Computer Science • 16 (3) 2015 http://dx.doi.org/10.7494/csci.2015.16.3.295

Michał Kąkol
Radosław Nielek

WHAT AFFECTS WEB CREDIBILITY PERCEPTION? AN ANALYSIS OF TEXTUAL JUSTIFICATIONS

Abstract *In this paper, we present the findings of a qualitative analysis of 15,750 comments left by 2,041 participants in a Reconcile web credibility evaluation study. While assessing the credibility of the presented pages, respondents of the Reconcile studies were also asked to justify their ratings in writing. This work attempts to give an insight into the factors that affected the credibility assessment. To the best of our knowledge, the presented study is the most-recent large-scale study of its kind carried out since 2003, when the Fogg et al. śHow do users evaluate the credibility of Web sites? A study with over 2,500 participants' paper was published. The performed analysis shows that the findings made a decade ago are still mostly valid today despite the passage of time and the advancement of Internet technologies. However we report a weaker impact of webpage appearance. A much bigger dataset (as compared to Fogg's studies) allowed respondents to reveal additional features, which influenced the credibility evaluations.*

Keywords credibility, world wide web, credibility ratings, reliability, credibility assessment heuristics, credibility evaluation factors

Citation Computer Science 16 (3) 2015: 295–310

1. Introduction

Our society increasingly relies on the vast amount of information available on the Internet. Information from the Internet is used not only for entertainment purposes, but also for making life-or-death decisions (e.g., Googling diagnoses – 'dr. Google'), looking for investment opportunities, etc. Nowadays, information credibility on the Internet is one of the main issues in developing the web. A better understanding of how people evaluate webpage credibility is crucial for all who are involved in publishing on the web, but not only for them. Automatic or semi-automatic tools for supporting credibility assessment (like Reconcile.pl, Mywot.com, or Factlink.com) are other possible applications.

More than ten years ago, [4] published the first large-scale study about factors that influence the assessment of web site credibility. In the meantime, the Internet has changed a lot. There are four times more Internet users (2.7 billion in 2013 as compared to 600 millions in 2003[1]) who have lightning-fast Internet connections (tens of megabits nowadays vs. a mere hundreds of kilobits at best in 2003) and use it everywhere, thanks to smartphones equipped with LTE or HSDPA.

This paper is focused on two research questions:

1. How did changes in technology and social environments influence the heuristics used by people for assessing web site credibility?
2. Is the 'Design' still the most-influential factor in web credibility assessment?

Although the methodology in this paper was previously used by [4] (thus making a direct comparison the easiest), some other works are also worth mentioning. [3] contributes to the knowledge of the underlying factors of perceived credibility by adding, inter alia, the viewer's Internet/web experience as one of the crucial factors. [6] directs attention to the role of social- and group-based tools used by viewers in their evaluations based on the research using focus-group data.

The authors of [8] show how features specific to twitter correlate with credibility. [12] identified eight credibility-related features, but also shows that only part of them can be detected automatically.

The computer-aided content analysis applied in this paper is already more than 40 years old. The first attempts were done by Philip Stone [13] in the late sixties. A good overview of the development of computer applications for content analysis and tagging can be found in [2]. The author of [11] discusse the pros (scalability, repeatability, objectivity) and cons (lack of transparency, doubtful) of automatic tagging.

2. Dataset

The dataset used in this paper has been collected as a part of a three-year- long research project focused on semi-automatic tools for website credibility assessment

[1] source: `http://www.internetworldstats.com/emarketing.htm`

– visit project's official site[2] or Sourceforge[3], where some of the project's results are available to download. All experiments were conducted using the same platform. Websites for evaluation were archived (including as well static as dynamic elements – e.g., ads) and served to users together with an accompanying questionnaire. Next, the users were asked to evaluate four more dimensions on the 5-point Likert scale (i.e., a site's appearance, information completeness, an author's expertise, and intentions) and justify their evaluation with a short comment (min. 150 characters).

2.1. C^3 study

The participants of the so-called C^3 study were recruited using the Amazon Mechanical Turk platform using money incentives and restricted to be located in an English-speaking country, excluding workers from India, Pakistan, China, Thailand, the Philippines, etc. A further description of respondent demographics and figures are available in section 2.3.

The corpus of web pages to be evaluated was gathered using three methods: – *manual selection*, *RSS feeds subscription*, and *customized Google queries*. It spans various topical categories: – *politics & economy*, *medicine*, *healthy life-style*, *personal finance*, and *entertainment*. The selection was aimed at achieving a thematically diverse and balanced corpus of a priori credible and non-credible pages – thus covering most of the possible threats on the Web.

