supporters of some charity organization, plus their circles of friends. The motivation to participate in the study was the promise of a donation to the charity organization. Moreover, the participants were not obliged to leave demographic information [6]. In our study, instead, we set up a contest for the participants to obtain a more balanced sample of participants. Demographic data and some additional user-related information was gathered as well.

The work of Schwarz and Morris [10] or the work of Yamamoto and Tanaka [14] use the approach of increasing the prominence of selected web content features in order to enable better credibility evaluations, e.g. by augmenting the search results with credibility information.

Building an automatic or semi-automatic credibility classifier is another approach applied in other works. Castillo et al. use that approach to asses tweets [1], and Chiao-Fang et al. deal with ranking comments from social web [3]. The work of Sondhi et al. [11] is an approach of building a semi-automatic credibility classifier based on site's features. Similar to our study, Sondhi focused on the credibility of Websites from the medical domain.

Finally, to evaluate the credibility using Web content features. some works focus solely on using the link structure and trust network [2,7].

3. THE STUDY

The first Reconcile study on evaluating Web credibility of a set of Polish Webpages was carried out in cooperation with IIBR[2], a company specialized in Web-based social opinion polling. The collaboration with IIBR allowed us to control the respondents group. As such, the group was diversified with respect to the factors that can influence subjectivity of user ratings.

The corpus of web pages to be evaluated was gathered manually. It spans various topical categories, including topics perceived as controversial.

1. Hormonal contraception
2. Aspartame
3. Breast feeding
4. Cannabis
5. Chemotherapy
6. Oral chemotherapy
7. Dukan diet
8. GMO
9. Homeopathy
10. Immunity
11. Diet during cancer treatment
12. Children bathing
13. Money investment
14. Picky eaters
15. "Sesja" diet supplement
16. Targeted therapy
17. Vitamin B17

The respondents evaluated the archived versions of the gathered pages. The archiving process included also the dynamically generated content (e.g., advertisements), so that respondents were viewing the "snapshotted" version of a page from a certain point of time.

The selection of respondents was carried out in a way that assured diversity with respect to socio-economic status and Internet efficacy of the respondents. 1,503 respondents (out of 2532 that were invited to participate) submitted their full evaluations.

3.1 Internet efficacy

Apart from credibility evaluations, the respondents were also asked to fill an additional questionnaire aimed at identifying their psychological traits and Internet efficacy (including how often and to what extent the respondent is using the Internet). We have analyzed a number of secondary sources and identified 9 Internet activities that are performed relatively rarely. Respondents were asked whether they perform the following activities at least once a month:

- creating and publishing own texts (e.g. blog, Wikipedia entry), graphics, music, photos, videos etc.
- creating or modifying WWW site (e.g. code changes, presentation changes)
- gathering materials/information required to learn or work
- gathering information for dealing with administrative matters
- buying products or services via Internet
- selling products or services via Internet
- commenting on blogs, writing on internet forums/discussion groups
- writing about/reviewing products or services
- using mobile banking

The Internet experience of the respondents was measured using the Web-Use Skill Index [8]. The Web-Use Skill Index is based on a list of 10 internet-related terms. Respondents were asked to rate their level of understanding of these terms on a 1-to-5 point Likert scale. The user's score on this scale is given by the sum of points of all the evaluations, and can take on any value between 10 and 50.

- Advanced search
- Tagging
- PDF
- Spyware
- Wiki
- Weblog
- JPG
- Cache
- Malware
- Phishing

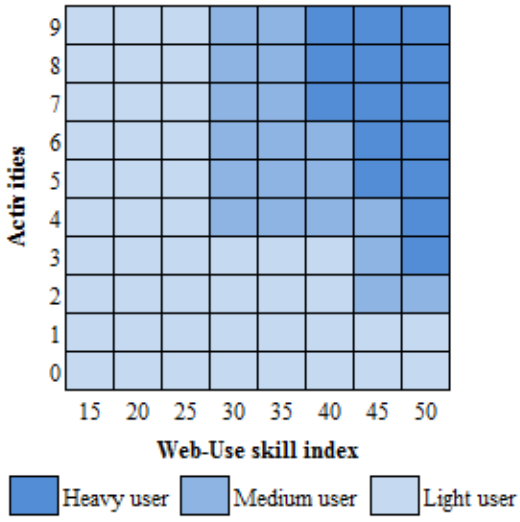

Figure 1. Internet efficacy groups based on Internet activities and Web-Use skill index

These two sets of questions enabled us to categorize the respondents into groups of heavy, medium and light Internet users. Figure 1 shows the classification of the respondents into the mentioned groups, based both on the scores achieved while answering questions related to Web-Use Skill Index and on the frequency of performing Internet activities. The horizontal axis represents the number of activities performed at least once a month, while the vertical axis is a sum of points representing familiarity with Internet-related terms.

