

Augmenting Web Pages and Search Results to Support Credibility Assessment

Julia Schwarz
HCII, Carnegie Mellon
Pittsburgh, PA, USA
julia.schwarz@cs.cmu.edu

Meredith Ringel Morris
Microsoft Research
Redmond, WA, USA
merrie@microsoft.com

ABSTRACT

The presence (and, sometimes, prominence) of incorrect and misleading content on the Web can have serious consequences for people who increasingly rely on the internet as their information source for topics such as health, politics, and financial advice. In this paper, we identify and collect several page features (such as popularity among specialized user groups) that are currently difficult or impossible for end-users to assess, yet provide valuable signals regarding credibility. We then present visualizations designed to augment search results and Web pages with the most promising of these features. Our lab evaluation finds that our augmented search results are particularly effective at increasing the accuracy of users' credibility assessments, highlighting the potential of data aggregation and simple interventions to help people make more informed decisions as they search for information online.

Author Keywords

Credibility, Trustworthiness, Web

ACM Classification Keywords

H.5.4 Information Interfaces and Presentation – Hypertext/Hypermedia: *User issues*.

General Terms

Human Factors.

INTRODUCTION

The internet is increasingly becoming a primary source of information for people around the world. While there is a great deal of useful information online, misleading and incorrect Web pages continue to proliferate. Assessing the *credibility* of Web pages is therefore becoming an increasingly important aspect of information literacy [25], albeit one that many end-users struggle with [17].

The difficulty of assessing Web sites' credibility manifests itself in several problematic phenomena. For instance, providing account information to malicious sites masquerading as authentic ones, as in phishing attacks, results in the

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.

CHI 2011, May 7–12, 2011, Vancouver, BC, Canada.

Copyright 2011 ACM 978-1-4503-0267-8/11/05...\$10.00.

loss of billions of dollars annually [7], despite the integration of phishing toolbars into mainstream browsers [33]. The presence of misleading, questionable, and factually incorrect information on the Web is yet another source of concern. For instance, misinformation campaigns portraying U.S. President Barack Obama as a member of the Muslim faith have resulted in substantial confusion among American voters [29]. This latter issue of incorrect information is the focus of this paper. *Non-credible* Web pages of this type can have serious consequences when people use information found online as the basis for decisions in critical domains such as politics, finance, and health.

Naïve user populations, such as school-age children, may be particularly at risk for being misled by such content; hence, credibility assessment is considered a high-priority topic by educators. Jenkins [21] names Web search as one of several key “new media literacies” for students, and identifies *judgment* (“the ability to evaluate the reliability and credibility of different information sources”) as a key part of that process. Leu and Zawilinski [25] note that “online reading” has altered the meaning of literacy to include Web-related skills such as discerning credibility. This challenge extends beyond childhood – even college-educated adults tend to visit non-credible Web pages when searching, because of a tendency to conflate high ranking in search result lists with credibility [17].

In this paper we present visualizations to augment search results and Web pages in order to help people more accurately judge the credibility of online content. After first discussing our creation of a public dataset of credibility-labeled URLs, we identify several page and site-specific features that are currently difficult or impossible for end-users to assess, and quantify their relationship with credibility. We then present two visualizations designed to assist users in their credibility judgments. Finally, we present findings from a user study that evaluates our visualizations' effectiveness in increasing credibility assessment accuracy. Our findings suggest that augmenting search results with information about expert user behavior is a particularly effective means of enhancing users' credibility judgments, resulting in a doubling of judgment accuracy.

RELATED WORK

Due to the dire economic consequences of phishing scams and malware sites, a great deal of effort has gone into in-

creasing end-users' awareness of insecure Web sites that steal users' personal information or spread computer viruses (e.g., [7, 33]). In this paper, we consider the issue of *credibility* as separate from *security* (other researchers, such as Lazar et al. [24], make a similar distinction); hence, we do not focus on sites that actively perpetrate criminal activities, but rather on sites that contain questionable, misleading, or factually incorrect information. Such non-credible sites may be particularly problematic for naïve Web consumers such as children and teens [21, 25], as well as for adults seeking critical information in unfamiliar domains.

Prior research on Web credibility includes research on understanding users' mental models when assessing credibility and on the development and evaluation of interventions to help people better judge credibility online.

Mental Models

Organizations put considerable effort into appearing credible to gain customers' trust. The field of captology [12] studies how technology can be designed to persuade end-users. Much prior work in the area of credibility approaches the topic from a captology perspective, with a goal of understanding how people evaluate credibility so as to help Webmasters and designers create sites that will appear more credible (regardless of their true information quality), e.g., Schneiderman's guidelines for designing trust online [30] and Ivory and Hearst's tool for high quality site design [19].

This line of prior research has shown that users consider many different pieces of information to help them evaluate the credibility of Web pages. Fogg categorizes this information into four types of credibility [12]:

1. *Presumed credibility* is based on general assumptions in the users' mind (e.g., the trustworthiness of domain identifiers like *.gov*).
2. *Surface credibility* is derived from inspection of a site, is often based on a first impression that a user has of a site, and is often influenced by how professional the site's design appears.
3. *Earned credibility* refers to trust established over time, and is often influenced by a site's ease of use and its ability to consistently provide trustworthy information.
4. *Reputed credibility* refers to third party opinions of the site, such as any certificates or awards the site has won.

Fogg et al. conducted large-scale studies to determine what factors people use to evaluate credibility [11, 14], ultimately determining that in practice the "look and feel" of a site has the greatest impact on users' credibility assessments. McKnight and Kacmar [27] also found that professionalism of site design heavily influences credibility perceptions.

Hargittai et al. [17] found that search result ranking is often interpreted by end-users as a key credibility indicator, despite that fact that search engine rankings primarily reflect keyword relevance, not factual correctness, and are influenced by factors like search engine optimization businesses [26] whose aims may be at odds with credibility. More than two-thirds of American internet users think search engines are a "fair and unbiased" information source [9].