As of May 2013, the dataset consisted of 15,750 evaluations of 5,543 pages from 2,041 participants. Users performed the evaluation tasks over the Internet on our research platform via Amazon Mechanical Turk. The respondents independently evaluated the archived versions of the gathered pages not knowing each other's ratings.

We also implemented several quality-assurance heuristics during the study. The evaluation time of a single web page could not be shorter than 2 minutes; the links provided by the users should not be broken; and, linking must be to an English-speaking webpage. Additionally, the textual justifications had to be at least 150 characters long and written in English. As a quality-assurance element, the comments were also manually monitored for spam. In fact, our respondents were quite generous in terms of the length of the comments they left, which is depicted in Figure 1.

Spam and error issues in crowdsourcing applications and collaborative systems, in general, are ubiquitous and need to be handled properly. A kind of trust management should be applied in order to prevent dishonest or ideologically-biased users from contributing noise to the gathered data. Stemming from Internet auction systems [5], a reputation system is a standard countermeasure. Several interesting variations of reputation systems performing such tasks are discussed in [1, 7], and [14].

[2]source: `http://reconcile.pjwstk.edu.pl`
[3]source: `http://sourceforge.net/projects/reconcile2011/`

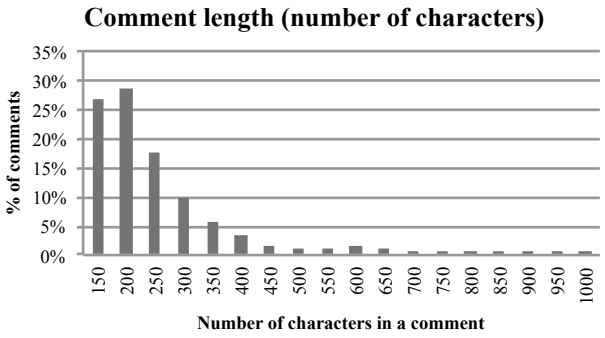

Figure 1. Comment's length distribution.

2.2. Analysis

This paper focuses on the qualitative analysis of the collected user-textual-rating justifications. More studies focused on other dimensions of the C^3 study can be found in [9, 10]. Additionally, for the reader's convenience, we will provide some insight into dataset characteristics in section 2.3.

To obtain a qualitative insight into the credibility assessment factors, a semi-automatic approach has been applied to the results of the C^3 study. In favor of manual labeling, we used text clustering in order to get hard disjoint cluster assignments and topic discovery for soft nonexclusive assignments for a better understanding of the credibility factors. These methods served the purpose of getting the preliminary insight and creating a codebook for future manual labeling. The natural-language processing was performed using SAS Text miner tools. Latent Semantic Analysis and Singular Value Decomposition were used to reduce the dimensionality of the term-document frequency matrix weighted by TF-IDF. Clustering was performed using the SAS expectation-maximization clustering algorithm instead of k-means; additionally, a topic-discovery node was used for Latent Semantic Analysis. Unsupervised learning methods enabled us to speed up the analysis process and, for now, reduce the prone-to-subjectivity tasks to interpretation only.

The semi-automatic analysis itself was performed by analyzing the list of descriptive terms returned as a result of all clustering and topic-discovery steps. This method produced not one labeled dataset but many. In this manner, we attempted to produce the most comprehensive list of reasons that underly segmented-rating justifications. We presume that segmentation results are of good quality, as the received clusters or topics could be easily interpreted in most cases as the thematic category of the commented page. This, in fact, turned out to be a shortcoming in this method. In order to lessen the impact of the page categories, we processed all of the comments as well as each of the categories at one time, used a list of customized subject-related stop-words, or advanced parsing like noun-group recognition. The findings presen-

ted below are a summary of the interpretations of all of the parsing and clustering scenarios taken.

2.3. C^3 dataset insight, interesting distributions

In this section, we will share some additional information about the C^3 dataset. 95% of respondents came from the USA, the number of female participants amounted to 40% of all respondents, and nearly 45% of the participants reported having a higher education (Master's or Bachelor's degree, see Fig. 3). More than a half of the respondents (i.e., 55%) reported their ages in a range from 20 to 30 years (see Fig. 2).

