Temporal, Cultural and Thematic Aspects of Web Credibility*

Radoslaw Nielek[1], Aleksander Wawer[2], Michal Jankowski-Lorek[1], and Adam Wierzbicki[1]

[1] Polish-Japanese Institue of Information Technology,
ul. Koszykowa 86., 02-008 Warsaw, Poland
{nielek, fooky, adamw}@pjwstk.edu.pl
[2] Institue of Computer Science Polish Academy of Science
ul. Jana Kazimierza 5, Warsaw, Poland
axw@ipipan.waw.pl

Abstract. Is trust to web pages related to nation-level factors? Do trust levels change in time and how? What categories (topics) of pages tend to be evaluated as not trustworthy, and what categories of pages tend to be trustworthy? What could be the reasons of such evaluations? The goal of this paper is to answer these questions using large scale data of trustworthiness of web pages, two sets of websites, Wikipedia and an international survey.

Keywords: trust, language, Wikipedia, temporal, national, credibility

1 Introduction

In the early 90s there was a need to organize an increasing number of websites. People had problems with navigating a still expanding Internet. Search engines and Internet catalogues have flourished in order to address this issue. At that time, not only content credibility but even Internet frauds were not serious issues. Increasing number of less proficient Internet users and lowering costs of publication became a driving force in this change and have stimulated the bloom of Internet frauds.

Warning against potentially harmful websites constitutes a very useful feature but, for some time now, no longer sufficient. The broad spectrum of information, ranging from completely non-credible (e.g. theories that earth is flat) to very credible, can be found on the Internet and, therefore, users need a support while deciding whether to trust a particular information or not. It is the subject of dynamic research (e.g. Reconcile project) and also some commercial and semi-commercial projects (Hypothesis, FactLink, Web of Trust).

On the one hand, web content credibility gets more and more important for the Net. On the other hand, not much is known about temporal, cultural and

* Research supported by the grant "Reconcile: Robust Online Credibility Evaluation of Web Content" from Switzerland through the Swiss Contribution to the enlarged European Union

thematic patterns of websites credibility. Since the early 50s concept of credibility has been widely studied by psychologists, media experts and economists. Most publications focus on either persuasive effect of source credibility[1-3] or media credibility[4] or importance of credibility for economic theories[5]. Most researchers agree that credibility is not a property of object, person or piece of information but it is rather a perceived quality[6].

Researchers argue whether credibility is a subjective or objective matter. Tseng[6] proposed four types of credibility: presumed, reputed, surface and experienced. The first two types are based on either stereotypes or third-party reports. The last two are derived from individuals own experiences. Some people can argue that those categories are an essentially heuristic use to assess credibility and do not define different types of assessed variable. This view seems to be strengthened by the definition of credibility as believability, given by Fogg[7].

People use heuristics to assess credibility. A physically attractive person is perceived more credible[8]. People constantly use some signals to estimate credibility of others and the same type of mechanism exists also in terms of assessing web content credibility. Study conducted on over 2500 Internet users at Stanford University revealed 18 areas that people notice while assessing web site credibility[9]. Almost 50% participants pointed at design and look, the one forth on information design and information structure. Bias of information and tone of the writing are present only in ca. 10% comments. On the other hand, asking people explicitly about features they use to assess credibility can only reveal heuristics they are aware of. People also adapt their heuristics specifically to web sites. They use position in an Internet search engine (higher position indicates more reliable information[10]) or following graphs and presence of shortened URLs for tweets[11].

The prominence-interpretation theory[12] tries to combine signals with peoples motivation and perceptibility. The theory assumes that first, user has to notice particular feature and only then he starts evaluating it. This process is repeated many times by a single user for a single web site and its efficiency depends strongly on users motivation and experience. The prominence-interpretation theory is mainly focused on conscious processing and ignores preapprehension and feelings in general. It is worth noticing the fact that some features can be difficult to notice by people (e.g. number of question marks or punctuations) but still may be highly correlated with content credibility[13].

