Improving Credibility Evaluation of Search Results by Analyzing Historical Content and Traffic Data

Adam JATOWT † Yukiko KAWAI ‡ Satoshi NAKAMURA † and Katsumi TANAKA †

† Graduate School of Informatics, Kyoto University Yoshida-Honmachi, Sakyo-ku
606-8501 Kyoto, Japan
‡ Kyoto Sangyo University Motoyama, Kamigamo, Kita-Ku
603-8555 Kyoto, Japan
E-mail: † {adam,nakamura,tanaka}@dl.kuis.kyoto-u.ac.jp, ‡ kawai@cc.kyoto-su.ac.jp

Abstract In this paper we propose improving credibility evaluation of Web search results by displaying information on historical traffic rates and on the past content of Web pages extracted from online Web archives. Data about visit rates is used for estimating trends in popularity of Web sites over time and for determining the overall popularity of searched topics. On the other hand, historical content of pages allows for determining long-term topics of pages and for estimating the age of their content elements. In addition, other important characteristics of pages are calculated such as changes in sentiment or commercial intent over time. In this demonstration we show the proof-of-concept system that displays the above measurements as additional information about returned Web search results.

Keyword Web content credibility, Web archive, traffic data, page history

1. Introduction

Conventional search engines do not provide sufficient information on the credibility of search results. Typically, search results contain only titles, URLs and extracted snippets of pages for showing pages' relevance to issued queries. Although it is believed that search engines use some kind of popularity measures (i.e., PageRank [4]) for constructing search results, usually, no absolute values of these popularity measures are provided for users. Thus users are left alone to judge how popular and also how credible a particular search result is. This is however usually not trivial [3]. Given that there is lots of unreliable information on the Web, we think that search engines should provide additional data on pages for users to be able to better judge their credibility and accuracy.

In this paper we explore the potential of using traffic data and historical content for evaluating page credibility. The former allows capturing relative popularity levels of search results and their trends. In reality, top search results sometimes contain highly relevant yet rather unpopular pages. We believe that users should be informed about the popularity of search results among Web surfers before they can decide whether or not to trust their content. It is well-known that people tend to trust recommendations of others and often follow the actions and choices of the majority. On the other hand, the historical content of pages can be used to generate additional indicators for page credibility evaluation beyond the ones based on the current page content. We propose the following measures based on the analysis of the past content of pages: a) age of page content, b) top terms characterizing page long-term topics, c) page sentiment level over time and d) page commercial intent over time.

We need to emphasize that the proposed measures apply only to certain aspects of credibility evaluation of Web pages, which is actually a complex and multi-dimensional process. In order to completely evaluate trustworthiness of search results one would need to examine many other factors [3,5]. Nevertheless, we believe that the proposed measurements should help users more accurately determine level to which they can trust particular search results.

2. Credibility evaluation based on traffic data

We present here two cases in order to demonstrate the usefulness of traffic data for credibility evaluation. Figure 1 shows two cases of search results presented together with their current popularity levels. In each case four search results are ordered by their relevance scores. In case 1 the current popularity levels of pages decrease linearly along with the decrease of page relevance. In this case, it is easy to imagine the trust levels users should put into the pages. However, in case 2 we observe a drastic drop in the popularity levels of pages 3 and 4. The levels of trust put to pages 3 and 4 should be now different than the corresponding ones in case 1. If users did not have any

information on the traffic rate of search results, then they would treat equally and put the same trust levels to the pages in both case 1 and 2.

Consider another situation in Figure 2. Here search results are associated with the traffic data measured over some predefined time spans. This allows for estimating popularity trends of search results. We can see that now the actual trust that users may put into the pages should differ from the one in case 1 of Figure 1. Since pages 1 and 2 actually have falling trends of traffic, while pages 3 and 4 have rising ones, one cannot surely claim that the latter search results are more trustworthy than the former ones. We think that displaying information on the current popularity levels of pages as well as their trends should increase the capability of users to evaluate page content from the viewpoint of its credibility.

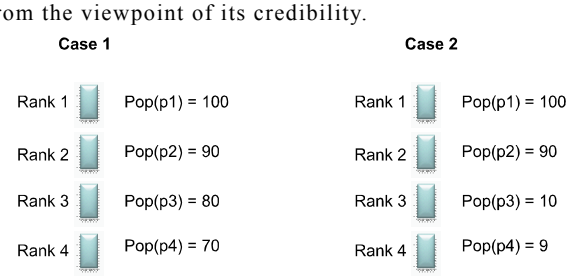

Figure 1 Two cases of search results' presentation with the information on their current popularity.

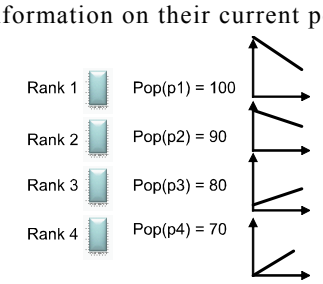

Figure 2 Case of search results' presentation with the information on trend of their popularity.