Educational institutions, consumer watchdog groups, and libraries offer guidance to users on techniques for credibility assessment, such as directing readers to consider a Web page's accuracy, authority, objectivity, currency, and coverage [22]. Whether users take the time to follow such advice, assuming they are aware of it, is influenced by factors such as the nature of their task [28].

Fogg proposed Prominence-Interpretation Theory [13] to model how people assess credibility online. P-I Theory posits that the impact an element has on perceived credibility is a product of its prominence (how likely it is to be noticed) and interpretation (what value or meaning people assign to that element). Factors such as user involvement [10], user task [28], and experience affect prominence, while a user's assumptions, knowledge level, and context affect interpretation [13]. Hilligoss and Rieh [18] present a framework which identifies aspects of credibility assessment independent of media, information type, and environment. This framework identifies three levels of credibility judgment: construct, heuristics, and interaction. Similar to P-I Theory, Hilligoss and Rieh observed that people repeatedly refine their credibility judgments as they are viewing a Web page.

Interventions

In addition to understanding how people evaluate credibility, there have also been efforts toward improving people's credibility assessment accuracy. Three main approaches are content analysis, prediction, and informing end users.

Content Analysis

One approach to helping people find factually correct content is to automatically identify false facts. Extracting factual information from the Web is an active research area. An example of such work is Open Information Extraction [1]. However, natural language approaches are not yet reliable or comprehensive enough for use on the open Web.

Some researchers have applied content analysis approaches to specific aspects of credibility. For instance, Dispute Finder [8] identifies contentious topics by looking for text such as "X is disputed," and BLEWS [15] provides insight into online news by analyzing the content and sentiment of blogs referencing particular articles.

Predicting Credibility

Developing algorithms to predict the credibility of a page is another promising approach. Two of the better known algorithms, TrustRank [16] and CredibleRank [4], use the link structure of the Web to determine a credibility score. However, these algorithms define non-credible pages in terms of Web spam, not in terms of information quality.

Informing End-Users

A third approach is to show end-users information that may help them form more accurate impressions of a page's quality. Examples of this type of intervention include augmenting Wikipedia pages with a visualization of edit history [23], and creating certifications or whitelists for trustworthy sites (such as HON certification [www.hon.ch]). A challenge of this approach is designing interventions that impact user behavior; for example, studies of anti-phishing browser

toolbars have found them to be ineffective because people don't notice or pay attention to the toolbar [33].

Our present work follows the “informing end users” approach. Since content-analysis and prediction approaches are not yet accurate enough to rely on for credibility analyses, this approach of supporting a user's reasoning process avoids the risk of incorrect classifications by automated techniques. This approach also permits more nuanced interpretations of credibility (e.g., situations where there is not a single accepted truth, or in which quality judgments may differ based on cultural norms). Another potential benefit of this approach is that informing end-users might encourage reflection on credibility, potentially enhancing users' credibility assessment skills over the long-term.

DATA SET

Creating and testing interventions for improving users' credibility assessments requires a data set of Web pages that have been labeled with a “ground truth” credibility score. To our knowledge, no such data set is publicly available. The following sub-sections describe our process for creating such a data set.

Page Selection

Hand-labeling Web pages for credibility is a time-consuming process (requiring several minutes to read the contents of each page plus time to look for additional information from other sources to support the assessment in cases where the factual correctness of a page was unclear). We chose 1,000 Web pages as the size of our data set to balance the desire to have a large sample size with the time constraints of generating meaningful credibility ratings.

To balance topical breadth and depth, we selected five topics by browsing directory headings and sub-headings in the Open Directory Project [dmoz.org], with the aim of selecting topics known to have substantial amounts of both credible and non-credible coverage online. The five topics selected were Health, Politics, Finance, Environmental Science, and Celebrity News. We then used Google Zeitgeist [google.com/zeitgeist] to identify more specific information needs in each of these areas, based on recent popular search trends. Based on the Zeitgeist trends, we developed five queries within each topic area (Table 1). We then issued these queries to a popular search engine, and used the URLs for the top 40 search results for each query as the URLs to label for our data set. This resulted in a total of 1,000 URLs (5 topics x 5 queries each x 40 results each).

Credibility Ratings

We associated a five-point Likert scale rating with each URL in our data set, with a score of 1 for “very non-credible” and a score of 5 for “very credible.” Based on a synthesis of the research literature [11, 14, 18, 22, 24, 27, 28], we used the following definition to operationalize credibility for the purposes of assigning a score: *A credible webpage is one whose information one can accept as the truth without needing to look elsewhere. If one can accept information on a page as true at face value, then the page is credible; if one needs to go elsewhere to check the validity of the information on the page, then it is less credible.*

Using this definition as a guide, one author of this paper rated all of the 1,000 pages in our data set. This rater is an expert Web user who interacts with the Web and search engines on a daily basis, holds a bachelor's degree in computer science, is enrolled in a graduate degree program in a related field, and is familiar with suggested pedagogies for credibility assessment (e.g., [22]). This process took approximately 20 hours. To ensure reliability, another author of this paper, with similar background and credentials, rated 50 randomly selected URLs from the data set. A Spearman correlation test to measure inter-rater reliability indicates high agreement for this overlapping sample ($\rho(50) = 0.7$).

To further ensure reliability, we solicited additional ratings from topic experts for the subset of 21 pages ultimately used in our evaluation sessions. As discussed in the “Evaluation” section, this subset consisted of 7 pages each on the topics of Finance, Politics, and Health. For the reliability check, we recruited two experts in each of these topic areas, and gathered their credibility scores for the pages in their topic. The finance experts were both professionals working in the banking and investment industry, the political experts were both volunteers on presidential political campaigns, and the health experts were both medical doctors. Agreement was calculated by averaging the two experts' scores for each page, and then comparing these to the ground truth scores using Spearman correlation. Agreement between the average of the experts' ratings in each domain and our ratings was $\rho(7) = 0.6$ for Finance, $\rho(7) = 0.7$ for Politics, and $\rho(7) = 0.9$ for Health.