[2] Interaktywny Instytut Badań Rynkowych, http://www.iibr.pl/

3.2 Study duration and design

The study took place from 9th to 25th September 2012 (22 days). During the study, 1,503 respondents submitted 4354 evaluations, and 154 web pages from 17 categories were evaluated, averaging 28 evaluations per page. Every user taking part in the study evaluated the same archived versions of the pages. Submitted evaluations were independent, as the users did not see the credibility scores submitted by others. The diversity of the respondents, combined with the independence of their ratings, is the most important criteria that enabled us to use the Wisdom of crowds approach [13]. The high number of evaluations per page makes this approach particularly effective.

4. ANALYSIS OF STUDY RESULTS

4.1 Basic distribution of credibility responses

The credibility of a assessed page was measured using a 5-point Likert item as follows: 1: completely not credible; 2: mostly not credible; 3: somewhat credible, although with major doubt; 4: credible, with some doubt; 5: completely credible; do not know. However this question in the questionnaire was not presented as a scale or slider, but as a dropdown list.

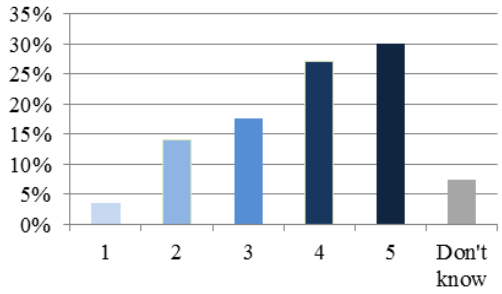

Figure 2. Distribution of all credibility ratings

The figure 2 shows the distribution of all credibility evaluations submitted by the respondents in all categories of pages. This distribution visibly has a negative skew and the values of credibility score are concentrated on the right side of the scale. Such a phenomenon can be due to biased responses that where submitted. We manually tried to balance the corpus of the pages for assessment in order to achieve equal number of credible and not credible webpages. This leaves space for an error and possibility that corpus eventually was not fully balanced.

4.2 Verification of hypotheses regarding subjectivity

During the study several socio-economic data of the respondents was gathered. Responses among different groups of gender, education and Internet experience have shown statistically significant differences. The discovered differences between respondents' groups by their features are taken as a sign of possible subjectivity--however it has only a slight impact on the rating distributions.

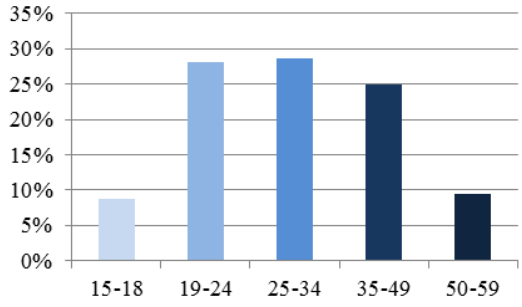

Figure 3. Respondents by age category

4.2.1 Age

The preparation of the respondents sample was done with caution also to include the middle-aged respondents and higher. The distribution of age among the respondents is shown on the figure 3.

As it is shown the respondents were divided into 5 age categories. The distributions of the credibility scores in each age group do not show any apparent differences. The Kruskal-Wallis test also does not let us to reject the null hypothesis on equality of those distributions ($p<0.1026$). Therefore we assume that age is not a factor leading to different credibility ratings.

4.2.2 Gender

Slight majority of the respondents were of female gender, as shown on the figure 4. The differences in submitted credibility scores among the genders were statistically significant ($p<0.0001$). Males seem to give less extremely positive credibility scores while most frequent female credibility submissions were 5, which is "completely credible", see figure 5.