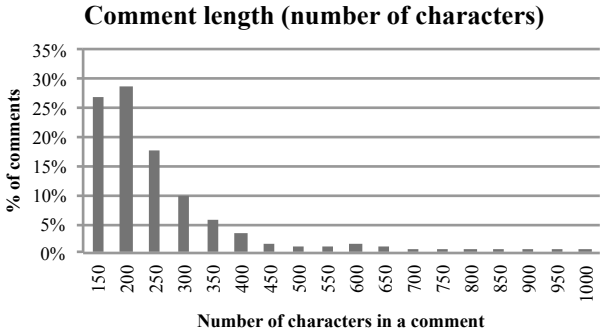

Figure 2. Respondent's age distribution.

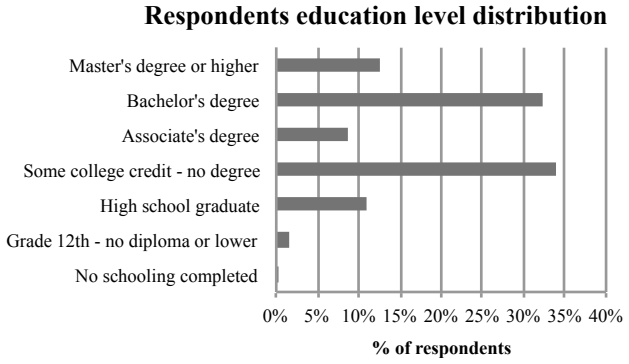

Figure 3. Respondent education level distribution.

The credibility-rating distributions are heavily negatively skewed. On the one hand, the credibility has the most-commonly so- called J-shaped distribution (see Figure 4). But on the other hand, there are also subcategories of the rated pages where the respondents did not come to a consensus on how to assess those particular pages. An example of a controversial topic with an almost uniform distribution of ratings are drug-related pages (e.g., Cannabis): see Figure 5. Figure 6 and Figure 7

depict the apparent effect of education and Internet experience levels on credibility ratings (i.e., low- level users tend to overrate the content).

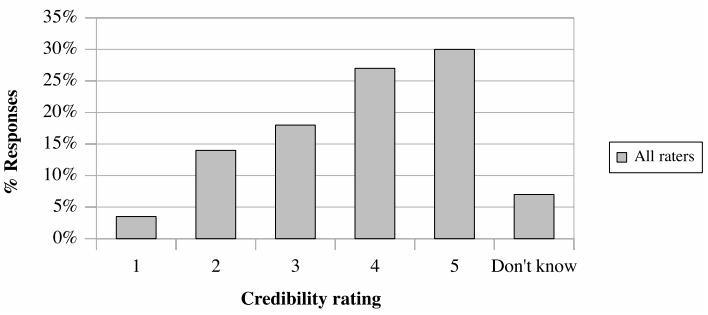

Figure 4. Credibility ratings, all responses.

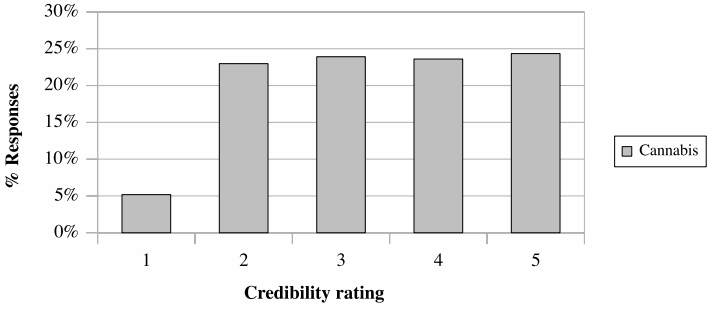

Figure 5. Credibility ratings for Cannabis/Marijuana-related pages.

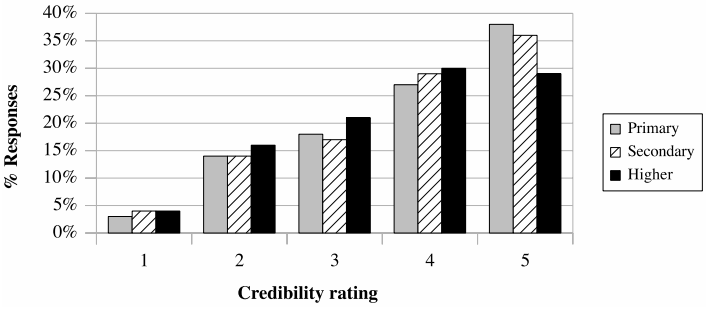

Figure 6. Credibility ratings by education level of respondents.

