Reliability Prediction of Webpages in the Medical Domain

Parikshit Sondhi, V.G.Vinod Vydiswaran, and ChengXiang Zhai

Department of Computer Science
University of Illinois at Urbana-Champaign, Urbana, IL 61801
{sondhi1, vgvinodv, czhai}@illinois.edu

Abstract. In this paper, we study how to automatically predict reliability of web pages in the medical domain. Assessing reliability of online medical information is especially critical as it may potentially influence vulnerable patients seeking help online. Unfortunately, there are no automated systems currently available that can classify a medical webpage as being reliable, while manual assessment cannot scale up to process the large number of medical pages on the Web. We propose a supervised learning approach to automatically predict reliability of medical webpages. We developed a gold standard dataset using the standard reliability criteria defined by the Health on Net Foundation and systematically experimented with different link and content based feature sets. Our experiments show promising results with prediction accuracies of over 80%. We also show that our proposed prediction method is useful in applications such as reliability-based re-ranking and automatic website accreditation.

1 Introduction

As the dependence on online content increases, it is important to also know what content is reliable. This is especially true in the medical domain, where patients try to find much of the knowledge online.There are some non-profit organizations, such as Health on Net Foundation (HON, *http://www.hon.ch/*) and Quackwatch (*http://www.quackwatch.com/*), that rate websites based on how reliable they believe the website is. They do it by manually looking through a site to determine if it satisfies some conditions, such as citing references, attributing articles to experts, and so on. This task is highly effort-intensive and hence, cannot scale up well to keep pace with the rapid growth of medical information on the Web.

In this paper, we want to explore if it is possible to automate this process of assessing the reliability of a webpage in the medical domain. As a first step in studying this novel problem, we focus on classifying webpages, to differentiate good informational pages from other less reliable ones. We cast the problem in a supervised learning setup and study the feasibility of learning to classify pages as reliable or not. We propose a variety of features defined based on both the content of a webpage and other information such as links and study how different features help in this classification.

A big challenge in studying this prediction problem is that no existing test collection is available for evaluation. To solve this challenge, we have created a labeled test set by leveraging the websites accredited by the Health on Net (HON) Foundation. We have also proposed appropriate measures to quantitatively evaluate this task.

Evaluation results on the dataset show that we are able to achieve an overall accuracy of over 80% in prediction. Thus, the proposed method can help significantly reduce manual labeling efforts currently in practice. Experiments also show that our prediction method works better than Google PageRank alone in reliability prediction and can be used for reranking search results based on reliability.

2 Related Work

The quality of medical information on the Web has attracted considerable attention from medical domain researchers. Matthews et al. [15] evaluated a set of 195 webpages pertaining to alternative cancer treatments and found nearly 90% have atleast one flaw. Related studies by Marriott et al. [13] and Tang et al. [18] also concluded that medical information quality on the Internet was variable.

The first attempt to automatically identify high quality health information on the Web was published in 1999, by Price and Hersh [16] who developed a simple rule based system which perfectly separated desirable and undesirable documents using a heuristic scoring function. However their dataset, comprising of only 48 documents, was too small to draw concrete conclusions on either the characteristics of medical webpages or the discriminative power of features. In recent related attempts, Aphinyanaphongs and Aliferis [2] used text categorization models for classifying pages discussing unproven treatments, Wang and Richard [21] used a regular expression based heuristic approach for measuring information quality and Gaudinat et al. trained classifiers to predict each of the Health on Net reliability criteria (e.g. presence of author names) using content based features [7] and only URL based features [6].

Some approaches for identifying low quality webpages have focused mainly on detecting spam webpages through link structures [9, 4], the most popular being Page Rank [5]. Our goal differs from spam detection approaches [3, 1, 22], since we attempt to directly assess reliability of legitimate webpages and analyze the utility of our learnt models in real applications. Other related works include [12, 17, 14].

