food	Margarine risk	Genetically-modified food risk
History	Yamataikoku location	Kamakura shogunate formation year
Law	Personal bankruptcy procedure	Succession of property procedure

users judge the credibility of Web search results by checking their surface information (title, snippet, and URL) and their scores on various credibility factors.

Materials

We manually selected ten topic categories and prepared two queries (and tasks) for each category by considering topics whose credibility is important to examine Web pages about. A set of the prepared queries and topics is shown in Table 1. We randomly selected ten relevant Web search results and their actual Web pages for each query from Google search results before this experiment and presented them to participants in the experiment.

Participants

A total of 960 participants were recruited via an online questionnaire company. The participants were recruited so that their sex and their age (from 20's to 50's), spread evenly. The participants were randomly assigned to one of 20 groups. The participants received $2 as compensation for their time.

Procedure

The participants joined this experiment through the Internet. In the experiment, the participants of each group conducted two tasks for each topic category in Table 1.

Before starting tasks, the participants were asked to answer two questions. The first question was about (1) how important they thought each of nine factors was to judge the credibility of Web information about one specific topic of the ten topics on a 5-point Likert scale (from -2="strongly not unimportant" to 2="strongly important"). The nine factors are *content expertise*, *social reputation*, *update frequency*, *content freshness*, *authority of page creator*, *content typicality*, *evidence presentation*, *objectivity*, and *accuracy*. The second question was about how familiar the participants were with the topic on a 3-point Likert scale (1="familiar", 2="neutral", and 3="unfamiliar").

After answering the two questions, the participants performed two tasks. Immediately before starting each task, we showed a brief description about the tasks like:

> *Suppose you or one of your family has meniere disease. You are now searching for Web pages. We will show some Web pages or Web search results about meniere disease. Please evaluate the credibility of them.*

In the first task, the participants rated each of 10 Web search results for either of two queries in the specific topic using the interface where credibility radar charts for Web search results were presented on Google search interface (called the +chart interface) on a 7-point Likert scale. They rated how credible they perceived each of the search results to be (from 1="strongly not credible" to 7="strongly credible"). In this process, the participants were limited to check the presented information to rate the credibility of Web search results without following their URL links. Next, as the second task, the participants were asked to rate each of actual Web pages for the Web search results in the same way. These tasks had no time limits for rating Web search results and their actual Web pages.

In this experiment, we controlled (1) topic categories, and (2) queries in each topic category to use for two tasks, to split them equally for participants. As a result, we divided the 960 participants into 20 groups of 48 people.

Results

First we report which factors the participants thought as important to judge the credibility of Web pages about the specific ten topics. Table 2 shows the average weights of multiple credibility factors for participants who were familiar with the topics (called knowledgeable participants) and those for participants who were not (called unknowledgeable participants). According to the table, for all topics, unknowledgeable participants thought accuracy as the most important credibility factor. Also, we found that they were not so confident about factors to judge the credibility (weight $<$ 1). On the other hand, the most important credibility factor for knowledgeable participants depended on the topic categories. Furthermore, knowledgeable participants thought several credibility factors to be very important in the six categories other than science, health, economics, and food (weight $>$ 1). In some topic categories, unknowledgeable participants thought a specific factor to be a little important while knowledgeable participants thought it to be a little unimportant (e.g. social reputation in economics), and vice versa (e.g. authority in food). In this way, credibility judgment models of knowledgeable participants and unknowledgeable participants differed.

Table 2. Average weights of factors to judge the credibility of information about topics in specific categories. Weight values range from -2 (strongly unimportant) to 2 (strongly important). Numbers without/with parenthesis are weights of factors for participants familiar/unfamiliar with specific topics. Bold value means the most important factor in specific topic. In (category x factor) with gray background, unknowledgeable participants' thoughts conflict with knowledgeable participants' thoughts.