In the remainder of this paper, we refer to the scores we assigned as the credibility “ground truth” for these 1,000 Web pages. Although we used a rater with high general expertise regarding credibility and the Web, the reader should bear in mind that assigning credibility ratings is to some extent a subjective process. However, our comparisons with a sample of topical specialists' ratings indicate that our ground truth ratings are a fair approximation of expert opinion on the credibility of Web pages. We have made our labeled data set available at <http://research.microsoft.com/credibility> to enable readers to assess the quality of our labels, as well as to build upon this data for future research.

FEATURE EXPLORATION

As discussed in the “Mental Models” section, prior research on users' credibility assessment practices indicates that people typically make credibility judgments based on features such as a page's search result ranking [17] and visual design [14]. We hypothesized that, in addition to these readily apparent features that people currently rely on, there are many features which are difficult or impossible for users to reflect on when viewing a Web page, but that might help them better judge credibility. We explored several such features, each of which fell into one of three categories: on-page, off-page, and aggregate features.

Following each feature description, we report how it correlates with our ground truth labels. We used a non-

Topic	Query Terms	Expert URL Filters	# of Users
Health	Atkins diet effectiveness P90x exercise program H1N1 vaccine side effects Alzheimer's genes Autism warning signs	ncbi.nlm.nih.gov/pubmed pubmedcentral.nih.gov	254,175
Finance	Is it a good time to invest in gold? What mutual funds to invest in Reduce personal debt Mortgage refinancing Is it a good time to invest?	bloomberg.com edgar-online.com hoovers.com sec.gov	201,014
Politics	Iran election rigged Cash for clunkers eligibility Obama birthplace Death Panels Tea Party	foreignaffairs.com theatlantic.com foreignpolicy.com hir.harvard.edu economist.com	66,155
Celebrity News	Lady Gaga Adam Lambert Nadya Suleman Floyd Landis Michael Jackson	ew.com usmagazine.com people.com	692,611
Environmental Science	Renewable energy Green jobs Climate change Cap-and-trade Organic Eating	pewclimate.org epa.gov rff.org nrdc.org whitehouse.gov/administration/ceq	83,476
All Users		(none)	50,473,520

Table 1: Our data set contained 1000 URLs – 200 in each of 5 topics, consisting of the top 40 search results for each of the 5 topical query terms). The “expert URL filters” for each topic are the sites used to heuristically identify topical experts. The final column shows the number of users in one month’s worth of browser toolbar data identified as experts by these filters.

parametric correlation measure (Spearman’s ρ) since our credibility labels were subjective and not continuous.

On-Page Features

On-Page features are present on a page but are difficult or time-consuming for a person to quantify or attend to.

Spelling Errors: We computed the number of spelling errors by writing a program to screen-scrape and spell check each URL in our data set. ($\rho[1000]=0.01$)

Advertising: Our program computed the number of advertisements on a page by searching the HTML for script tags from several popular advertisers such as Google AdWords [adwords.google.com] and DoubleClick [doubleclick.com]. ($\rho[1000]=-0.20$)

Domain Type: Because users tend to focus on the page contents rather than the browser’s address bar [33], the domain type (.com, .gov, etc.) of a page may not be salient; we collected this as a categorical feature. ($\rho[1000]=0.19$)

Off-Page Features

Off-page features require the user to leave the target page and look elsewhere for supplementary data.

Awards: We collected information about what awards and certifications sites had received from three agencies. A Webby Award is “the leading international award honoring excellence on the internet” [webbyawards.com]. We counted the number of Webby Awards a site had received for the year 2009. Alexa [alexa.com] (an organization that monitors internet traffic) publishes a list of the top one million most popular sites on the internet. We used the Alexa rank of a site as another feature. Health on the Net (HON) [www.hon.ch] promotes reliable health information online,

and certifies reliable health-related Web sites. We treated the presence or absence of a HON award as a binary feature for the “Health” subset of our URLs. (Alexa rank: $\rho[1000]=-0.15$, number of Webby awards: $\rho[1000]=0.11$, HON certification: $\rho[200]=0.35$)

PageRank: A Web page’s PageRank [3] is not generally visible to end-users, unless they install browser toolbars or look up URLs on dedicated sites such as those run by search engine optimization companies [26]. We gathered the PageRank of each URL in our data set. When the PageRank for a particular page was unavailable, we used the rank of its parent site. A related feature we also gathered was a major search engine’s ranking of the URL for the queries used in generating our data set. ($\rho[1000]=0.30$)

Sharing: We also gathered information about how frequently a URL was shared using publicly available sharing and click information from Bit.ly [bit.ly]. We obtained the number of times a link to a Webpage was shared, liked, commented on, and clicked from Facebook, as well as the number of times a shortened version of the URL (commonly used when sharing links on Twitter) was clicked. We also counted the number of users that bookmarked a URL [del.icio.us]. (bookmarks: $\rho[1000]=0.22$, bit.ly clicks: $\rho[1000]=0.17$, Facebook shares: $\rho[1000]=0.29$)

Aggregate Features

Aggregate features are not generally available to end users, though they could be made available by companies such as search engines who often log user behavior via browser toolbars. Our aggregate features were computed from one month’s worth of anonymous browsing data (June 2010)

from 50,473,520 users who opted to provide data via Microsoft's browser toolbar.