3 Notion of Medical Reliability

For identifying reliable pages, we define our reliability guidelines based on the eight HONcode Principles[1] . These principles are generally accepted by experts in medical community worldwide (e.g. [7]). We assume the reliability of a webpage to be a binary value (1 for reliable and 0 for unreliable), judged based on

[1] *http://www.hon.ch/HONcode/Conduct.html*

the HONcode. In reality, reliability may have multiple degrees, but similar to relevance judgments in information retrieval, assuming a binary notion of reliability makes it easier to create judgments. Further, it is not clear what principles an intermediate class ("moderately reliable") should satisfy. Manually defining criteria for such additional classes would lead to the problem of evaluating the criteria themselves. Other potential formulations, such as generating a real valued reliability score or estimating a reliability probability, also run into the same definition and evaluation hurdles. As a first step in exploring this problem, we thus restrict our study to a binary notion of reliability.

It can also be argued that even if a webpage is deemed reliable based on the HONcode principles, the content may still be inaccurate; e.g. an article based on (and citing) an inaccurate published research study. At this point, we must distinguish between *reliability* and *veracity*. Being able to extract potential facts from text and judge their veracity is not the goal of this work, but has been explored elsewhere (e.g. [20]).

4 Supervised Learning for Reliability Prediction

We cast the problem of reliability prediction as a supervised binary classification problem. In a supervised setting, reliability of a webpage is defined as a binary function over computable features that model the abstract HONcode principles. We present a wide range of features in Sec. 4.1 and learn a Support Vector Machine classifier [19, 11] to label webpages as reliable or not.

Once the reliability of individual webpages is determined, the reliability of a website W is computed as the fraction of webpages in W found to be reliable. Thus the reliability of a website is not binary, but a real value. Our formalism fits well with the nature of the Web, where we often find a mix of reliable and unreliable pages in a website. For example, a commercial website may have some reliable pages with information about diseases, and other less reliable pages that advertise their products. Other examples include sites where both doctors and laypersons may post articles, or where some articles are properly referenced while others are not. In such cases, a binary classification of websites is insufficient to capture the diversity of the Web.

4.1 Features

In this section, we provide a detailed description of our proposed features. Apart from the PageRank-related feature set, all other features are calculated over individual pages.

1. Link-based Features: Links can often give a good indication on the type of webpage. For example, a reliable site is likely to contain a large number of internal links, whereas a small unreliable site is more likely to be dominated by external links of advertisements. We also defined two boolean features based on the presence of contact and privacy policy links, that are inspired by the HON reliability criteria. The absence of such information usually means the website is less reliable. The five link-based features we defined are: (a)Normalized count of internal links($\frac{\#(\text{internal links})}{Z_1}$), (b)Normalized count of external links($\frac{\#(\text{external links})}{Z_1}$), (c)Normalized count of total links($\frac{\#(\text{total links})}{Z_1}$), (d)Presence of a Contact Us Link and (e)Presence of a Privacy Policy Link. The first three are normalized count features, while the last two are binary features.

Classification models tend to perform well when all the features have nearly similar range of values. Since the number of links often vary considerably across webpages. We normalize the first three features by a sufficiently large factor Z_1. For our experiments, we set the value of $Z_1 = 200$, by observing a random sample of the dataset. (Normalizing by the maximum feature values in the dataset doesn't necessarily help as we don't know the range of values in the unseen test examples).

2. Commercial Features: Commercial interests often indicate unreliability. For example, information about a drug on a company's website may be commercially biased, and hence unreliable. To estimate if there is a commercial bias involved, we define two features based on the number of commercial keywords and commercial links:(a)Normalized count of commercial links and (b)Normalized frequency of commercial keywords in the webpage. To compute these features, we manually compiled a list of commercial words, such as *buy, sell, cheap, deal, free, guarantee, shop, price*, etc.

3. PageRank Features: PageRank provides an indication of relative "importance" of a website and has been successfully used to improve Web search performance. Moreover, unreliable sites are more likely to link to low PageRank-ed sites as compared to the reliable ones. We generated six features the first feature below represents the PageRank of the website to which the webpage belongs. The next five features are essentially a five-point representation of PageRank values of all external links [8]. We used Google PageRank(via WWW::Google::PageRank perl package) to get the PageRank values in $[0, 10]$, and we normalize it by 10 to get the values in $[0, 1]$.