In our implementation we retrieve traffic data of web sites to which pages from search results belong. This data is obtained from external sources like Alexa[1] or Compete[2]. Such sources track the number of visitors by distributing toolbars to users and gathering information on their visited sites in real time. Note that the above sources usually provide only traffic rates to whole sites. Data on traffic rates to individual pages within sites could be however approximated based on site structures.

In our implementation, graphs showing visit popularity of sites over time are displayed below search results (Figure 6 in Appendix). In addition, trend lines are calculated by applying the linear regression and shown next to the traffic data graphs. We also calculate a unified popularity score for each site based on the current and past visit rates (Equation 1).

$$S^{pop} = \alpha * r_M + \beta * mean_{1<i<M}(r_i) + \varepsilon * (r_M - r_{M-1}) \qquad 1$$

Equation 1 is derived from the concept of proportional–integral–derivative (PID) controllers. The first part of the equation depends on the current popularity rate r_M of the page (proportional part); the second one is based on the average traffic rate over time (integral part), while the last one depends on the amount of the recent change in the traffic rate (derivative part). For calculating Equation 1 we use M values of traffic rates of sites included in search results with a default granularity level of 3 months. Similar solution was also used for computing user trust levels in online communities [1].

By adjusting α, β and ε users can choose the degree to which each part of Equation 1 influences the final popularity score of pages (see the top part of Figure 6 in Appendix). In addition, the trend and popularity score of the aggregated traffic rate over N top sites containing search results are shown as a reference at the top. Those search results that have slopes of traffic trend lines or popularity scores lower than 2-3 standard deviations from the mean values will be indicated in red background color for alerting users.

3. Credibility evaluation based on page history

In this section we discuss additional measures for improving the credibility evaluation of search results based on page historical content. Past page snapshots are obtained from the Internet Archive[3], which contains about 2PB of data crawled since 1996. Our system retrieves M past snapshots of pages from search results with a default granularity of 3 months. Based on the downloaded snapshots we then calculate the following measures:

- age of page content
- top terms characterizing page content over time
- page sentiment level over time
- page commercial intent over time

[1]Alexa search engine: http://www.alexa.com/

[2]Compete: http://www.compete.com/

[3]Internet Archive: http://www.archive.com/

Similarly to the case of displaying the traffic rate, search results are indicated with a red background color if the page sentiment or commercial intent levels differ 2-3 standard deviations from the agglomerated measures over all *N* search results.

3.1 Content age

The age of page content is calculated by sequentially comparing consecutive snapshots of pages retrieved from the Internet Archive. The creation date of a given content element is approximated as the timestamp of the oldest snapshot that contains this element. Content is divided according to DOM structure of pages with the lowest HTML nodes determining single content elements. The system analyzes the occurrence of the same HTML nodes in consecutive past snapshots of pages starting from the latest one until it finds the age of all content elements published on the current page version (Figure 3). In order to speed up the whole process a modified binary search algorithm may be used. More details on the age detection method are provided in [2].

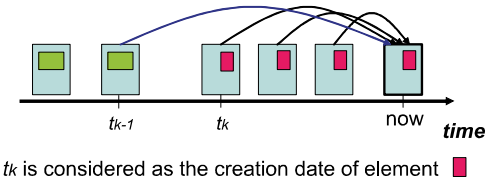

Figure 3 Calculation of content age by sequentially comparing past page snapshots with the present page version.

In Figures 4 and 5, we show examples of pages annotated with creation dates of content elements using a prototype system. Content elements are framed by a red line containing approximate content creation date attached to the left-down corner. In addition, the average age of page content and the age of the whole page are shown. The former is calculated by averaging dates of particular content elements on a page with weights depending on the size of the area that they occupy. The latter is taken as the timestamp of the oldest page snapshot available in the Internet Archive.

Page shown in Figure 4 is about Tim Berners-Lee, the inventor of the Web. Each content element on the page including the photo has been created before May 2000. Clearly, considering that there has been no update since that time, the page features a high risk of providing obsolete or untrue information. Page in Figure 5 is also old (05/06/1997 as the page age), however, it is more up-to-date (last modification date 20/03/2007). For example, the photo of Tim Berners-Lee has been inserted around 02/02/2004.