General Popularity: To calculate overall popularity, we counted the number of unique user IDs visiting the page in the time period covered by our logs. ($\rho[1000]=0.38$)

Geographic Reach: To approximate the popularity of a page among a broad demographic, we computed the number of different geographic locations visitors to the site originated from using zip code information. ($\rho[1000]=0.32$)

Dwell Time: We computed the average length of time users kept a URL open in their browser as a proxy for the amount of time spent viewing a page. ($\rho[1000]=0.001$)

Revisitation Patterns: Returning to a page can be considered an implicit vote for its quality. We calculated on average how much each page was re-visited. ($\rho[1000]=0.36$)

Expert Popularity: Fogg's classification of credibility into presumed, surface, earned, and reputed [12] suggests that not everybody is able to evaluate credibility equally well. For example, people unfamiliar with a topic, such as medicine, have little opportunity to evaluate the *earned credibility* of a given medical site because they have spent little time on medically-related websites. Indeed, research comparing search strategies of experts and non-experts in topics such as health reveal that topic experts are more effective searchers in their expertise area because they use previously-encountered, high-quality URLs as starting points in the information-seeking process [2]. Therefore, behavior of experts within a particular domain may provide an especially useful source of information regarding credibility.

Based on the approach described by White et al. [31], we classified each of the fifty million user IDs in our log data with regards to expertise in the five topic areas in our URL data set. White et al. [31] found that a heuristic-based approach to defining expertise (users who visit a set of whitelisted URLs identified by a professional in the target topic area) is effective at differentiating users according to several standards of expert behavior used in the information retrieval community (e.g., [2, 32]). Using this approach with whitelists from [31] (health, finance) and discussions with topic professionals and enthusiasts (politics, environmental science, and celebrity news), we labeled users who visited any of the whitelisted sites in a particular topic area more than ten times to be "experts" in that topic (Table 1). Using this metric, we then calculated the number of topic experts who had visited each page in our data set in the period covered by our log data. (health: $\rho[200]=0.5$, politics: $\rho[200]=0.4$, environment/celebrities/finance: $\rho[200]=0.3$)

Feature Selection

While showing users all the features we gathered for each page would contain the most information, Prominence-Interpretation theory suggests that showing a user all of this information may not help people evaluate credibility. Not only would each of the features be less prominent, but the inevitable clutter that would result from so much information would make our entire augmentation less prominent

on a page. Additionally, such an information-heavy intervention may distract users from their primary intent. Consequently, we reduced the size of our feature set by first measuring how well each feature correlated with our ground truth. As expected based on the difficulty of predicting nuanced concepts such as credibility [4, 16], the highest of these correlations were in the "moderate" range of $\rho = .3$ to $\rho = .5$. These moderate correlations indicate that predicting an objective credibility score using the features above would lead to only moderate success; however, they also indicate that these features contain signals that can inform end users in their interpretation of a page's credibility.

We then selected a few features that we felt would be intuitive to end-users, would support an interesting visualization, and that correlated well with ground truth. Features such as spelling errors were omitted due to low correlation, and features such as awards, though sparse, were included because of relatively high correlation. The features we chose were: overall popularity of a Website; popularity of a site among domain experts; the number of zip codes people accessed a site from; receipt of awards and certifications; and PageRank. We then used these features to create visualizations for augmenting Web pages and search results.

VISUALIZATIONS

Prior work indicates that subtle cues (such as color changes to indicate safety levels in anti-phishing toolbars [33]) may not be salient enough to alter users' perceptions of a Web page. However, work on augmenting Wikipedia pages with visualizations of edit history [23] has succeeded in altering users' opinions, perhaps due to the more salient nature of the latter visualizations. Hence, we decided to create a rich visualization to represent our chosen features to end users.

In addition to presenting the features adjacent to Web pages themselves, we created a second, more compact version of the visualization to augment search results. We felt that augmenting search results was particularly important given recent findings that many users make determinations of credibility based on search results pages [17]. Such assessments are particularly in need of assistance, due to the limited information scent [4] of search results as compared with Web pages. We hypothesized that a search result visualization might also have more impact than one on the Web page itself, since the sparseness of search results pages would boost the likely impact of such visualizations according to Prominence-Interpretation theory [11], and since once users have seen a Web page itself, it is difficult to overcome the first impressions they form based on the professional appearance of a page's design [14, 27].

We built a Web application to augment the Web pages and search result snippets for the 1,000 URLs in our data set. The Web page visualization (Figure 1) appears adjacent to the Web page, so that it is visible regardless of scroll positioning. The increasing popularity of wide-screen displays and multi-monitor setups facilitates this horizontal expansion of the browser footprint. The visualization uses color and font size to draw attention to a page's domain type, and includes icons to indicate whether a page has received a

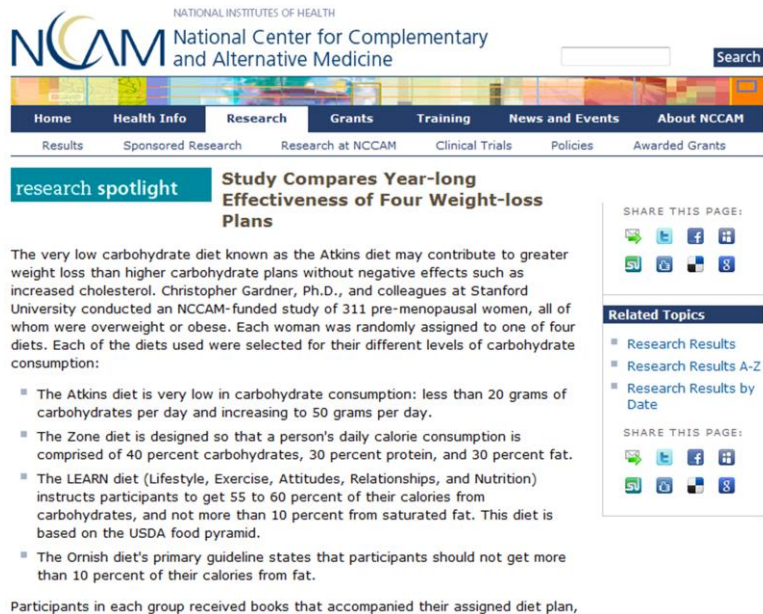
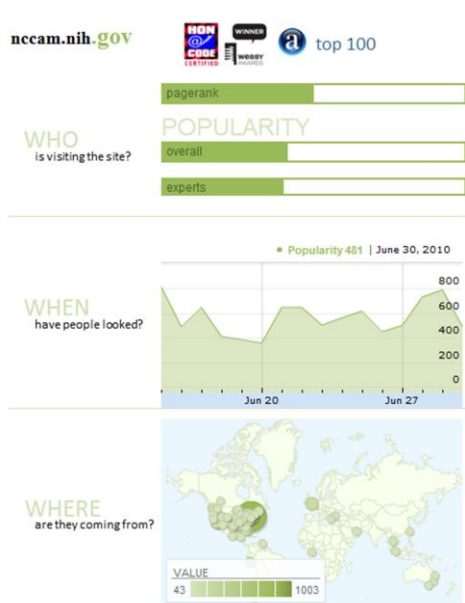


Figure 1. Web page augmented with our visualization.