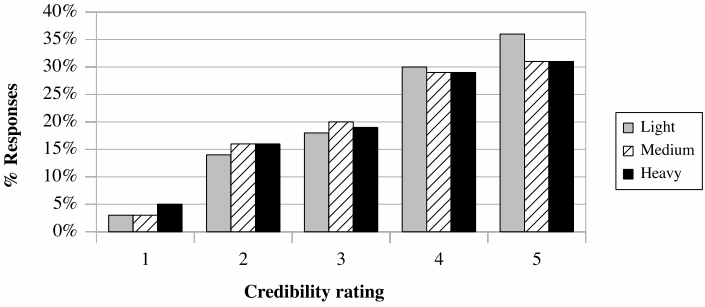

Figure 7. Credibility ratings by Internet experience of respondents.

3. Findings

Just as in [4], we not only present the features influencing credibility assurance that are noticed by the users, but also give some insight into the way of thinking as related to these features. Presented below is a listing of credibility factors together with sample comments describing the identified factors.

3.1. PJIIT Pilot Study

Prior to carrying out the main study, a pilot program was conducted on Computer Science students at PJIIT in Warsaw, Poland, in order to test the platform. Based on a preliminary analysis of approximately 1,000 comments, we introduced an initial codebook for comment labeling, and we made a couple of interesting findings. Firstly, CS students (probably being professionally biased) reported 'the amount of required resources to run the site' as a possible factor of credibility assessment (please note that this factor will not appear in the main study). Secondly, the same page feature could have a positive or negative effect on the student's rating. Because of this, the samples are marked with a sentiment indicator in parentheses (+/- depending on the context of a particular comment; e.g., '- *there are too many ads*' or '+ *there are no ads at all*').

The pilot program was followed by a larger-scale study on Polish-speaking participants, which covered 1,400 participants, 4,000 comments, and revealed 11 possible factors defined by the examples below in this section. The results were manually labeled, thus enabling us to calculate the incidence of labels similar to [4]. We used the codebook from the pilot in practice and decided to improve and extend the set of possible labels for the final C^3 study on English native speakers. As this paper focuses the most on the results of the final C^3 study, the results of studies prior to C^3 are presented only in a fragmentary fashion to give a general overview of all the of the work related to this paper.

(1.1) Author: (+) *'author has his own experience because he works at school as a psychologist"* (+); *'to be a doctor, you have to have a credible information"* (+);

(-) *'British researchers' reports are generally not true"*; (-) *'lack of information about author"*

(1.2) Domain and web site: (+) *'official newspaper"*; (+) *'private blog on the big blogging service"*; (-) *'strange domain name"*; (-) *'I do not believe in specialized content placed on general web sites"*

(1.3) Commercials: (+) *'lack of ads"*; (-) *'focus on selling"*, (-) *'commercials are too strongly connected with the subject of this web page"*

(1.4) References: (+) *'author gives information about the place where this particular research was done"*; (-) *'lack of references to relevant sources"*; (-) *'Medicine based on the bible? No thanks."*

(1.5) Predefined position: (+) *'pros and cons"*; (-) *'page is about legalizing marihuana; thus, cannot be credible"*; (-) *'text is subjective; thus, non-credible"*

(1.6) Language: (+) *'I like the scheme: thesis – antithesis"*; (+) *'web page contains difficult words"*; (-) *'careless language"*; (-) *'youth slang"*

(1.7) Broad verification: (+) *'it is heavily visited page"*; (-) *'it is not a professional web page; thus, only a small number of people read it, and it can contain errors"*; (-) *'It is the opinion of a well-known researcher, but it is only one person."*

(1.8) Informativity: (+) *'there is so much information that it has to be credible"*; (-) *'Page is about everything"*; (-) *'nothing new, only some well-known rumors"*

(1.9) Design: (+) *'picture at the top of the page boosts credibility"*; (+) *'clear design and easy-to-read text"*; (-) *'discouraging look"*; (-) *'old and neglected webpage"*

(1.10) Required resources (motivation): (+) *'author wants to help readers"*; (+) *'a lot of information"*; (-) *'poor text written in 5 minutes"*

(1.11) User experience and knowledge: (+) *'it is true because I've heard about it earlier"*; (+) *'examination sessions are tough times for students, so, I'm using most of these supplements"*; (-) *'my girlfriend uses it, but positive effects are difficult to notice"*; (-) *'it is contradictory to what I have learned in school".*

3.2. C^3 Study credibility factors

Based on the previous codebook and the findings using the methodology described in 2.2, we analyzed the C^3 dataset and revealed 23 credibility-assessment factors, which are apparently easily grouped into 6 assessment heuristics. These groups are enumerated below in this section.

What is it?