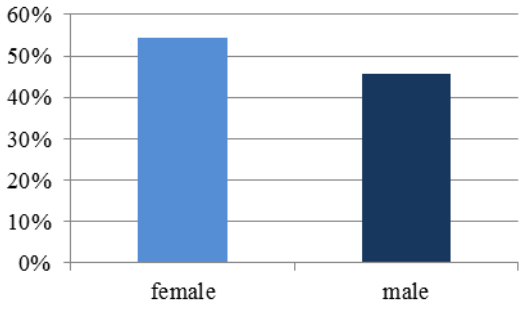

Figure 4. Distribution of genders among the respondents

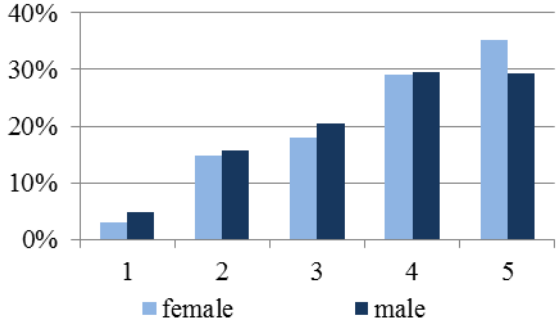

Figure 5. Credibility ratings by genders

4.2.3 *Education*

The differences in credibility scores submitted by the respondents of different education categories are also visible and statistically significant ($p<0.0001$).

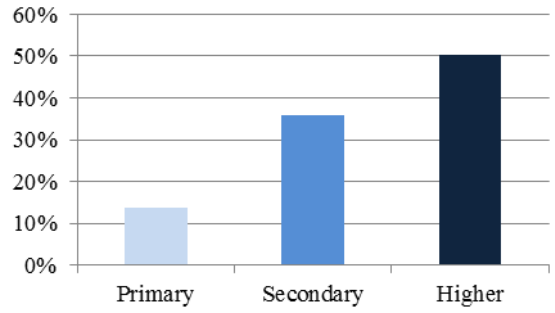

Figure 6. Distribution of education categories among respondents

Half of the respondents reported that they have higher education. The higher the education level was the more female respondents were present in this group, which can be seen in figures 6 and 7. The general conclusion from distribution of credibility scores by education level, see figure 8, is that respondents group of higher education level has more evenly distributed ratings. The lower the education level the more respondents tend to concentrate their credibility evaluation on the right side of the scale.

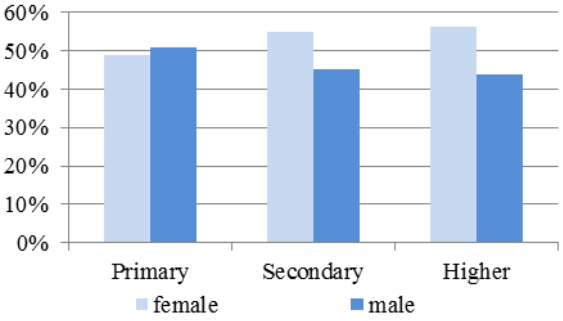

Figure 7. Education categories by genders

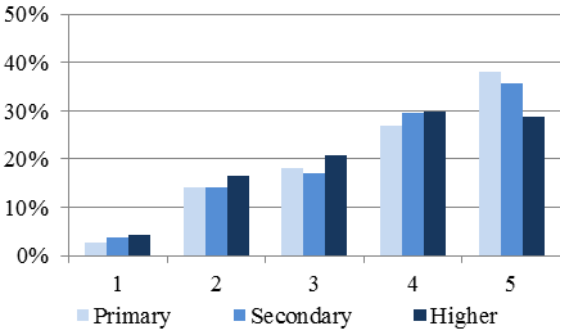

Figure 8. Distribution of credibility ratings by education categories

4.2.4 *Internet experience*

Using Web-use skill index and additional questions regarding frequency of using the Internet the majority of the respondents were classified as heavy users (>30%), see figure 9.

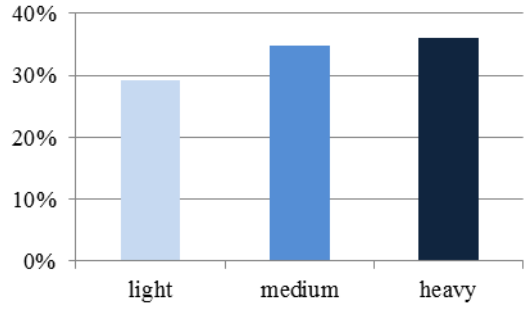

Figure 9. Distribution of Internet efficacy levels

Most of those heavy users were of male gender. The frequency of primary education respondents decreased with the ascending level of Internet efficacy. Again statistically significant differences in credibility evaluations among experience groups were observed ($p<0,0017$). Light users tend more to use the extreme positive scores – completely credible on contrary to groups "medium" and "heavy", see figure 12. The group of heavy users had the most evenly distributed credibility ratings.