Most studies are focused on an attempt to understand factors influencing a credibility at an individual level (either web site or person) but large-scale systems supporting credibility evaluation build in the last few years create an opportunity to take a closer look on credibility of the Internet (or at least a huge amount of web sites) and dependencies interrelation between time, subject of a web site, language and credibility. Many questions seem important from sociological point of view. Among them:

- Are the web sites becoming more credible?
- What is the most credible subject on the Internet?
- Is web sites credibility evaluation related with a trust level in societies?

This paper is devoted to an attempt to answer these questions. Such answers may also have many practical applications, namely can be used for improving automatic credibility evaluation by incorporating additional context information (e.g. subject, language etc.).

No large-scale study of trustworthiness and credibility of web sites, which focus on such dimensions like language, time or culture, exists. Therefore, the paper is intended to fill this gap. The main assumption was to analyze existing credibility ratings and real web sites instead of orchestrating a dedicated surveys or craft special content. The particular attention was placed in assuring the scale that justifies generalization and makes drawing conclusions for the whole Internet possible. In total, more than 600 thousands web sites have been analyzed.

The rest of the paper is organized as follow. In the next chapter, are described datasets used in this paper . Results obtained for these datasets are presented in chapter three. Chapter four is focused on discussions about hypothesis that may explain the results. The last chapter summarizes the paper and proposes some interesting topics for further investigation.

2 Datasets

2.1 Article Feedback Tool

For our analysis we have used dataset build upon results of Wikimedia Foundation experiment with the feature to capture reader quality assessments of articles. Article Feedback v4 (AFT) was allowing users of English Wikipedia to rate every article with 4 different dimensions (trustworthy, objective, complete, well-written). AFT is a survey for article feedback to engage Wikimedia readers in the assessment of article quality. Reader is presented with short survey below every article and he can submit his ratings about four different aspects of article by choosing on 0 to 5 stars scale.

Original AFTv4 dump contains over 11M articles ratings based on 5.6M different revisions of over 1.5M distinct articles collected between July 2011 and July 2012. For comparison we have selected two subsets of data containing first 3 and last 3 months of article ratings and aggregated them grouping by article. Then we excluded pages that werent present in both datasets.

2.2 Web of Trust

WOT is a crowdsourcing system started in 2005 by two post-graduate students from Finland. Every logged user can evaluate each visited web site on four dimensions (trustworthiness, privacy, vendor reliability, and safety for children) and may also add comment to the evaluation. WOT aggregates all evaluation and show a pair of values (ranging from 0 to 100) for each dimension level of confidence and value of evaluation. However, the WOT does not disclose how this numbers are calculated. According to the company's blog numbers presented to the users are the average of left evaluations weighted with rater's credibility.

Up to now users have evaluated 43 million web sites and every month more than 500 thousands new web sites are evaluated.

WOT neither publish the dataset nor make it open for scientists. The only way to access information about web site credibility is via an open API. A researcher can send a question about a particular web site and he will obtain the same answer as plugin users. Datasets studied in this paper have been collected by multiple sending requests to WOT API and recording results. There is no publicly available list of web sites evaluated in WOT system, so an external lists have to be used: 1M most popular web sites from the Alexa and the DMOZ catalogue.

Only a part of domains from both lists have evaluations and this fact is also strongly correlated with its popularity. For the first few thousands most popular domains only a small fraction does not have evaluation (less than 5%) which is not very surprising (the more people visit the web site it is more likely that someone will evaluate it). On the other hand only one-third of domains in the second half of the Alexa rating list have at least one evaluation. In average, 41% domains on the list are evaluated.

Quite often web sites have been evaluated at only one or two dimensions (instead of four). Correlation level between different evaluation dimensions is very high and exceed 0.95 for all but child safety dimmension. That is why in this paper only two dimensions are studied – trustworthiness and child safety.