Webby award, HON certification, or high (top-10, top-100, or top-1000) Alexa ranking. Horizontal bars indicate the relative value of the current page's PageRank, general popularity, and popularity among experts for the page's topic (normalized based on the minimum and maximum values in our data set). Overall popularity is further broken down to reveal temporal and geographic patterns in separate charts. These indicators are grouped thematically according to the interrogative questions "who," "when," and "where."

The search result visualization (Figure 2) is more compact than the Web page visualization, to reflect the space constraints of search result pages. Only items from the "who" category are shown in this condensed view.

EVALUATION

In addition to assessing users' subjective reactions, we also sought to evaluate the following hypotheses:

H1: Users' credibility ratings will be more accurate when the visualizations are available to them.

H2: Users will feel more confident in the accuracy of their ratings when the visualizations are available to them.

H3: The impact (on both accuracy and confidence) of augmenting search results will be greater than the impact of augmenting Web pages.

Alzheimers Disease Genetics - WebMD

Information on Alzheimers disease and the genetic link. ... Will you ever get Alzheimer's disease? Genetics may have the answer. The genes you've inherited carry most of the risk ...

<http://www.webmd.com/alzheimers/guide/20061101/who-gets-alzheimers-genes-hold-key>



Figure 2. Search result augmented with our visualization.

H4: Teenagers' ratings will receive more benefit from these interventions than adults' will.

Method

We conducted a within-subjects experiment to evaluate our hypotheses. We recruited 26 paid participants (13 female) from the Seattle metropolitan area. Participants came to our lab for a one-hour session to complete the study. Participants' ages ranged from 13 to 40 (mean = 21). 15 of our participants were middle-school or high-school students. Adult participants had a variety of occupations unrelated to programming, Web design, or usability. Occupations included freelance photographer, sales representative, and graphic artist. All participants had experience browsing the internet and using search engines (most used them daily).

From our data set of 1,000 credibility-labeled URLs, we selected 21 URLs for evaluation in our study. These 21 URLs were selected by choosing seven URLs associated with each of three queries (Politics: "Obama birthplace", Finance: "Is it a good time to invest in gold?", and Health: "H1N1 vaccine side effects"). For each of these three queries, we selected the seven URLs (from the 40 available) in a manner that enabled us to achieve broad coverage of search result rankings and ground-truth credibility ratings for each query. As discussed earlier in the "Credibility Ratings" section, the ground-truth ratings for these 21 URLs were verified for agreement with ratings from experts in these three topic areas. The Web pages and associated search result snippets for our data set were cached at the time of ground truth data collection to ensure consistency.

For each participant, the experimenter first provided a tutorial session using URLs that were not part of the evaluation set to demonstrate both visualizations and explain the meaning of each visualization component in appropriate terms (e.g., "PageRank measures how many other Web pages link to this page."; "Expert popularity measures how many people who visit specialized sites on a topic also visit

Block	Condition	Topic
1	Basic Search Results	Politics
2	Basic Search Results	Finance
3	Augmented Search Results	Politics
4	Augmented Search Results	Finance
5	Basic Web Page	Politics
6	Basic Web Page	Health
7	Augmented Web Page	Politics
8	Augmented Web Page	Health

Table 2. Study Design. The seven URLs within each block were shown in random order.

this page.”). The experiment procedure was explained in detail, including the rating scale and definition of credibility discussed earlier in the “Credibility Ratings” section.

The study was organized into 8 blocks, with 7 URLs per block (Table 2). The order of URLs within blocks was randomized to reduce order effects. The order of blocks was constant for all participants. Counterbalancing conditions was not desirable, due to the progressively increasing level of information revealed by varying conditions. For instance, it would not make sense to ask a user to rate the credibility of a given URL in the “basic Web page” condition if they had already rated that same URL in the “augmented Web page” condition, since they may remember the additional information revealed by the visualization. Similarly, rating the credibility of a URL in either of the “search result” conditions after already having viewed the full Web page would present a confound.

To enable comparisons on the relative influence of the visualization on search results versus Web pages, the URLs from one topic (Politics) were repeated in all four conditions. Consequently, for the Politics URLs, information is introduced and later removed (between the “augmented Search result” and “basic Web page” conditions). To mitigate this potential learning effect on our data, we tested the other two sets of URLs in one condition rather than cross-conditions, using Finance for search results and Health for Web pages. Additionally, we alternated topics with each block to facilitate forgetting of previously seen URLs.

At the beginning of each block, participants were reminded of the query associated with the current topic, and were instructed to pretend they were conducting that query, in order to contextualize their credibility judgments in a realistic task. Then, for each of the seven URLs on the current topic, participants viewed either the basic search result, augmented search result, basic Web page, or augmented Web page, according to the current condition. They rated the target page’s credibility on a five-point Likert scale, as well as their confidence in their credibility assessment (also on a five-point Likert scale).