(2.1) Type of internet content: (+) *"I like the information I get on forums, because most times I know the moderators of these forums safeguard against spammy posts"*; (-) *"This is just a blog, so I don't know if it is all that credible"*

(2.2) Celebrity gossip: (+) *"It is celebrity gossip, which is credible but also overly sensationalized"*; (-) *"Celebrity gossip often has gaps in its information"*

(2.3) News source: (+) *"I recognize CBS as a reputable news source"*; (-) *"Highly opinionated and subjective news from largely unknown sources"*

(2.4) Scientific study: (+) *"The article is based on scientific study by a reputed organization"*; (-) *"The support given for this idea in the article was not scientifically substantiated and was based on private studies by a commercial company"*

Is it of commercial character?

(2.5) Advertising: (+) *"Adobe doesn't normally buy advertising off any bad sites"*; (-) *"There is much advertising that kind of throws off the design"*

(2.6) Sales offer: (+) *"There are no visible commercial ties"*; (-) *"As with anyone trying to sell you a product, there will be at least a slight push towards sales"*

Who is the author or publisher?

(2.7) Known author: (+) *"The author signed their name to the article"*; (-) *"The lack of attribution to the author and a lack of review from a MD leads the viewer to have doubts"* (2.8) Authority of author: (+) *"The author of this particular article, Maria Golia, is a well known author with a wealth of information on Egypt"*; (-) *"The author does not present him/herself as an authority on the topic"*

(2.9) Official page: (+) *"This appears to be an official, direct from the game studio website"*; (-) *"It is not official website, so it is not completely credible"*

(2.10) What is the source: (+) *"Johns Hopkins is a nationally known hospital and education center"*; (-) *"It did not have .edu or .gov suffix, so that was not in favor of credibility"*

How does it look like?

(2.11) Broken links: (+) *"All pictures and videos worked and loaded correctly"*; (-) *"There are a lot of broken links and the website has a very poor layout"*

(2.12) A lot of links: (+) *"It is also credible because it has a lot of links to relevant information"*; (-) *"There are no links to further information"*

(2.13) Contact information: (+) *"Contact information is also included, which I called to verify"*; (-) *"Site also lacks contact info or FB/Twitter links"*

(2.14) Content organization: (+) *"Health sites like these are well organized"*; (-) *"Very unorganized content"*

(2.15) Design: (+) *"The site looks pretty good, but it is also a bit basic"*; (+) *"Governmental organization with a sleek, professional-looking website"*; (-) *"Very poor design and appearance"*; (-) *"Lacks professional look and feel'*

Is it good to read?

(2.16) Easy to read: (+) *"Summarizes study with easy-to-read language"*; (-) *"I found it too long and rather boring"*

(2.17) Well-written – language: (+) *"The articles looks well written and informational"*; (+) *"it's an interesting read and did not bore me'*; (-) *"The multiple misspellings throughout the article made me somewhat skeptical of all of the details"*

(2.18) Informativity, completeness: (+) *"The article is detailed with a lot of information"*; (-) *"I don't want reviews from fans that are not accurate"*

Is it verifiable?

(2.19) Easy to google out: (+) *"Google searches do legitimize the site"*; (-) *"A Google search returns mixed reviews on the site"*

(2.20) Obejctivity – personal opinion, review: (+) *"Just a review, so it's basically their opinion or point of view"*; (-) *"There is also a certain amount of personal opinion that seems to contain a negative bias"*

(2.21) References – referring credible sources: (+) *"The article is very detailed and cites its sources"*; (-) *"Information is provided by one person, no references listed"*

(2.22) Freshness, date of publishing: (+) *"The website provides a current publishing date"*; (-) *"It seems credible but just offers outdated information"*

(2.23) Viewer's experience: (+) *"I am advocating an understanding of my experience"*; (-) *"I'm not experienced about this topic and could be wrong"*

4. Discussion

4.1. How the changes in technology and social environments influence heuristics used by people for assessing web site credibility?

We tried to match tags proposed in this paper and Fogg's reported factors in an attempt to answer this first research question.

Matching both tag sets with respect to their interpretations, we find out that all of the Fogg's factors can be matched with C^3 tags. 11 make direct one-to-one matches, while seven match with two or three C^3 tags. 18 Fogg factors were represented using 14 C^3 tags; i.e., *Advertising*, *Sales offer*, *Known author*, *Authority of author*, *Official page*, *What is the source*, *Broken links*, *A lot of links*, *Contact information*, *Content organization*, *Design*, *Easy to read, Well written* – language, *Informativity*, *Completeness*, *Easy to google out*, *Objectivity* – personal opinion, *Review*, *References* – referring credible sources, *Freshness*, *Viewer's experience*. The result of matching is depicted in Figure 8, with the numbers representing the identification numbers of the tags found in previous sections. There are eight (white rows) C^3 factors that do not find representation in [4].