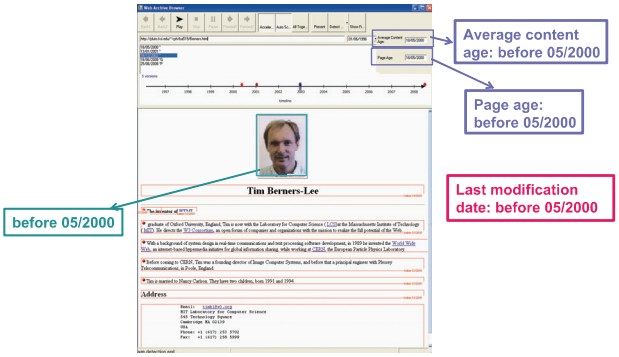

Figure 4 Example of content annotation with approximate age of content elements for a page having a high probability of being obsolete.

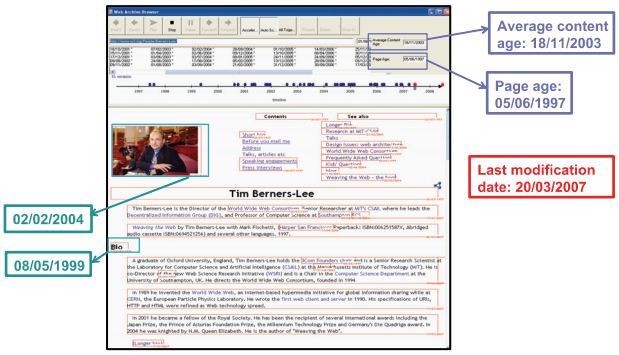

Figure 5 Example of content annotation with approximate age of content elements for rather up-to-date page.

3.2 Top terms characterizing typical page content

Using retrieved page snapshots we detect the most common terms appearing on pages over time in order to shed light on their typical content and themes. This should provide users with the means to evaluate the long-term relevance of pages for the purpose of facilitating detection of highly authoritative sources.

The calculation method is as follows. First, the content of the retrieved snapshots of a given search result is stripped from HTML and other markup. Next, common terms are eliminated via a stop word list. Then the remaining terms are scored according to the following equation.

$$S^{j} = \frac{1}{M} * \sum_{i=1}^{i=M} \frac{c_i^{j}}{c_i} \log\left(\frac{N}{DF_i^{j}}\right) \qquad 2$$

Equation 2 is derived from a well-known *tf*idf* scoring scheme used in information retrieval [6]. c^j_i is the number of instances of a term j in a page snapshot i, while c_i is the total number of terms in this snapshot. N is the total number of evaluated search results and DF^j_i is the number of documents whose snapshots at time point i contain term j.

Figure 7 in Appendix shows top 5 terms for example search results.

3.3 Page sentiment level over time

Sentiment level is another metric for evaluating page credibility. The main assumptions here are that pages with relatively high total sentiment levels may lack objectivity and pages with frequently changing sentiment levels may be written by authors with unstable opinions on the same topic. Total sentiment levels of pages are calculated using Equation 3, in which c^{ps}_i is the number of positive sentiment terms and c^{ns}_i is the number of negative sentiment terms in a page snapshot i. We use here an external sentiment dictionary containing both positive and negative sentiment words.

$$S^{\text{int}} = \frac{1}{M} * \sum_{i=1}^{i=M} \frac{c_i^{ps} + c_i^{ns}}{c_i} \qquad 3$$

Equation 4 is used for measuring the volatility of sentiment levels of pages over time.

$$S^{vol} = \frac{1}{M} * \sum_{i=1}^{i=M} \left(\frac{c_i^{ps} - c_i^{ns}}{c_i} - \frac{1}{M} * \sum_{i=1}^{i=M} \frac{c_i^{ps} - c_i^{ns}}{c_i} \right)^2 \qquad 4$$

Sentiment levels at single time points $(c^{ps}_i - c^{ns}_i)/ c_i$ are displayed on a graph under search results and, in addition, the total sentiment scores and their volatility over time are shown (Figure 7 in Appendix).

3.4 Page commercial intent

We propose page commercial intent as the last credibility evaluation metric discussed in this paper. The main assumption here is that pages with commercial intent may lack objectivity and provide inaccurate or unreliable information as their main objective is to sell products or services rather than to provide information.

Commercial intent at time point i is calculated as the aggregated frequency of terms expressing product selling or service providing intent in the page snapshot retrieved at i. We have manually prepared a dictionary of 339 terms expressing commercial intent that are commonly encountered on the Web (e.g., product, shipment, payment, credit card, etc.). The system displays the levels of commercial intent over time on a graph as well as the overall commercial intent score of each page calculated as the average commercial intent of its retrieved past snapshots.

4. Conclusions

Credibility of Web content is an important factor determining its usefulness. However, conventional search engines do not sufficiently approach the problem of trustworthiness evaluation of Web search results.

In this paper we proposed using page traffic data and historical content for improving credibility evaluation of search results. We have introduced several credibility indicators and demonstrated the prototype application for supporting users in evaluating the credibility of Web search results. The usefulness of the proposed credibility metrics for different kinds of Web pages has to be however tested in user studies. This forms a part of our future work.