To enable completion within participants’ one-hour lab visit and to simulate realistic circumstances (in which users rarely devote substantial time to credibility assessment [28]), there was a limited time to view the contents: 20 seconds for a search result and 60 for a Web page. These timings were selected based on prior findings [6] that users typically spend 10 seconds reading search results. Once the time

Type	Feature	Rating
Search Result	<i>Expert Popularity</i>	4.5
	Summary	4
	URL	3.9
	<i>Awards</i>	3.9
	Title	3.8
	Result Rank	3.5
	<i>PageRank</i>	3.2
	<i>Overall Popularity</i>	3.1
	Web Page	Factual Correctness
<i>Expert Popularity</i>		4.4
Citations		4
Familiarity with Site		3.9
Title		3.8
<i>Domain Type</i>		3.8
Look & Feel		3.7
Author Information		3.6
<i>Awards</i>		3.5
<i>PageRank</i>		3.3
<i>Overall Popularity</i>		3.3
<i>Popularity Over Time</i>		2.8
Number of Ads		2.6
<i>Where People Are Visiting From</i>		2.5

Table 3. Mean Likert ratings of relative feature usefulness in the credibility assessment process, including features of our visualizations (in italics) and features deemed important by prior studies of credibility perceptions.

expired, the content disappeared and the participant was forced to make a rating. Participants rated pages before time expired 90% of the time in the basic search results condition ($\mu=12$ sec.), 94% of the time with augmented search results ($\mu=9$ sec.), 97% of the time with the basic Web page ($\mu=18$ sec.), and 99% of the time with the augmented Web page ($\mu=13$ sec.). Each participant completed 56 ratings.

At the conclusion of the session, participants completed a short questionnaire about the usefulness of the visualizations, utility of different features, and demographics.

RESULTS

We collected a total of 1456 credibility ratings, 1456 confidence ratings, and 26 questionnaires from our study. Using these data, we revisit our initial hypotheses and questions. Due to the subjective (and potentially non-equidistant) interpretations participants may attribute to Likert scales, and the fact that ratings were not normally distributed, we use non-parametric tests when analyzing Likert responses. Specifically, we used Spearman’s ρ to measure inter-rater reliability between user ratings and ground truth, and Wilcoxon tests to evaluate differences in ratings between conditions.

Questionnaire Data

Table 3 summarizes the credibility-assessment-utility participants reported for several features of search results and Web pages. In addition to attempting to judge the correctness of a page’s contents, users found the expert popularity to be most helpful both when viewing full Web pages and also search results. For instance, P18 mentioned “The expert bar was very useful in finding credible information. I would be very happy to have one on search engines I use.” P17 said “I really like the experts bar...this is because if experts don’t use the site...then it must be wrong.”

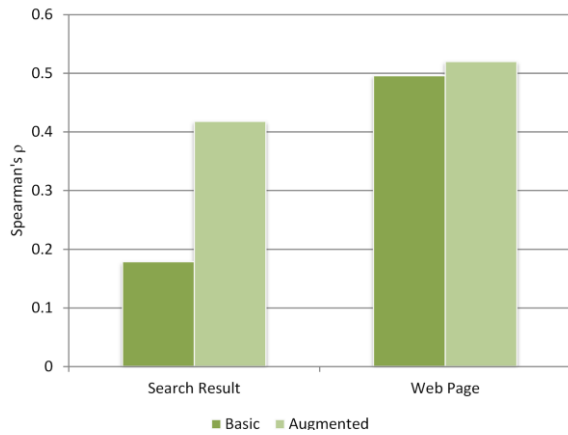


Figure 3. Agreement between participant ratings and ground truth.

Performance Data

Not only did our visualization *seem* useful (receiving a mean utility score of 5.9 on a 7-point Likert scale), our data suggest that our visualization made a significant impact on participants' ability to evaluate credibility. We present our results in the context of our initial hypotheses:

H1: Users' credibility ratings will be more accurate when the visualizations are available to them.

Our data supports this hypothesis for search results. To assess accuracy, we first measured the distance (absolute value of the difference) between users' credibility ratings and ground truth for a Webpage. We also measured the correlation between participant ratings and our ground truth.

Distance from Ground Truth

Our results indicated a significant improvement between the basic search results and augmented search results condition, ($z = -2.50, p < 0.05$). The mean distance between credibility and ground truth in the basic search results condition was 1.25; this distance was reduced to 1.09 in the augmented search results condition. There was no significant difference in the basic Web page (0.84) and augmented Web page (0.92) conditions. These results indicate that the visualization significantly impacted people's ratings when viewing search results but not a Web page.

Agreement with Ground Truth

Not only did our visualization decrease the distance from ground truth in the search result condition, it also improved participants' agreement with the ground truth by a factor of two (Figure 3). The average agreement in the search results condition was 0.17, while the average agreement in the augmented search results condition was 0.42. This difference is statistically significant ($t(25) = -4.75, p < 0.01$).

Not only does our visualization significantly improve users' ability to evaluate the credibility of Web pages when viewing search results, it makes users as accurate as if they were viewing the entire page. Figure 4 shows average agreement between participant ratings and ground truth for the seven "Politics" URLs that were shown in all four conditions. A repeated measures ANOVA test ($F(3) = 2.97, p < 0.05$) and follow-up pairwise t-tests indicated a significant difference

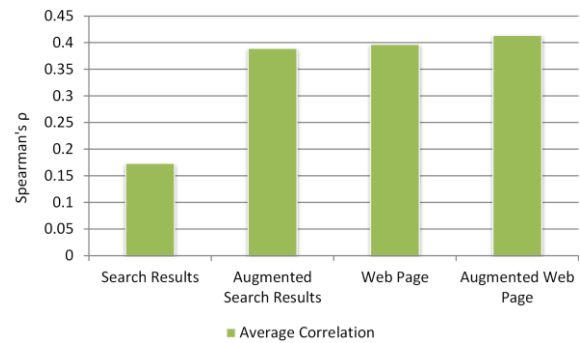


Figure 4. Agreement between participant ratings and ground truth for the subset of URLs (Politics) that were present in all conditions.

between the basic and augmented search results ($t(25) = -2.38, p < 0.05$) as well as the basic search result and augmented Web page ($t(25) = -2.21, p < 0.05$) and (marginally) the basic search result and basic Web page ($t(25) = -1.93, p = .07$), while showing no support for differences between any of the other conditions.