2.3 Category detection

All domains have been assigned to categories with help of the AlchemyAPI `www.alchemyapi.com`, a services which is based on NLP tools. For academic accounts Alchemy limits the number of requests to 30 thousands per day. The motivation to use AlchemyAPI instead of other methods (e.g. TF IDF with a manually tagged corpus) was the intention to make these results easily replicable for other scientists.

2.4 Nation level data

The procedure to obtain the dataset was as follows. We started with all tokens, known as correct words in each language in the Aspell library. That limited our list of languages. We then replaced accented characters with their non-accented (ascii) equivalents, because of domain names restrictions. The lists of tokens generated this way were submitted to WOT query API. Unfortunately, this seemed the only plausible procedure since WOT does not enable to browse the data it collects and one needs to query for specific domain.

3 Results

3.1 Effect of time on web sites credibility

Freshness is one of factors that influence content credibility[14]. Time can affect web site credibility in many ways. Even if the web site content has not change,

new facts or discoveries may make it irrelevant or wrong (the same effect can be observed for science as well as for sport). Many web sites are regularly updated and each update may change web sites credibility (people may struggle to improve published content but may also use previously gained reputation to sell products or misinformation).

An attempt to trace credibility changes of a particular web site faces many difficulties. The most obvious is that evaluations have to be done at many points of time (we cannot ask people for evaluate an old version even if we have stored content, because the passage of time might change content credibility). To solve this problem, credibility rating from WOT has been collected every two days for almost four months. Results for first and last run are presented in Table 1 (for Alexa) and Table 2 (for DMOZ catalogue).

Table 1. Average trustworthiness (trust) and child safety (safety) for two snapshots in time ("old" – September 2012 and "new" – January 2013) for domains from the Alexa. Statistically significant differences are denoted with stars.

Category	no. of domains	Trust_old	Trust_new	Safety_old	Safety_new
Arts&entertainment	29257	79.88	79.92	73.11	73.09
Business	36218	78.14	78.18	80.60	80.59
Computer&internet	41497	76.09*	76.03*	76.40*	76.36*
Culture&politics	16098	73.17	73.20	60.98	60.96
Gaming	7491	79.90	79.91	77.46	77.44
Health	6122	77.43	77.45	77.85	77.87
Law&crime	1535	74.43*	74.24*	70.97	70.83
None	66125	75.55	75.57	75.60	75.62
Recreation	34522	76.51	76.49	72.46*	72.42*
Religion	5251	80.52	80.45	80.58	80.50
Science&technology	11149	81.84	81.86	82.18	82.13
Sports	7389	82.60*	82.70*	82.87*	88.96*
Weather	131	86.32	86.24	88.85	88.83
All	263444	77.19	77.19	75.46*	75.44*

A relatively short time span between measurements, high number of web sites and the fact that WOT returns only an aggregated credibility score for all evaluations (very old and relatively new) causes that big differences in aggregated credibility for categories should not be expected. In fact, differences are small and in most cases statistically not significant. As can be seen in Table 1 , average credibility has increased for eight categories but only for one – sport – the difference is statistically significant. Among five categories, which have lower credibility in a second run, only for Computer&Internet and Law&Crime differences are statistically significant. Similar, small changes can be also observed for the dimension child safety. The main difference is that an average child safety for all web sites has slightly decreased but this change in opposition to trustworthiness is statistically significant.

In Table 2 is presented a comparison of trustworthiness level for two runs for domains from the DMOZ catalogue. Only for two categories – culture&politics and science&technology — can be observed statistically significant differences and, in both cases, web sites are getting more trustworthy. Very small change in average trustworthiness of all domains is visible but cannot be confirmed as statistically significant.