Topic	Content expertise	Social reputation	frequency	freshness	Authority of page creator	Content typicality	Evidence presentation	Content Objectivity	Content Accuracy
Disease	0.923 (0.413)	0.231 (0.217)	0.077 (0.196)	0.923 (0.543)	0.923 (0.152)	0.154 (0.239)	**1.231** (0.413)	1.153 (0.500)	0.929 **(0.935)**
Politics	0.857 (0.419)	0.357 (0.210)	0.357 (0.081)	0.643 (0.306)	0.357 (0.016)	0.286 (0.177)	0.929 (0.613)	**1.000** (0.500)	**1.000** **(0.726)**
Science	0.800 (0.175)	0.200 (0.079)	0.300 (0.048)	0.900 (0.444)	-0.300 (-0.206)	0.200 (-0.063)	0.800 (0.333)	0.400 (0.317)	**1.300** **(0.730)**
Health	0.440 (0.353)	0.100 (0.382)	0.400 (0.265)	0.720 (0.441)	-0.600 (-0.118)	-0.140 (0.029)	0.440 (0.324)	0.480 (0.088)	**0.980** **(0.588)**
Sightseeing	0.667 (0.440)	0.400 (0.100)	0.800 (0.400)	**1.333** (0.720)	0.000 (-0.600)	0.133 (-0.140)	1.000 (0.440)	0.533 (0.480)	1.133 **(0.980)**
Economics	0.429 (0.397)	-0.071 (0.293)	0.214 (0.259)	0.571 (0.379)	0.071 (-0.017)	-0.214 (0.121)	0.071 (0.345)	0.071 (0.328)	**0.643** **(0.793)**
Goods	0.800 (0.508)	0.500 (0.220)	1.000 (0.220)	0.900 (0.542)	0.300 (-0.017)	0.400 (0.152)	**1.400** (0.458)	0.900 (0.492)	1.200 **(0.881)**
Food	0.659 (0.160)	0.561 (0.040)	0.463 (0.160)	0.756 (0.400)	0.146 (-0.200)	0.244 (0.040)	0.707 (0.480)	0.756 (0.440)	**1.000** **(0.680)**
History	1.000 (0.308)	0.200 (0.077)	0.650 (0.128)	1.100 (0.462)	0.150 (0.077)	0.200 (-0.026)	1.100 (0.359)	1.150 (0.205)	**1.450** (0.538)
Law	**1.136** (0.441)	0.455 (-0.088)	0.545 (0.294)	1.000 (0.618)	0.273 (0.029)	0.409 (0.206)	1.045 (0.676)	0.955 (0.382)	**1.136** **(0.912)**

Table 3. Mean number of consistent credibility judgments for different familiarity levels.

Familiarity level	Consistent credibility judgment		
	Good-Good or Bad-Bad	Neutral--Neutral	Total
Not familiar (470 people)	2.95 (2.15/0.80)	3.55	6.51
Neutral（309 people）	2.72 (1.94/0.78)	3.76	6.48
Familiar (181 people)	4.00 (3.00/1.00)	2.50	6.50

Next we focused on the participants' credibility ratings to examine whether the credibility judgments of Web pages were consistent or not when the participants checked the crediblity of Web pages when using the +chart interface and when viewing the pages' content only. "Good-Good" means that participants judged a Web page as credible (rate = 5, 6, 7) on both the +chart interface and its own URL. In addition, "Bad-Bad" means that participants judged a Web page as not credible (rate = 1, 2, 3) on both the +chart interface and its own URL. "Neutral-Neutral" means that participants judged the credibility of a Web page as unknown (rate = 4) on the +chart interface and its own URL. The larger the number of "Good-Good", "Bad-Bad", and "Neutral-Neutral" judgments, the more consistent credibility judgment for Web pages when checked on the +chart interface and on their own URLs. That is, we can assume that the +chart interface enables users to judge the credibility of Web pages in a list of Web search results more consistently without checking them on their own URL.

We examined the mean number of consistent credibility judgments per 10 Web pages for each participant, for different levels of participants' familiarity to topics. Table 3 shows the results focusing on the familiarity with search topics. According to the table, as for the total number of "Good-Good", "Bad-Bad", "Neutral-Neutral" credibility judgments, there was no statistically significant difference depending on the participants' familiarity with search topics (not familiar: 6.51, neutral: 6.48, familiar: 6.50). However, we made an interesting finding in the analysis on different types of consistent credibility judgment. Focusing on the total number of "Good-Good" and "Bad-Bad" credibility judgments (that is, black-and-white judgments), that of the participants who were familiar with the search topics was significantly larger than that of the participants that were not so familiar with the topics (familiar vs. not familiar: $4.00 > 2.95$, $t(649) = 3.316$, $p < 0.001$) (familiar vs. neutral: $4.00 > 2.72$, $t(488) = 3.915$, $p < 0.001$). This suggests that if users familiar with search topics use the +chart interface, they can make a consistent and clear judgment on the credibility of Web search results, while users unfamiliar with the topics often make an unclear credibility judgment.