The factors affecting credibility evaluations of the participants of Fogg's et al. study were reported as follows: (1) *Design Look*, (2) *Information Design/Structure*, (3) *Information Focus*, (4) *Company Motive*, (5) *Usefulness of Information*, (6) *Accuracy of Information*, (7) *Name Recognition & Reputation*, (8) *Advertising*, (9) *Bias of Information*, (10) *Tone of the Writing*, (11) *Identity of Site Sponsor*, (12) *Functionality of Site*, (13) *Customer Service*, (14) *Past Experience with Site*, (15) *Information Clarity*, (16) *Performance on a Test*, (17) *Readability*, (18) *Affiliations* [4].

Factors like *Celebrity gossip*, *News source*, and *Scientific study* are tightly linked to the subject of the assessed page, but still lead C^3 respondents to give specific credibility ratings – those factors are available in the C^3 list only. The dependence of a web-credibility rating on the subject also finds confirmation in [3]. Surprisingly, Fogg

et al. did not find *Type of Internet content* as a significant Factor. C^3 respondents paid attention to what kind of content they evaluate (e.g., personal blog or public forum), which can it be compared to using different evaluation heuristics for different media types. More of a surprise is the fact that none of the factors reported in [4] concern references and citations in a page's text, which in C^3 study is named *References*. Another tag to find only in C^3 is *A lot of links*, which might be similar to *References*, but some of the respondents seemed to treat citations and links differently, with an emphasis on the number of links (high or low) existing on a page. *Freshness*, which is the importance of the publishing date or the fact of whether the assessed information is up-to-date), also does not exist in the compared list of factors. The last factor found in C^3 but absent in [4] is *Easy to google out*. What is seen in C^3 comments, the participants attempted to verify the information on the presented page mostly by using search engines like Google and querying the subject of the assessed page. Apparently, the respondents did not perform the task in complete separation, but reached for external sources while evaluating. This C^3 factor additionally confirms the findings of [6].

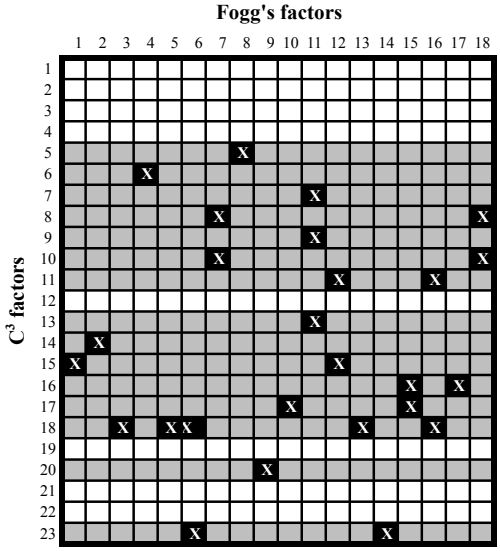

Figure 8. Confusion matrix of C^3 tags vs. Fogg tags, with the numbers representing the identification numbers of the tags found in previous sections.

We can answer the first research question that, according to the study findings, the heuristics used in credibility assessment did not change significantly over the last decade. The exception here is the use of external resources (like search engines) helping the evaluation, which is an emanation of technological advances. Moreover, the different steps of our studies were performed on participants from different cultural circles (English/Non-English speaking). We successfully reused the codebook created

on findings from one circle to another; thus, we may conclude that the factors affecting credibility seem to be culturally independent.

Contrary to Fogg's findings (where frequencies of the comments tagged with a particular label were given), we do not present such quantitative data in this paper. As this work may seem preliminary, it was itself a thorough preparation for manually labeling the dataset that will bring a qualitative perspective on user comments – (which is planned in future work to extend this article). In addition, we want to emphasize the size differences between both studies, as C^3 covered 55 times more pages than [4] and gathered six times the number of comments. Compared to studies presented in [3, 6] C^3 also happens to be of a bigger scope. A larger scope of the study most likely made it possible to find a wider range of diversified credibility-assessment factors.

4.2. Is the 'Design' still the most influential factor in web credibility assessment?

At this point, the C^3 dataset (i.e., textual justifications of the ratings) is not labeled in the sense of Fogg's research. But still using the 'Design' factor identified descriptive terms, we are able to identify the responses where the 'Design' affected the respondent's assessment. The keywords we searched for in the comment text were as follows: *design*, *layout*, *laid out*, *graphics*, *interface*, *appeal*, *attractive*. The ratings with corresponding user's comments containing the above-mentioned keywords were considered as 'Design' affected assessments.