Although not statistically significant, the general trend in our data was that people tend to overestimate the credibility of non-credible pages, and underestimate the credibility of credible pages, and our visualization helped them correct for this over/underestimation.

H2: Users will feel more confident in the accuracy of their credibility ratings when the visualizations are available to them.

Participants consistently felt somewhat confident in their ratings across all conditions. The mean confidence rating (using a five-point Likert scale) across all conditions ranged from 4.1 to 4.3, and there was no correlation between confidence and accuracy. The discrepancy between participants' confidence scores and their actual accuracy further highlights the importance of helping people evaluate information credibility because people trust their judgments regardless of whether they are correct or not.

H3: The impact (on both accuracy and confidence) of augmenting search results will be greater than the impact of augmenting Web pages.

Our data indicates that the impact of augmenting search results is greater than that of augmenting Web pages in terms of accuracy but not confidence. As discussed above, augmenting search results significantly improves credibility accuracy, while augmenting Web pages has little effect, and participant confidence was high in all conditions.

H4: Teenagers' ratings will receive more benefit from our intervention than adults' will.

We found no significant differences in accuracy improvements between teenagers and adults. However, teens found expert popularity to be more helpful than adults when viewing search results. Teens rated expert popularity as more important, on average, than adults ($z = -2.74, p < 0.01$).

DISCUSSION

Our findings suggest that augmenting search results with additional features can help users make more accurate as-

assessments about the credibility of the target Web pages. Aggregate information about topical experts' visitation patterns appears particularly promising in this respect.

While our visualization significantly improved users' ability to evaluate credibility when viewing search results, it had little effect on page-level credibility assessment. The increased impact of the visualizations in the search results context harmonizes with the predictions of Prominence-Interpretation theory [13]. It is particularly striking that the augmentation of search results increased users' credibility assessment performance to the same level as viewing an entire Web page, demonstrating that our visualizations added valuable information scent [4] to search results. Improving the utility of search results for credibility assessment is particularly important, as users increasingly use search engines as their entry point to the Web [9] and already attempt to infer credibility from search engine result lists [17].

The lack of additional benefit of our visualization in the Web page condition is disappointing, but not altogether surprising in light of the large influence a page's visual design has on credibility assessments [14, 27] and prior findings in related areas such as the effectiveness of anti-phishing toolbars [33]. Prominence-Interpretation theory would suggest that, given the amount of information competing for attention when viewing a page, our visualization was not sufficiently salient. The visualization did not detract from performance, however, and therefore may be valuable to retain for cases in which users are making critical decisions (e.g., medical choices) whose motivations are difficult to simulate in laboratory studies, but which have been shown in prior work to influence users to devote additional time and attention to credibility assessment [28].

Educators' emphasis on the need for students to be taught about Web credibility (e.g., [21, 22, 25]) led us to believe that the teens in our study would be less accurate in their credibility assessments (and therefore benefit more from our visualizations) than adults; however, their performance was comparable. Overall, both teens and adults had room for improvement in their credibility assessment skills (and both groups were overconfident in the correctness of their ratings), suggesting that credibility assessment education and interventions may need to be directed to the population at large rather than only to youth.

Our results indicate that search engine companies could add value to their services by augmenting search results with visualizations that reflect features indicative of credibility that are typically not available to end-users, but which are available in aggregate in the opt-in log data such companies already collect. In particular, our participants indicated that page popularity among topical experts was one of the most informative aspects of our visualization. Due to the desire of many search engines to fit as many results "above the fold" as possible, and to represent results compactly so as to reserve space for advertising, even our condensed visualization may be prohibitively large for real-world deployment. Reducing the footprint of the visualization further by only showing the "expert popularity" bar that participants found

most persuasive may be a reasonable compromise. For highly motivated users (e.g., [28]), revealing the other aspects of the visualization upon hover may be acceptable.

In addition to displaying credibility-correlated features (particularly expert behavior) to end-users, search engine companies might consider integrating such data into their ranking algorithms, particularly given user mental models that already assume that ranking is a proxy for credibility [17]. Alternatively, credibility visualizations could be used as interactive advanced query operators, enabling users to sort and filter search result lists by criteria they deem important (e.g., indicating minimum popularity criteria for search results returned, or that they would like to see pages popular in specific regions of the world, or pages that have won certain award categories).

In order to realize these design visions on arbitrary Web pages and search result sets, further research is required on algorithms for automatically determining the topic of a Web page (a partial solution could involve reverse lookup in labeled sets such as the Open Directory Project [dmoz.org] or crowd-sourced efforts such as tagging via social bookmarking tools). Improving upon the heuristic method of identifying topical experts proposed by White et al. [31] is another direction for future research. Their whitelist approach, though simple, provided value to the participants in our study; more sophisticated expertise identification algorithms may further enhance users' credibility assessments.

Long-term assessments of search result augmentation in ecologically-valid settings are also an important area for further research, in order to answer questions such as whether exposure to such visualizations helps users learn over time to be more discerning Web consumers, or whether augmented search results will change users' tendency to predominantly click through to the top-ranked result [20].

The objective versus subjective nature of truth is a topic of extensive philosophical debate, which extends far beyond the scope of this paper. Our approach navigated this nuanced space by creating an "objective" credibility score that has some additional perspectives as sanity-checks (ratings by outside professionals in health, finance, and politics for the pages in our test set), and then using an "informing end-users" approach (as opposed to a predictive approach) to present this data to users in order to support subjective interpretations. However, our definitions of expertise and choice of expert raters may inherently reflect a particular worldview that is not shared by all Web searchers. The fraught politics of defining credibility may make mainstream search engines hesitant to explicitly promote such indicators; implicitly incorporating expertise information into ranking algorithms or displaying it only on educational search portals may be more expedient.