One needs to remember that the complete credibility evaluation of Web pages is a multi-dimensional process and should incorporate various different metrics. Thus the metrics that we showed apply only to certain aspects of credibility of Web search results.

Acknowledgements

This research was supported in part by the National Institute of Information and Communications Technology, Japan, by the MEXT Grant-in-Aid for Scientific Research in Priority Areas entitled: Content Fusion and Seamless Search for Information Explosion (#18049041, Representative: Katsumi Tanaka), by the Kyoto University Global COE Program: Informatics Education and Research Center for Knowledge-Circulating Society (Representative: Katsumi Tanaka) and by the MEXT Grant-in-Aid for Young Scientists B (#18700111, Representative: Adam Jatowt; #18700110, Representative: Yukiko Kawai)

Publications

[1] Caverlee, J. Liu, L., and Webb, S. “Socialtrust: tamper-resilient trust establishment in online communities.” Proceedings of the 8th ACM/IEEE-CS Joint Conference on Digital Libraries, pp. 104-114, 2008.

[2] Jatowt, A., Kawai, Y. and Tanaka, K. “Detecting age of page content.” Proceedings of the 8th International Workshop on Web Information and Data Management, pp. 137–144, 2007.

[3] Nakamura, S., Konishi, S., Jatowt, A., Ohshima, H., Kondo, H., Oyama, S. and Tanaka, K.

"Trustworthiness Analysis of Web Search Results," Proceedings of the 11th European Conference on Research and Advanced technology for Digital Libraries, Springer LNCS 4675, Budapest, Hungary, pp. 38-49, 2007.

[4] Page, L., Brin, S., Motwani, R., and Winograd. T. "The pagerank citation ranking: Bringing order to the Web." Technical report, Stanford Digital Library Technologies Project, 1998.

[5] Rieh, S. Y. and Danielson, D. R. "Credibility: A multidisciplinary framework." In B. Cronin (Ed.), Annual Review of Information Science and Technology (Vol. 41, pp. 307-364). Medford, NJ: Information Today, 2007.

[6] Salton, G. and Buckley, C. "Term-weighting approaches in automatic text retrieval." Information Processing and Management: an International Journal, 24:5, pp. 513-523, 1988.

Appendix

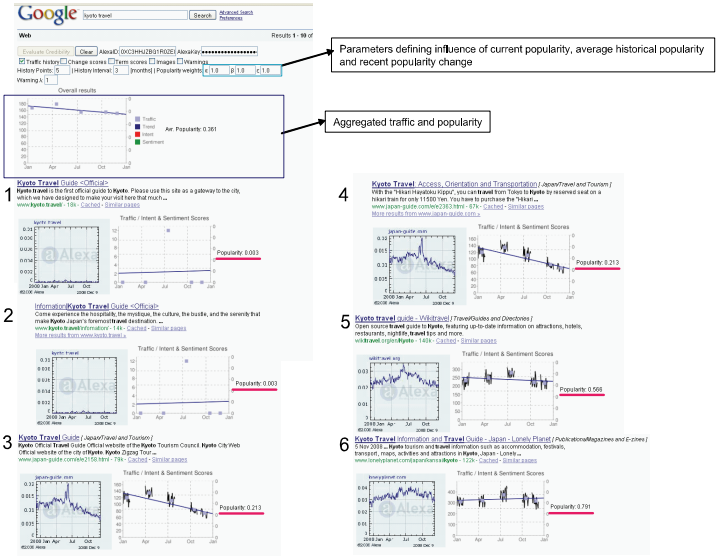

Figure 6 Presentation of search results with displayed traffic rates and trends over time as well as their popularity scores.

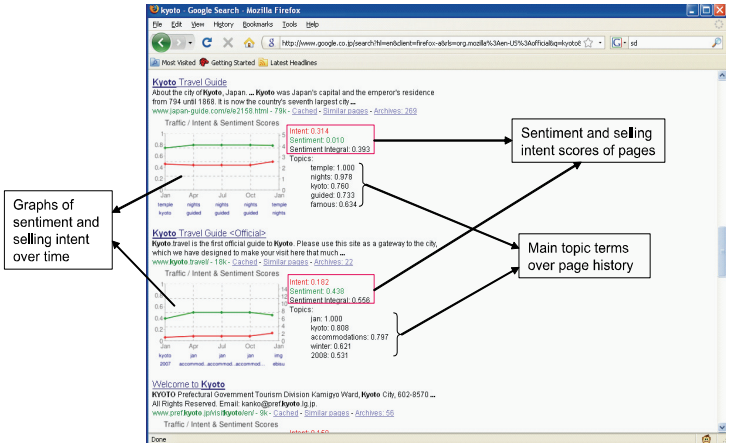

Figure 3 Examples of search results with displayed top terms over time, sentiment and commercial intent scores.