On the other hand very interesting pattern exists for the 20% most trustworthy web sites in the Alexa. Trustworthiness has increased from 95.28 to 95.32 and this change is statistically significant on the level 0.005. The same effect can be observed for the AFT dataset where average trustworthiness has risen from 84.13 to 84.35 (statistically significant on the level 0.00003). It may be an interesting point in the discussion about the rich get richer hypothesis. For the AFT dataset such an effect does not occur.

Other interesting source making a temporal analysis of trustworthiness evolution for a big number of web sites possible, is the AFT dataset , described in details in previous chapter. In table 4 are presented results for two subsets (first three months and last three months). In opposition to results for the Alexa and the DMOZ for almost all categories average trustworthiness is decreasing and results are statistically significant. The same effect can be also observed for all articles.

Table 2. Average trustworthiness for two snapshots in time (old – September 2012 and new – January 2013) for random sample of domains from the DMOZ catalogue. Statistically significant differences are denoted with stars.

Category	no. of domains	Trust_old	Trust_new
Arts&entertainment	17656	73.93	73.96
Business	34754	71.81	71.81
Computer&internet	25660	73.08	73.02
Culture&politics	10570	74.19*	74.28*
Gaming	1923	75.68	75.79
Health	7113	72.27	72.28
Law&crime	2225	72.32	72.30
Recreation	15706	72.47	72.48
Religion	7963	73.39	73.42
Science&technology	8634	75.30*	75.46*
Sports	8045	73.45	73.46
Weather	54	77.07	77.76
All	141232	73.05	73.04

3.2 The most credible websites are about...

All web sites in both datasets have been assigned to one of twelve categories (plus None). Differences in trustworthiness between all categories presented in

Table 3. Average trustworthiness for two snapshots in time (old – first three months in dataset and new – last three months) for random sample of 50% articles with evaluation from the AFT dataset. Statistically significant differences are denoted with stars.

Category	no. of domains	Trust_old	Trust_new
arts&entertainment	53461	3,02*	2,99*
business	17151	2,80*	2,75*
computer&internet	12626	2,73*	2,66*
culture&politics	27691	2,83*	2,78*
gaming	3131	3,19*	3,16*
health	6725	2,74*	2,70*
law&crime	830	2,65	2,75
recreation	4739	2,99*	2,94*
religion	14992	2,86*	2,83*
science&technology	46790	2,79*	2,74*
sports	8414	3,13	3,12
ALL	196550	2,89*	2,84*

the tables 1 and 2 are statistically significant. Web sites about weather forecast are evaluated as the most credible for both datasets. On first sight it may look a little bit counterintuitive but it may be explained with a limited expectation. People know that weather forecasts are only a scientifically supported guess and do not validate them (although it is very easy).

Another plausible explanation with evidence in the data is the view that trustworthiness correlates with the amount of intentional deception possible in a category. People tend not to believe cultural and political communication, have little trust in legal and crime-related websites. In all these types of communication deception occurs intentionally. Perpahs, the most striking evidence of importance of the intentional trust component is the highest trust put in weather forecast websites. The content of weather forecasts is related to the reality in a limited way as the accuracy of weather models is still far from being perfect. However, the "deception" of weather forecasts is unintentional. Moreover, it is likely that no spam or other dubious activity takes place around weather forecasts, as opposed to politics and crime categories, for instance.

A comparison between trustworthiness ratings for DMOZ and Alexa datasets reveals some similarities but also many big differences. Categories like law & crime and health are among less credible in both datasets (Science & technology next to weather forecasts are very trustworthy) but for others, like sport, religion or business, rating are missing a common pattern. Business is the least credible category for DMOZ catalogue and in the middle for the Alexa. Although the category for both datasets is the same, there are huge differences concerning business web sites between the DMOZ catalogue and the Alexa. Putting web sites into the DMOZ catalogue requires just filling out the form. Being in the Alexa is reserved only for established companies with a considerable number of visitors.