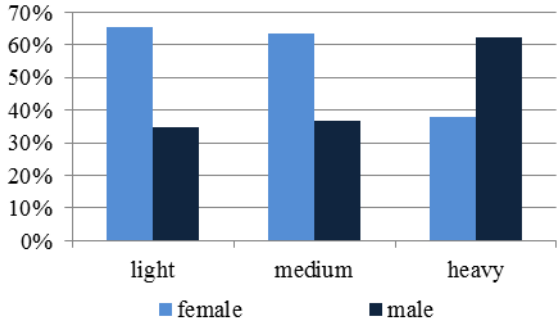

Figure 10. Internet efficacy levels by genders

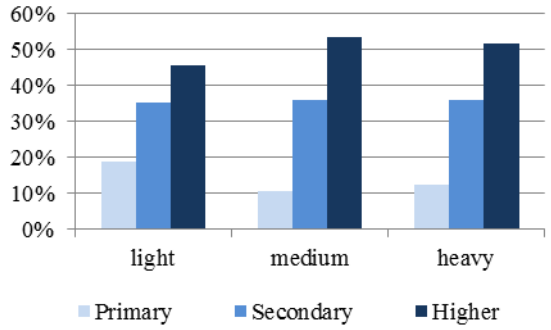

Figure 11. Internet efficacy levels by education categories

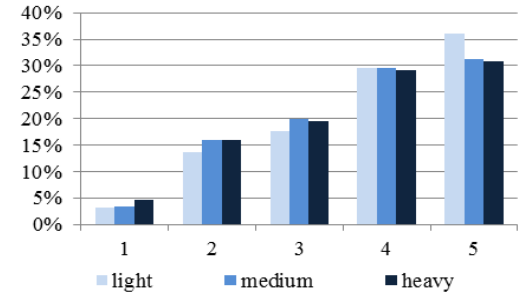

Figure 12. Distribution of credibility ratings by Internet efficacy levels

4.2.5 Psychological factors

Psychological traits were measured using scales from International Personality Item Pool. Those traits were: need for cognition[3] and trust[4].

Table 1. Sample questions measuring 'need for cognition' and 'trust' respondent traits

Need for cognition	
Positive	e.g. *I like to solve complex problems*
Negative	e.g. *I avoid philosophical discussions.*
Trust	
Positive	e.g. *I believe that others have good intentions*
Negative	e.g. *I am wary of others.*

For both traits, factor analysis revealed two factors related to the way the questions were asked. First factor is composed of positive and the second of negative statements.

Scales constructed for need for cognition based on these two factors are significantly positively correlated with each other (cor=0.354, $p<0.0001$). They do not correlate with credibility evaluations (on population of credibility evaluations). However, evaluations of those, who have very low need for cognition, do not use the low end of the credibility scale. Let us define overrating of a page as assigning an evaluation score at least one category higher than median evaluation of the given page. Using such definition we can say that there is significant negative relationship between high need for cognition and tendency to overrate websites (cor=-0.0875; $p<0.0174$).

Scales constructed for trust are also significantly positively correlated with each other (cor=0.2996, $p<0.0001$). There is weak but significant correlation between positive measure of trust and credibility evaluations (cor=0.0421, $p<0.007$). It indicates that evaluations given by people with greater willingness to trust are expected to be slightly higher. High credibility ratings are related with higher average score on positive trust scale and low credibility ratings are related with lower average score on negative trust scale.

The constructed measures of need for cognition and of trust were weakly related with credibility evaluations and should be validated further.

4.3 Verification of hypotheses regarding bias

We establish the ground truth about examined web pages using both the Wisdom of crowds approach and the median of the ratings (when sufficient number of evaluations is gathered). Under such conditions, it is reasonable to check the outcome of the study against the experts' ratings.

After gathering the study data, we managed to invite several experts who rated the same web pages assessed by the respondents. Unfortunately, the number of experts willing to help was not sufficient to examine all the pages. Experts evaluated 119 of the 154 pages from the study, mainly from medicine related topics. The group of experts that cooperated consisted of 7 medical doctors, 3 midwives and 1 investment broker, who evaluated the pages from categories related to their profession (excluding the following categories: Cannabis, Dukan diet, GMO).