CONCLUSION

Assessing Web page credibility is an increasingly important literacy as people turn to the Web for information in a variety of critical domains. In this paper, we made several contributions toward the goal of enhancing users' ability to

assess Web page credibility, including (1) creating a publicly available data set of 1,000 Web pages with associated credibility ratings, (2) identifying features not readily available to end-users that relate to credibility, and quantifying the degree to which they do so, (3) designing visualizations to augment Web pages and search results that convey the most promising of these features, (4) evaluating the effectiveness of these visualizations in a laboratory study, and (5) offering design suggestions and future research directions based on these findings.

Our findings indicate that visualizing features not readily available to most users (particularly page popularity among topical experts) on search results is an effective way to improve the accuracy of users' credibility assessments. Augmenting search results in this manner makes users' credibility judgments as accurate as if they were viewing the target page in its entirety. These results demonstrate that a simple visualization can help people make more informed decisions as they search for content online, and point the way for future research addressing this important issue.

REFERENCES

1. Banko, M., Cafarella, M.J., Soderland, S., Broadhead, M., and Etzioni, O. Open information extraction from the Web. *IJCAI 2007*, 2670-2676.
2. Bhavnani, S. Domain-Specific Search Strategies for the Effective Retrieval of Healthcare and Shopping Information. *CHI 2002 Extended Abstracts*, 610-611.
3. Brin, S. and Page, L. The anatomy of a large-scale hypertextual Web search engine. *ISDN 1998*, 107-117.
4. Cavarlee, J. and Liu, L. Countering Web spam with credibility-based link analysis. *PODC 2007*, 157-166.
5. Chi, E., Pirolli, P., Chen, K., and Pitkow, J. Using information scent to model user information needs and actions and the Web. *CHI 2001*, 490-497.
6. Cutrell, E. and Guan, Z. What are you looking for? An eye-tracking study of information usage in Web search. *CHI 2007*, 407-416.
7. Dignan, L. Gartner Puts Phishing Tab at \$3.2 Billion. *ZDNet*. December 17, 2007.
8. Ennals, R. Byler, D., Agosta, J.M. and Rosario, B. What is Disputed on the Web? *WICOW 2010*, 67-74.
9. Fallows, D. Search Engine Users. *Pew Internet and American Life Project*, 2005.
10. Ferebee, S. An examination of the influence of involvement level of Web site users on the perceived credibility of Web sites. *PERSUASIVE 2007*, 176-186.
11. Fogg, B.J. Marshall, Laraki, O., Osipovich, A., Varma, C. Fang, N., Paul, J., Rangekar, A., Shon, J., Swani, P., and Treinen, M. What makes Web sites credible?: A report on a large quantitative study. *CHI 2001*, 61-68.
12. Fogg, B.J. Persuasive Technology: Using Computers to Change What We Think and Do. *Morgan Kaufmann*, 2002.
13. Fogg, B. J. Prominence-interpretation theory: explaining how people assess credibility online. *CHI 2003 Extended Abstracts*, 722-723.
14. Fogg, B.J., Soohoo, C., Danielson, D.R., Marable, L., Stanford, J., and Tauber, E.R. How do users evaluate the credibility of Web sites?: A study with over 2,500 participants. *DUX 2003*, 1-15.
15. Gamon, M., Basu, S., Belenko, D., Fisher, D., Hurst, M., and Konig, A.C. BLEWS: Using Blogs to Provide Context for News Articles. *ICWSM 2008*.
16. Gyongyi, Z., Molina, H.G., and Pederson, J. Combating Web spam with TrustRank. *VLDB 2004*, 576-587.
17. Hargittai, E., Fullerton, F., Menchen-Trevino, E. and Thomas, D. Trust online: young adults' evaluation of Web content. *International Journal of Communications*, 2010.
18. Hilligoss, B. and Rieh, S.Y., Developing a unifying framework of credibility assessment: construct, heuristics, and interaction in context. *Information Processing and Management*, 44(4), 2008, 1467-1484.
19. Ivory, M. and Hearst, M. Statistical profiles of highly-rated Web sites. *CHI 2002*, 367-374.
20. Joachims, T., Granka, L., Pan, B., Hembrooke, H., and Gay, G. Accurately interpreting clickthrough data as implicit feedback. *SIGIR 2005*, 154-161.
21. Jenkins, H. Confronting the Challenges of Participatory Culture: Media Education for the 21st Century. 2006.
22. Kapoun, J. Teaching Undergrads WEB Evaluation: A Guide for Library Instruction. *C&RL News*, July/August 1998, 522-523.
23. Kittur, A., Suh, B. and Chi, E. Can you ever trust a wiki?: Impacting perceived trustworthiness in Wikipedia. *CSCW 2008*, 477-480.
24. Lazar, J., Meiselwitz, G., and Feng, J. Understanding Web Credibility: A Synthesis of the Research Literature. *Foundations and Trends in HCI*, 1(2), 2007.
25. Leu, D.J. and Zawilinski, L. The New Literacies of Online Reading Comprehension. *New England Reading Association Journal*, 43(1), 2007, 1-7.
26. Lohr, S. This Boring Headline is Written for Google. *The New York Times*, April 9, 2006.
27. McKnight, D. and Kacmar, C. Factors and effects of information credibility. *ICEC 2007*, 423-432.
28. Metzger, M. J. Making sense of credibility on the Web: Models for evaluating online information and recommendations for future research. *Journal of the American Society for Information Science and Technology*, 58(13), 2007, 2078-2091.
29. Pew Research. Growing Number of Americans Say Obama is a Muslim. August 18, 2010.
30. Schneiderman, B. Designing trust into online experiences. *CACM*, 43(12), 2000, 57-59.
31. White, R., Dumais, S., and Teevan, J. Characterizing the influence of domain expertise on Web search behavior. *WSDM 2009*, 132-141.
32. White, R. and Morris, D. Investigating the querying and browsing behavior of advanced search engine users. *SIGIR 2007*, 255-262.
33. Wu, M., Miller, R., and Garfinkel, S. Do security toolbars actually prevent phishing attacks? *CHI 2006*.