The same patterns are not clearly noticeable in articles from Wikipedia. The first problem is that some categories are not represented in this dataset (or there are only a few pages in a category e.g. one page in weather forecast category). Second, subjects of articles in the Wikipedia are not related one to one with web sites either from the DMOZ catalogue or the Alexa. On the other hand, in AFT dataset, similarly as in the Alexa, web sites about sport are among the most and law & crime among the least trustworthy.

Spearman rang correlation calculated for average trustworthiness for categories between the Alexa and the DMOZ is 0.517. Even higher is the correlation between the AFT dataset and the Alexa – 0.536. Positive correlation shows that web site category can be used at least as provisional filter which can help to identify web sites for an in-depth study (with other methods e.g. NLP).

3.3 Search for information in Estonian

European Social Survey (ESS) is a rigorous cross-national attitude survey that tries to trace changes in social behaviors in time and spots differences between countries. The biggest advantage of ESS is that results from many countries can be compared, because there has been used the same methodology. Results are freely accessible[3] . Data from the round five (finished in 2010) has been used in this paper.

Spearman rank correlation between average trustworthiness for countries top level domains and percentage of people in population who selected values from 8 to 10 as an answer for a question Most people try to be fair is 0.264. Figure 3 presents average results of WOT data, aggregated for each country, plotted against average results of trust question of the ESS 2010. The question, a 10-point scale depicting whether people do not trust others (0) or are trustful (10).

High levels of ESS trust can be found in northern Europe countries such as Finland, Sweden, but also Netherlands, Estonia and Switzerland. From this group, only web contents of Estonia and to the lesser degree Sweden, are also highly trusted according to WOT. Central and eastern Europe countries such as Poland, Germany, France, Czech Republic and Slovakia are similar in their WOT and ESS trust levels.

4 Discussion

Trustworthiness of the AFT (wikipedia) data is on average decreasing. This observation has limited relevance for the Wikipedia and may be related to various phenomena related to artices life cycle. In DMOZ, differences between time points are not significant and in Alexa overall trustworthiness did not change. Although a specific subset of web sites (20% most credible) trustworthiness increase may be observed. As it is visible for the Alexa and the DMOZ catalogue but not for the AFT dataset it may be a consequence of the fact that in the

[3] http://ess.nsd.uib.no/ess/round5/

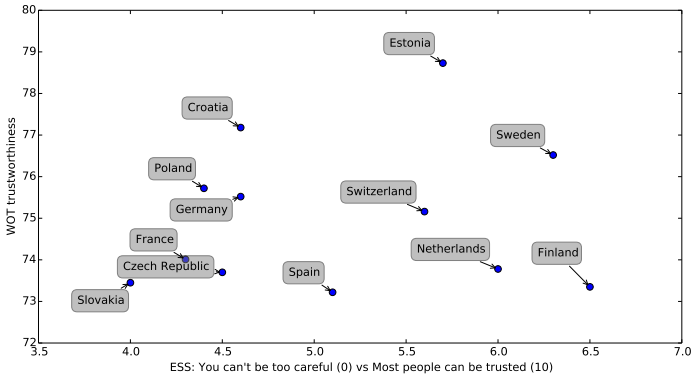

Fig. 1. Nation-level trust data: European Social Survey (ESS) and Web of Trust (WOT).

WOT (in opposition to Wikipedia) users know an existing evaluation before they evaluate themselves. This so-called anchor effect may particularly strong for very credible web sites. The conclusion to draw from these observations is that it is very likely that the internet, overall, is not getting more trustworthy. This conclusion needs to be treated with certain caution: the datasets analysed represent large and important portions of the web, nevertheless still miss certain spots of communication, such as social media or microblogs falling out of the scope of this paper. Assuming that the observation is nevertheless true, we may further hypothesize that the web has reached its trust limits, alternatively no mechanisms that could successfully raise the level of trust towards web contents were introduced.