Experiment #2: Laboratory study

One of our system's expected advantages is that users will be able to find credible Web pages from Web search results more efficiently. To evaluate search efficiency of our system, we compared how many credible Web pages could be selected from search results within a limited time using our system with the number obtained using a conventional Web search interface. In addition, we investigated how users use our system to search for credible Web pages.

Design and materials

Across Web search results for two queries in the same topic category, we contrasted the participants who used our system + Google interface against a baseline condition involving just the original Google search interface. Participants were asked to select credible Web pages from Google's results for one query of a category using one interface and to select credible Web pages from search results for the other query in the category using the other interface.

We used the same query set as the previous online study and

prepared two tasks for each of the ten topic categories to check the credibility of Web search results.

Participants

We recruited ten students from our university as participants: three undergraduates, and seven postgraduates. The participants received $20 as compensation for their time. We randomly assigned them to either of group 1 or 2.

Procedure

Participants did our experiment individually. We briefly introduced credibility judgment tasks and our system. The participants also practiced answering a task by using our system on their own PC for Web search results for a sample query.

After the task introduction and the practice, the participants went through 10 tasks using our system + Google and 10 tasks using only Google. Before starting each task, we briefly described the task like this:

> *Suppose you or one of your family has meniere disease. You are now searching for Web pages that Google returned for "meniere disease treatment". Please select as many credible Web pages as possible within three minutes.*

After the participants read the description and started each task, the system automatically showed the top 50 Google Web search results for a prepared query about the task. The participants were asked to select as many credible Web pages as possible within three minutes by using the normal Google search interface or our system + Google interface (we fixed which interface users used for each query in advance). If necessary, participants were allowed to follow links from search results. We asked the participants not to modify the query. In this experiment, if the participants found credible Web pages from search results, they were asked to double-click their search results. When search results were double-clicked, the experimental system changed their background color. When three minutes passed, the participants stopped performing the task and the experimental system stored the information about selected Web search results. In using our system, the participants' credibility feedback was stored for analysis of users' behavior to judge the credibility later.

The experimental system assigned twenty tasks to the participants in the same order and alternated between query set A and B (disease A $\rightarrow$ disease B $\rightarrow$ politics A $\rightarrow$ politics B $\rightarrow \cdots \rightarrow$ law B). The five participants in group 1 performed the tasks for query set A and B using only Google and then Google + our system, respectively. The five participants in group 2 performed the tasks for query set A and B using Google + our system and then only Google, respectively.

Results

Table 4 presents the mean and the standard deviation of the number of credible Web pages the participants selected by using the two different interfaces (only Google vs. our system + Google). According to the table, the mean of the number of selected credible Web pages in using our system + Google were significantly larger than when using only Google ($t(99) = 3.703$, $p < 0.01$). This result suggests that our system enabled the participants to find larger number of credible Web pages than Google interface within a limited time.

Table 4. Mean and standard deviation of number of selected credible Web pages within three minutes.

	Interface type	
	Google	**Our system + Google**
Mean	4.07	5.21
SD	2.44	3.13

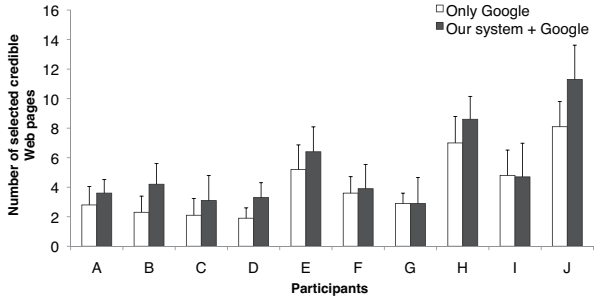

Figure 5. Mean number of credible Web pages each participant selected from a list of search results within three minutes.

Figure 5 presents results of the same analysis for each participant. According to this figure, in the results for participants F, G, and I, only Google and our system + Google do not differ significantly, but most participants could use our system + Google to find larger number of credible Web pages within a limited time than when using only Google. From the results of Table 4 and Figure 5, we found that our system can help users to find credible Web pages from a list of Web search results more efficiently or more quickly than conventional Web search interfaces.