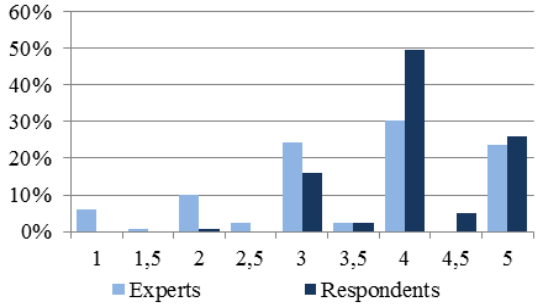

Figure 13. Median ratings of experts and respondents

Every pair or three of experts were evaluating pages from 3 categories. The average percentage agreement among the experts' ratings was 80%. When experts did not reach consensus of a page, their ratings differed maximally by one category. At the moment of writing this article, the goal of reaching the consensus on every assessed page was not met, and another round of expert evaluations is needed. So far, only the median experts' ratings were used.

When compared to experts' ratings , respondents' median ratings are far more concentrated on the right side of the credibility scale, as shown in Figure 13. Expert scores are more evenly distributed over the whole scale. While respondents' medians tend to be mainly positive, the experts' credibility medians also consist of extremely low ratings (e.g., 1, 2).

We performed agreement analysis using Cohen's kappa. For median ratings agreement, simple kappa reached 0.02, while weighted kappa (considering the credibility scale as ordinal) reached 0.2. This level of agreement can be considered as "slight" [9]. Together with calculating Kappa measures, we prepared the contingency table with the median credibility scores of pages, evaluated both by respondents and experts (Table 2). The diagonal of the table represents the frequency of matching median scores among respondents and experts. In that case 26,9% of pages has equal median credibility scores as well from experts as from respondents. What is also visible in the table is that, in comparison to experts' median scores, the respondents median scores are higher. For example, pages with experts' scores equal to 3 constituted 24% of all pages. Respondents evaluated those pages higher, because only 4% percent of pages got median 3 from both experts and respondents, when 17,7% of all pages got a respondents' median of 4 vs. median of 3 from the experts.

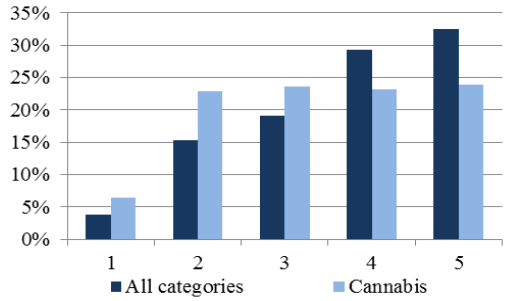

Figure 14. Respondents credibility ratings in all categories versus "Cannabis" category

[3] CHS: Cacioppo & Petty, 1982, http://ipip.ori.org/newCHSKey.htm#Need-for-Cognition

[4] NEO: A1, http://ipip.ori.org/newNEOKey.htm#Trust

The distributions of respondents' ratings are generally negatively skewed. The respondents seem to choose the right, positive part of the scale. This can be due to the acquiescence bias that makes the users most likely to consider the page "credible". But this is not always the case, as this transition to the right end of scale depends also on the thematic category. In fact, users could not achieve a consensus on the pages belonging to categories perceived as controversial. Figure 14 shows the difference between the distribution of ratings in all categories and ratings from the "Cannabis" category. In the distribution of "Cannabis" related evaluations, the characteristic skew is no longer visible. This can be taken as a sign that the users do not choose the credibility ratings randomly, but still they avoid negative ratings. Rating 1 (completely not credible) is the least frequent category of rating.

Table 2. Contingency table of pages median credibility scores of respondents and experts

			Table of experts medians by respondents medians									
			respondents median									Total
			1	1.5	2	2.5	3	3.5	4	4.5	5	
experts median	1	%	0,0%	0,0%	0,8%	0,0%	5,0%	0,0%	0,0%	0,0%	0,0%	5,9%
	1.5	%	0,0%	0,0%	0,0%	0,0%	0,0%	0,0%	0,8%	0,0%	0,0%	0,8%
	2	%	0,0%	0,0%	0,0%	0,0%	0,8%	0,8%	6,7%	0,0%	1,7%	10,1%
	2.5	%	0,0%	0,0%	0,0%	0,0%	0,8%	0,0%	1,7%	0,0%	0,0%	2,5%
	3	%	0,0%	0,0%	0,0%	0,0%	4,2%	0,0%	17,7%	1,7%	0,8%	24,4%
	3.5	%	0,0%	0,0%	0,0%	0,0%	0,0%	0,0%	2,5%	0,0%	0,0%	2,5%
	4	%	0,0%	0,0%	0,0%	0,0%	5,0%	1,7%	10,9%	0,8%	11,8%	30,3%
	4.5	%	0,0%	0,0%	0,0%	0,0%	0,0%	0,0%	0,0%	0,0%	0,0%	0,0%
	5	%	0,0%	0,0%	0,0%	0,0%	0,0%	0,0%	9,2%	2,5%	11,8%	23,5%
Total		Freq.	*0*	*0*	*1*	*0*	*19*	*3*	*59*	*6*	*31*	*119*
		%	0,0%	0,0%	0,8%	0,0%	16,0%	2,5%	49,6%	5,0%	26,1%	100%