5 Conclusion

When the temporal aspect is considered, results reveal that the levels of trust are either constant in time (Alexa, DMOZ) or slightly decrease (Wikipedia). However, certain types of categories do become more trustworthy. Also, the trust to the most trustworthy websites tends to increase, which leads to the conclusion that overall variability of WOT trust levels increases.

Trust levels tend to exhibit certain patterns between topics or categories. Perhaps the most surprising finding is that the most trustworthy websites are weather forecasts. A possible explanation of this fact, as well as other patterns of trust differences between categories, is that people relate trust, at least partly, to intentional rather than factual dimension.

The paper investigated also nation-wide trust measurements of the European Social Survey to find them to be only partially reflected in the WOT data.

Reported results have strong foundations because size of the datasets is significant. The conclusions are based on comparing millions of webpages, identified

with well-established data sources such as Alexa and DMOZ, and backed by many individual trust evaluations of each website.

Some of the reported results demand further investigations and pose many research questions. For instance, it is not entirely clear what are the reasons of trust decrease in the case of Wikipedia pages. In the case of other data sets, why are trust levels constant and why is trust variability increasing. Perhaps, some of the answers could be formulated by measuring the influence of linguistic and textual content of webpages on their trust levels. Another issue that remains to be addressed is how to measure the state of trust in social media and microblogs in a way comparable to WOT ratings for the web.

References

1. Sternthal, B., R. Dholakia, and C. Leavitt, The Persuasive Effect of Source Credibility: Tests of Cognitive Response. J. of Consumer Research, 1978. 4(4): p. 252-260.
2. Hovland, C.I. and W. Weiss, The Influence of Source Credibility on Communication Effectiveness. Public Opinion Quarterly, 1951. 15(4): p. 635-650.
3. Pornpitakpan, C., The Persuasiveness of Source Credibility: A Critical Review of Five Decades' Evidence. J. of Applied Social Psychology, 2004. 34(2): p. 243-281.
4. Gaziano, C. and K. McGrath, Measuring the Concept of Credibility. Journalism Quarterly, 1986. 63(3): p. 451-462.
5. Sobel, J., A Theory of Credibility. Rev. of Economic Studies, 1985. 52(4): p. 557-573.
6. Tseng, S. and B.J. Fogg, Credibility and computing technology. Commun. ACM, 1999. 42(5): p. 39-44.
7. Fogg, B.J. and H. Tseng, The elements of computer credibility, in Proceedings of the SIGCHI conference on Human Factors in Computing Systems1999, ACM: Pittsburgh, Pennsylvania, USA. p. 80-87.
8. Patzer, G.L., Source credibility as a function of communicator physical attractiveness. Journal of Business Research, 1983. 11(2): p. 229-241.
9. Fogg, B.J., et al., 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 experiences2003, ACM: San Francisco, California. p. 1-15.
10. Schwarz, J. and M. Morris, Augmenting web pages and search results to support credibility assessment, in Proceedings of the SIGCHI Conference on Human Factors in Computing Systems2011, ACM: Vancouver, BC, Canada. p. 1245-1254.
11. Morris, M.R., et al., Tweeting is believing?: understanding microblog credibility perceptions, in Proceedings of the ACM 2012 conference on Computer Supported Cooperative Work2012, ACM: Seattle, Washington, USA. p. 441-450.
12. Fogg, B.J., Prominence-interpretation theory: explaining how people assess credibility online, in CHI '03 Extended Abstracts on Human Factors in Computing Systems2003, ACM: Ft. Lauderdale, Florida, USA. p. 722-723.
13. Olteanu, A., et al., Web credibility: features exploration and credibility prediction, in Proceedings of the 35th European conference on Advances in Information Retrieval2013, Springer-Verlag: Moscow, Russia. p. 557-568.
14. Dai, N. and B.D. Davison, Freshness matters: in flowers, food, and web authority, in Proceedings of the 33rd international ACM SIGIR conference on Research and development in information retrieval2010, ACM: Geneva, Switzerland. p. 114-121.