(a) *Normalized internal PageRank*: $PR_{int} = \frac{PageRank(\text{parent website})}{10}$

(b) *Normalized external PageRank features* ($ExtPR$): We computed the PageRank of all websites linked from the webpage, and derived 5 features based on the five-point summary (mean, minimum, maximum, and first and third quartiles) of the values.

4. Presentation Features: Authoritative and reliable websites often seem to clearly present information, while the unreliable ones are usually cluttered with advertisements. With this idea, we define two simple presentation related features. We use elinks (*http://elinks.or.cz/*), without the frames option, to generate a text version of the webpage. Webpages cluttered with a large number of advertisements and poor presentation, when converted to text, tend to have a large number of blank lines between small scattered chunks of text. Consequently, the first feature, *Percentage of Coherent Text* ($\%CT$) is the fraction of document lines that do not have a blank line on either side. The second feature, *Percentage of Spread-out Text* ($\%ST$) is the opposite (i.e. $1 - \%CT$).

5. Word Features: The textual content and the writing style used in a webpage are usually good indicators of its reliability. For a document D, each unique word is an independent feature taking the normalized word frequency ($\frac{\#(w,D)}{\max_{w' \in D}(\#(w',D))}$) as its value.

5 Test Set Construction

Next, we wanted to build a balanced dataset that was representative of the typical webpages an Internet user might encounter. For the positive set, we used 32 medical websites that had been accredited by the HON staff during Sep–Oct 2009.[2] We applied our reliability criteria on pages from these sites and randomly selected 180 reliable pages. Since the websites had already been thoroughly reviewed and certified by experts, the task of finding reliable pages was simplified. We removed the HON seal from these pages at the time of feature generation.

For the negative set, however, we could not use this approach, since the HON website does not provide information on websites that failed the certification process. So, the negative set had to be built by directly searching for unreliable pages on the Web. We initially considered several "simple" approaches for this purpose. Intuitively, it is relatively easy to find a large number of unreliable websites by simply searching for queries like "disease name"+"what your doctor doesn't want you to know" or "disease name"+"miracle cure", etc. In addition, it is easy to find websites that promote treatments banned by the FDA [2], or the ones criticized on Quackwatch. However, it is important to ensure topical overlap between the reliable and unreliable sets of documents, so as to prevent a simple classifier from discriminating documents based solely on topic-specific keywords. Similarly, simply picking unreliable pages from obscure websites could bias the classifier to choose Page Rank as the most discriminating feature.

Therefore, for the unreliable set, we first compiled a list of topics (keywords representing diseases/conditions), covered by the 32 reliable websites. We then searched Google for (a) the topic keyword, (b) the topic keyword + "treatment", and (c) the topic keyword + "treatment" + a randomly chosen keyword from {"*cure*", "*miracle*", "*latest*", "*best*"}. For each query, we manually analyzed the webpages appearing in both the general results and advertisements, and ultimately selected 180 webpages from 35 websites that failed comprehensively on one or more of our reliability criteria. Finally, for all positive and negative pages in our dataset, we ensured that some medical information was present on the page.

[2] Information on recent certification activity is available at the "Health on Net Foundation Recent Activity" page, *http://www.hon.ch/HONcode/Patients/LatestActivity/*.

Thus, our dataset (available at *http://timan.cs.uiuc.edu/downloads.html*) consists of a total of 360 webpages divided evenly into two classes – reliable and unreliable. The size of our dataset was mainly restricted by the amount of labor needed to judge the negative documents. Since reliability analysis requires reasonable amount of expertise in understanding the criteria and the content, we chose not to use Amazon Mechanical Turk (*https://www.mturk.com/mturk/*) for data quality concerns, even though the entire process of compiling the dataset took over two weeks. Nevertheless, we believe the dataset is sufficiently large for experimenting with binary classifiers and features for reliability prediction in the sense that even with 5-fold cross validation, we still have over 72 test cases in the held-out set, which would give us a meaningful average of performance.