We analyzed credibility feedback logs to examine how the participants used our system to select credible Web pages from a list of Web search results. In the experiment, the ten participants sent credibility feedbacks a total of 121 times, meaning the average per participant was 12.1. Participant A did not re-rank Web search results by sending credibility feedback. We analyzed the credibility judgment model of each participant through his/her credibility feedback logs. Table 5 presents the results. In the table, each number represents the weight of each factor for credibility judgment. If a specific factor's weight is high on the credibility judgment model of a participant, we presume the participant considered the factor as important to judge the credibility of Web search results. The table shows the factors important to judge the credibility of Web search results vary among participants. For example, the important factor for participants B and C was social reputation, while those of participant G and J were update frequency and freshness of Web pages.

Like participant A, sometimes users might be able to find credible Web page efficiently from a list of Web search re-

Table 5. Predicted credibility judgment model of each participant. SR, RI, CT, TC, CF, and UF mean social reputation, referential importance, content typicality, topic coverage, content freshness, and update frequency, respectively. Bold numbers are the most important aspects for each participant. Participant A sent no credibility feedback.

Pariticapnt	SR	RI	CT	TC	CF	UF
B	**0.1004**	0.0636	0.0601	0.0630	0.0514	0.0477
C	**0.0929**	0.0653	0.0466	0.0631	0.0543	0.0605
D	0.0650	**0.0781**	**0.0722**	0.0588	0.0545	0.0609
E	0.0641	**0.0843**	0.0651	0.0499	0.0587	0.0669
F	0.0609	**0.0747**	0.0262	0.0187	0.0384	0.0447
G	0.0509	0.0675	0.0561	0.0547	**0.0772**	**0.0875**
H	**0.0767**	0.0555	0.0651	**0.0777**	0.0654	0.0524
I	**0.0885**	**0.0700**	0.0484	0.0441	0.0659	0.0651
J	0.0530	0.0535	0.0421	0.0489	**0.1068**	**0.0927**

sults by only checking credibility radar charts of Web search results without using the re-ranking function. However, it is difficult sometimes to find many credible Web pages in a visible list of Web search results by only using the information on radar charts. In such cases, the re-ranking function becomes more important. We checked the participants who sent less feedbacks to the system than the average (12.1): participants F(4), G(5), and I(7). We know from Figure 5 that they found almost the same number of credible Web pages within the limited time using either only Google or our system + Google. From this, we think that users can find credible Web pages from the vast Web more quickly and efficiently by checking credibility radar charts and using the re-ranking function through credibility feedbacks.

DISCUSSION

We found from experiment #1 and #2 the following results:

- If users are familiar with search topics, the additional information about scores of Web search results on credibility factors as radar charts is useful for them to consistently make a black-and-white judgment about their credibility. On the other hand, if users are not familiar with the topics, users often cautiously judge the credibility of Web search results as "unknown" by checking the additional information on the Web search results.
- Our system enables users to find credible Web pages from a list of Web search results more efficiently than conventional Web search interfaces.

These results suggest that our system can be more useful for users — particularly those familiar with search topics — to search for credible Web pages in a list of Web search results from the viewpoint of efficiency than conventional Web search interfaces.

Our system is also useful for users who are not familiar with search topics to search for credible Web pages. However, the results of experiment #1 indicate they cannot judge the credibility of Web pages as clearly as users who are familiar with the topics. In real life, users often search for Web pages to learn about topics with which they are not familiar. Therefore it is very important to help such users judge the credibility of Web search results more clearly and obtain credible Web search results more efficiently.

According to ELM theory proposed by Petty et al. [18], if people have enough knowledge to understand information about a topic, they often pay attention to it. On the other hand, if people do not have enough knowledge to understand information about the topic, they often determine whether to accept the information or not with poor judgment. Considering this ELM theory and the results of the experiments, unknowledgeable users can make a consistent but poor credibility judgment on Web search results about search topics. In the worst case, they can be misled, if their credibility criteria are different from those of knowledgeable users. To avoid this, we need to think about providing help such as recommending knowledgeable users' credibility judgment models to unknowledgeable users so that unknowledgeable users can judge the credibility of Web search results more appropriately and obtain credible Web pages.