5. Conclusions and future work

Basing on our study, the hypothesis of strong subjectivity of users' evaluations has been found not to hold. While the subjectivity of Web credibility evaluations due to the considered factors is statistically significant, it has only a slight impact on the rating distributions. However, the hypothesis that credibility is subject to strong bias is supported by our results. The distributions of Web credibility ratings are shifted towards the positive values. This shift could be due to the overall high quality of examined content, despite our efforts to diversify the quality of selected Web pages. However, a comparison of the user ratings with expert ratings shows that this is not the case. Distributions of expert ratings are much more evenly positioned on the evaluation scale.

Such effect can be due to an information bias of some kind that affected the respondents. We suspect that one of the main causes is represented by the acquiescence bias. The analysis of the responses from the perspective of the psychological features reveals that high need for cognition and trust can lead to overrating the examined pages.

Using the experience gathered while preparing and running the presented study, an extended version of the study was planned. As this article is being redacted, the new study is already being carried out. The new study will cover only English-language Web pages, and a greater number of respondents will be asked to participate. Moreover, a bigger sample of archived pages for evaluation will be balanced in order to cover an equal number of credible and not credible Web pages. Differently from the conditions in the presented study, a balanced sample will allow us to compare the credibility ratings results not only with experts' ratings, but also with sound assumptions about the Web pages corpus.

6. REFERENCES

[1] Castillo, C., Mendoza, M., and Poblete, B., 2011. Information credibility on twitter. In *Proc. of WWW*.

[2] Cavarlee, J. and Liu, L., 2007. Countering Web spam with credibility-based link analysis. *PODC 2007*, 157--166.

[3] Hsu, C., Khabiri, E., and Caverlee, J., 2009. Ranking Comments on the Social Web. In *Proc. of CSE*, Volume 04, pages 90-97.

[4] Diaz, J. A., Griffith, R. A., Ng, J. J., Reinert, S. E., Friedmann, P. D. and Moulton, A. W., 2002. Patients' Use of the Internet for Medical Information. *Journal of General Internal Medicine*, volume 17: pages 180–185.

[5] Fogg, B. J., 2003. Prominence-interpretation theory: explaining how people assess credibility online. In *Proc. of CHI*.

[6] Fogg, B. J., Soohoo, C., Danielson, D. R., Marable, L., Stanford, J., and Tauber. E.R., 2003. How do users evaluate the credibility of web sites?: a study with over 2,500 participants. In *Proc. of DUX*.

[7] Gyöngyi, Z., Garcia-Molina, H, Pedersen. J., 2004. Combating web spam with trustrank. In *Proc. of VLDB*.

[8] Hargittai, E., and Hsieh, Y., P., 2011. Succinct Survey Measures of Web-Use Skills, *Social Science Computer Review,* February 2012 30: 95-107, first published on February 28.

[9] Landis, J. R., Koch, G. G., 1977. The measurement of observer agreement for categorical data. Biometrics 33:159-174.

[10] Schwarz, J., and Morris, M., 2011. Augmenting web pages and search results to support credibility assessment. In *Proc. of CHI*.

[11] Parikshit Sondhi, V. G. Vinod Vydiswaran, and ChengXiang Zhai, 2012. Reliability prediction of webpages in the medical domain. In *Proc. of ECIR*.

[12] Surowiecki, J., The Wisdom of Crowds, Anchor, 2005.

[13] Wagner, C., and Vinaimont, T., 2010. Evaluating the wisdom of crowds. In *Proceedings of Issues in Information Systems*, volume XI, no. 1, pages.724 -732.

[14] Yamamoto, Y., and Tanaka, K., 2011. Enhancing credibility judgment of web search results. In *Proc. of*