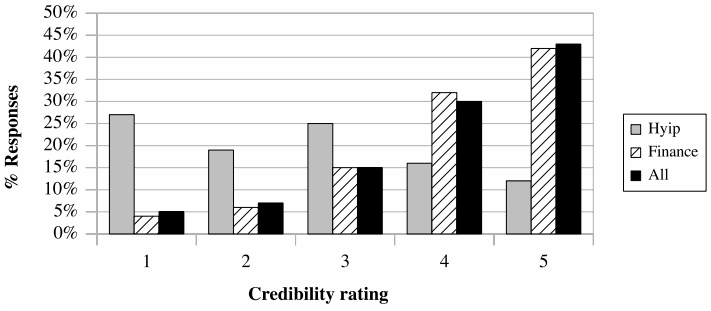

Figure 9. Credibility ratings distribution by different sites categories.

For the purpose of analysis, we are going to use a specific subset of the pages (i.e., *Personal finance category*). This category includes a subcategory of High-Yield Investment Programs (HYIP), which are classic Internet scams offering an incredibly high return rate. HYIP pages can easily be assumed as a priori highly non-credible, which is neatly depicted in Figure 9, where HYIP-related ratings are visibly positively skewed. The distributions of gathered credibility and appearance ratings for HYIP pages in comparison to other categories are available in Table 1.

Table 1

Credibility and presentation rating values correlation by page categories and design keywords *(alternative hypothesis: true correlation is not equal to 0).*

Group	**Design (keywords)**	**Pearson correlation**	
		p-value	**Estimate**
Hyip	All	2.20E-16	0.73
	No design	2.20E-16	0.72
	Design	2.20E-16	0.73
Finance	All	2.20E-16	0.55
	No design	2.20E-16	0.54
	Design	2.20E-16	0.62

The design keyword (thus, comments affected by the 'Design' factor) are more likely to be found among HYIP pages – 19% of HYIP evaluations are affected by 'Design' in comparison to 12% of other Finance pages (see Table 2). Despite the fact that the corpus of pages used in our study is different than the one used in [4], it is still worth mentioning that this level of incidence is much lower than the 46% reported by Fogg. The linear correlation between credibility and appearance ratings is significant for all of the compared groups. The correlation is moderately high and positive. The correlation value apparently increases for the 'Design'-affected groups. This can be interpreted as the users not only rating their appearance higher, but also reported this fact and eventually gave a higher credibility rating for the page. The particular case of the HYIP group shows a higher correlation than average, almost independent from the occurrence of 'Design' keywords. Which leads to the assumption that 'Design' factor is especially important in the case of a priori non-credible pages – those malicious pages can apparently deceive the user mostly by their looks.

Table 2

Credibility and appearance ratings for HYIP and Finance categories grouped by 'design affected' assessments.

Group	**Design**	**Credibility** [%]					**Appearance** [%]				
		1	**2**	**3**	**4**	**5**	**1**	**2**	**3**	**4**	**5**
Hyip	All	27	19	25	16	12	23	18	20	20	18
	No design	29	19	26	14	11	23	19	21	21	15
	Design	17	21	20	25	17	21	17	15	17	31
Finance	All	4	6	15	32	42	5	12	15	33	36
	No design	4	6	15	32	43	4	12	15	34	36
	Design	4	10	16	31	40	9	15	15	25	36
Other	All	4	7	14	30	45	7	14	13	30	36
	No design	4	7	14	30	46	6	13	13	31	37
	Design	5	10	15	31	39	13	19	11	23	34
All	All	5	7	15	30	43	7	14	13	30	36

5. Conclusions and future work

The study presented in this paper shows that the factors affecting website-credibility assessment identified over 10 years ago are still valid. Appearance, the most important factor indicated by [4], is still strongly correlated with credibility evaluation. However, the effect that this factor had on assessment might be weaker than a decade ago. Thus, a further investigation and manual labeling of the dataset is planned in order to measure the effect of all credibility factors reported in this paper. Due to the large scale of the study, we have been able to not only confirm many findings from multiple related papers, but also reveal some new heuristics used by people. Contrary to other research known to the authors (where only manual tagging and manual analysis were performed), we reached for semi-automatic methodology.