CONCLUSION AND FUTURE WORK

In this paper, we proposed the system to help users judge the credibility of Web search results and find credible Web pages from a list of Web search results. The proposed system has the following three functions for credibility-oriented Web searches: (1) visualization of score of Web search results on each of credibility factors, (2) prediction of users' credibility judgment model through users' credibility feedback for Web search results, and (3) re-ranking of Web search results according to predicted users' credibility judgment models.

We conducted two experiments to evaluate our system. Experimental results shows that our system is more useful for users — particularly users who are familiar with search topics — to efficiently find credible Web pages from a list of Web search results than conventional Web search interfaces. However, we need to consider about some issues. First, we need larger-scale experiments for more rigorous evaluation of our system. Then we also need controlled experiments where images in Web pages are eliminated, because visual designs of Web pages often affect participants' quality judgment as Zheng et al. pointed out [29]. The second issue is about making it easier for users to provide credibility feedbacks for Web pages. As in the case of participant A in experiment #2, some users are unwilling to explicitly send credibility feedbacks. Therefore it is important to think about designing better interfaces to promote users' credibility assessments or implicit credibility feedbacks for Web pages.

For users who are not familiar with search topics, we have to think about additional help for more appropriate credibility judgments of Web search results. As one possible solution, we plan to study recommending credibility judgment models of users who are familiar with the specific topics by analyzing other users' credibility feedbacks. We also think search technology for evidence information is also important as well as analysis of Web pages on the main credibility factors, so that users can make a final judgment of Web page's credibility.

To safely and efficiently obtain Web information from the vast Web, search systems focusing on credibility will be-

come more important in the future, as well as conventional relevance-oriented and popularity-oriented ones. We believe our proposed system can contribute to credibility-oriented Web searches.

ACKNOWLEDGEMENTS

This work was supported in part by the following projects and institutions: Grants-in-Aid for Scientific Research (No. 18049041) from MEXT of Japan, a Kyoto University GCOE Program entitled "Informatics Education and Research for Knowledge-Circulating Society," the National Institute of Information and Communications Technology, Japan, and Grants-in-Aid for Scientific Research (No. 09J01243) from JSPS.

REFERENCES

1. B. T. Adler and L. de Alfaro. A Content-Driven Reputation System for the Wikipedia. In *Proceedings of the 16th international conference on World Wide Web (WWW 2007)*, pages 261–270, 2007.
2. E. Agichtein, C. Castillo, D. Donato, A. Gionis, and G. Mishne. Finding High-Quality Content in Social Media. In *Proceedings of the international conference on Web search and web data mining (WSDM 2008)*, pages 183–194, 2008.
3. J. Cho, S. Roy, and R. E. Adams. Page Quality: in Search of an Unbiased Web Ranking. In *Proceedings of the 2005 ACM SIGMOD international conference on Management of data (SIGMOD 2005)*, pages 551–562, 2005.
4. P. Denning, J. Horning, D. Parnas, and L. Weinstein. Wikipedia Risks. *Communication of ACM*, 48(12):152–152, 2005.
5. M. Eisend. Source Credibility Dimensions in Marketing Communication—A Generalized Solution. *Journal of Empirical Generalisations in Marketing Science*, 10(2):1–33, 2006.
6. R. Ennals, B. Trushkowsky, and J. M. Agosta. Highlighting Disputed Claims on the Web. In *Proceedings of the 19th international conference on World wide web (WWW 2010)*, pages 341–350, 2010.
7. G. Erkan and D. Radev. LexRank: Graph-based lexical centrality as salience in text summarization. *Journal of Artificial Intelligence Research*, 22(2004):457–479, 2004.
8. B. J. Fogg, J. Marshall, O. Laraki, A. Osipovich, C. Varma, N. Fang, J. Paul, A. Rangnekar, J. Shon, P. Swani, and M. Treinen. What Makes Web Sites Credible? A Report on a Large Quantitative Study. In *Proceedings of the SIGCHI conference on Human factors in computing systems (CHI 2001)*, pages 61–68, 2001.
9. B. J. Fogg and H. Tseng. The Elements of Computer Credibility. In *Proceedings of the SIGCHI conference on Human factors in computing systems (CHI 1999)*, pages 80–87, 1999.
10. R. Guha, R. Kumar, P. Raghavan, and A. Tomkins. Propagation of Trust and Distrust. In *Proceedings of the 13th international conference on World Wide Web (WWW 2004)*, pages 403–412, 2004.
11. C. Hovland and W. Weiss. The Influence of Source Credibility on Communication Effectiveness. *Public Opinion Quarterly*, 15(4):635–650, 1951.
12. Y. Liu, B. Gao, T.-Y. Liu, Y. Zhang, Z. Ma, S. He, and H. Li. BrowseRank: Letting Web Users Vote for Page Importance. In *Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval (SIGIR 2008)*, pages 451–458, 2008.
13. M. Meola. Chucking the Checklist: A Contextual Approach to Teaching Undergraduates Web-Site Evaluation. *portal: Libraries and the Academy*, 4(3):331–344, 2004.
14. M. Metzger, A. Flanagin, K. Eyal, D. Lemus, and R. McCann. Credibility for the 21st century: Integrating perspectives on source, message, and media credibility in the contemporary media environment. *Communication yearbook*, 27:293–336, 2003.
15. M. J. Metzger, A. J. Flanagin, and L. Zwarun. College student Web use, perceptions of information credibility, and verification behavior. *Computer & Education*, 41(3):271–290, 2003.
16. S. Nakamura, S. Konishi, A. Jatowt, H. Ohshima, H. Kondo, T. Tezuka, S. Oyama, and K. Tanaka. Trustworthiness Analysis of Web Search Results. In *Proceedings of the 11th European Conference on Research and Advanced Technology for Digital Libraries (ECDL 2007)*, pages 38–49, 2007.
17. M. Nakatani, A. Jatowt, H. Ohshima, and K. Tanaka. Quality Evaluation of Search Results by Typicality and Speciality of Terms Extracted from Wikipedia. In *Proceedings of the 14th International Conference on Database Systems for Advanced Applications (DASFAA 2009)*, pages 570–584, 2009.
18. R. Petty and J. Cacioppo. The elaboration likelihood model of persuasion. *Advances in experimental social psychology*, 19(1):123–205, 1986.
19. P. Pirolli, E. Wollny, and B. Suh. So You Know You're Getting the Best Possible Information: A Tool that Increases Wikipedia Credibility. In *Proceedings of the 27th international conference on Human factors in computing systems (CHI 2009)*, pages 1505–1508, 2009.
20. S. Rieh and D. Danielson. Credibility: A Multidisciplinary Framework. *Annual Review of Information Science and Technology*, 41(1):307–364, 2007.
21. S. Y. Rieh. Judgement of Information Quality and Cognitive Authority in the Web. *Journal of the American Society for Information Science and Technology*, 53(2):145–161, 2002.
22. G. Salton and C. Buckley. Improving Retrieval Performance by Relevance Feedback. *Journal of the American society for information science*, 41(4):288–297, 1990.
23. E. Sillence, P. Briggs, L. Fishwick, and P. Harris. Trust and Mistrust of Online Health Sites. In *Proceedings of the SIGCHI conference on Human factors in computing systems (CHI 2004)*, pages 663–670, 2004.
24. J. Stiff and P. Mongeau. *Persuasive communication*. Guilford Publications, 2002.
25. B. Suh, E. H. Chi, A. Kittur, and B. A. Pendleton. Lifting the Veil: Improving Accountability and Social Transparency in Wikipedia with WikiDashBoard. In *Proceeding of the twenty-sixth annual SIGCHI conference on Human factors in computing systems (CHI 2008)*, pages 1037–1040, 2008.
26. M. A. Suryanto, E. P. Lim, A. Sun, and R. H. L. Chiang. Quality-Aware Collaborative Question Answering: Methods and Evaluation. In *Proceedings of the Second ACM International Conference on Web Search and Data Mining (WSDM 2009)*, pages 142–151, 2009.
27. J. Teevan, S. T. Dumais, and E. Horvitz. Personalizing Search via Automated Analysis of Interests and Activities. In *Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval (SIGIR 2005)*, 2005.
28. B.-Q. Vuong, E.-P. Lim, A. Sun, M.-T. Le, and H. W. Lauw. On Ranking Controversies in Wikipedia: Models and Evaluation. In *Proceedings of the international conference on Web search and web data mining (WSDM 2008)*, pages 171–182, 2008.
29. X. S. Zheng, I. Chakraborty, J. J.-W. Lin, and R. Rauschenberger. Correlating Low-Level Image Statistics with Users' Rapid Aesthetic and Affective Judgments of Web Pages. In *Proceedings of the 27th international conference on Human factors in computing systems (CHI 2009)*, pages 1–10, 2009.