Understanding the heuristics and motives underlying web credibility assessment may help us to develop automatic tools for credibility assessment; thus, contributing to a better Web in which users are provided with tools that support them in credibility evaluation and keeps them protected from unreliable content. Confirmation of the previously reported factors and the search for new ones is a step closer to providing Internet users with such tools. To complement the qualitative analysis presented in this paper, we plan carrying out further studies. After getting the insight to the large set of credibility rating justifications, we were able to produce a comprehensive list of factors leading to certain credibility evaluations. The same dataset is planned to be manually tagged in a crowdsourcing environment (e.g., Amazon Mechanical Turk) in order to quantify the importance of the discovered factors. This will make it possible to make direct comparisons with previous findings about what impacts credibility assessment and make confident recommendations on credibility-oriented design and web-browsing safety measures. Furthermore, the experience gained in the iterative acquiring and labeling of credibility rating justifications, together with the final crowdsourced and labeled dataset, is planned to be used as training input for the automatic classifier. The classifier will be used as an automated tagger of comments left by users in the Reconcile.pl system, thus serving the purpose of this work (helping users stay safe on the Internet).

Acknowledgements

This work was supported by the grant Reconcile: Robust Online Credibility Evaluation of Web Content through the Swiss Contribution to the enlarged European Union.

This work was financially supported by the European Community from the European Social Fund within the INTERKADRA project

References

[1] Borzymek P., Sydow M., Wierzbicki A.: Enriching trust prediction model in social network with user rating similarity. In: *Computational Aspects of Social Networks, 2009. CASON'09. International Conference on*, pp. 40–47, IEEE, 2009.

[2] Diefenbach D. L.: Historical Foundations of Computer-Assisted Content. *Theory, method, and practice in computer content analysis*, vol. 16, p. 13, 2001.

[3] Flanagin A. J., Metzger M. J.: The role of site features, user attributes, and information verification behaviors on the perceived credibility of web-based information. *New Media & Society*, vol. 9(2), pp. 319–342, 2007.

[4] Fogg B., Soohoo C., Danielson D.R., Marable L., Stanford J., Tauber E.R.: How do users evaluate the credibility of Web sites?: a study with over 2,500 participants. In: *Proceedings of the 2003 conference on Designing for user experiences*, pp. 1–15, ACM, 2003.

[5] Kaszuba T., Hupa A., Wierzbicki A.: Advanced feedback management for internet auction reputation systems. *Internet Computing, IEEE*, vol. 14(5), pp. 31–37, 2010.

[6] Metzger M. J., Flanagin A. J., Medders R. B.: Social and heuristic approaches to credibility evaluation online. *Journal of Communication*, vol. 60(3), pp. 413–439, 2010.

[7] Morzy M., Wierzbicki A.: The sound of silence: Mining implicit feedbacks to compute reputation. In: *Internet and Network Economics*, pp. 365–376, Springer, Berlin, Heidelberg, 2006.

[8] O'Donovan J., Kang B., Meyer G., Hollerer T., Adalii S.: Credibility in context: An analysis of feature distributions in twitter. In: *Privacy, Security, Risk and Trust (PASSAT), 2012 International Conference on and 2012 International Confernece on Social Computing (SocialCom)*, pp. 293–301, IEEE, 2012.

[9] Rafalak M., Abramczuk K., Wierzbicki A.: Incredible: Is (Almost) All Web Content Trustworthy? Analysis of psychological factors related to website credibility evaluation. In: *Proceedings of the companion publication of the 23rd international conference on World wide web companion*, pp. 1117–1122, International World Wide Web Conferences Steering Committee, 2014.

[10] Rafalak M., Bilski P., Wierzbicki A.: Analysis of Demographical Factors Influence on Websites Credibility Evaluation. In: *Human-Computer Interaction. Applications and Services*, pp. 57–68, Springer, International Publishing, 2014.

[11] Scharkow M.: *Automatische Inhaltsanalyse und maschinelles Lernen.* epubli, 2012.

[12] Shariff S. M., Zhang X., Sanderson M.: User Perception of Information Credibility of News on Twitter. In: *Advances in Information Retrieval*, pp. 513–518, Springer, International Publishing, 2014.

[13] Stone P.: Improved Quality of Content Analysis Categories: Computerized Disambiguation Rules Forhigh Frequency English Words, presented at an National Conference on Content Analysis, 1967.

[14] Wierzbicki A.: The case for fairness of trust management. *Electronic Notes in Theoretical Computer Science*, vol. 197(2), pp. 73–89, 2008.

Affiliations

Michał Kąkol
Polish-Japanese Academy of Information Technology, Warsaw, Poland,
michal.kakol@pjwstk.edu.pl

Radosław Nielek
Polish-Japanese Academy of Information Technology, Warsaw, Poland, nielek@pjwstk.edu.pl

Received: 20.01.2015
Revised: 20.03.2015
Accepted: 20.03.2015