6 Experiment Design

6.1 Evaluation Measures

Our evaluation criteria are based on two prominent application settings. In the first setup, which we call as the webpage classification task, we assume that the user is surfing the Web and the classifier is required to classify every new page that the user observes. In this setting, the utility of a classifier would depend on its classification accuracy. The classifier will make two types of errors – mislabel a reliable page as unreliable (type I error) and mislabel an unreliable page as reliable (type II error). Intuitively, the type II errors would cost more. In order to account for this bias, we measure the utility of our classifiers by a weighted accuracy function, parametrized by λ:

$$\textit{Weighted Accuracy}(\lambda) = \frac{(\lambda \times TP) + TN}{\lambda \times (TP + FN) + TN + FP}$$

where unreliable pages are labeled positive, reliable pages are labeled negative, and TP, TN, FP, and FN are the numbers of true positives, true negatives, false positives, and false negatives, respectively. The function assumes that cost of making a type II error is λ times the cost of making a type I error. We measure the utility of our classifiers with three different utility functions corresponding to $\lambda \in \{1, 2, 3\}$, for unbiased, moderately biased, and heavily biased setup, respectively.

In our second application setting, the system helps a human expert in labeling webpages as reliable or unreliable. We term this the webpage re-ranking task. The system generates an ordering of all webpages by ranking the reliable documents higher than all unreliable documents and the user can then look at this ordering and correct the mistakes. Ideally, the user would only need to choose a single cut-off threshold separating all reliable pages from the unreliable ones. The utility of a classifier depends on the number of mistakes that need to be corrected. This is similar to the problem of evaluating relevance ranking and, therefore, we use Mean Average Precision (MAP) as the evaluation measure for this setting.

6.2 Experiment Procedure

For our experiments, we used the SVMlight toolkit [10] to train an SVM classifier on different feature set combinations with varying amounts of training data, for all three bias settings. For evaluation, we used 5-fold cross validation. Each fold consisted of 288 training pages (144 reliable and 144 unreliable) and 72 test pages (36 reliable and 36 unreliable). In each case, the train and test examples belonged to different sets of websites. The overall weighted accuracies and MAP scores were calculated by averaging the five values. When measuring the weighted accuracy for $\lambda \in \{2, 3\}$, the SVM classifiers were trained to account for the bias. This was realized by setting the "`-j`" parameter in SVMlight to λ. The interpretation of the parameter is the same as our interpretation of λ.

7 Experiment Results

In this section, we first describe the results of our different lines of experiments and then present a thorough analysis of the observations. In particular, we are interested in identifying feature set combinations that lead to high performance while being robust towards amount of training data and different bias settings.

7.1 Effectiveness of feature sets

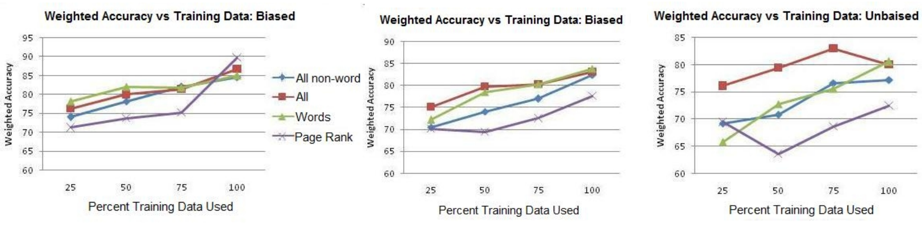
In our first set of experiments, we measured the performance of different feature set combinations based on overall accuracy and MAP scores. Table 1 shows the variation of weighted accuracy and MAP for the three bias settings over all feature sets, using SVM classifier.

Among the feature sets, word features tend to be the most discriminative, reinforcing our observation that authors of reliable and unreliable content tend to have different writing styles. PageRank features perform better than link-based features, especially when the bias is high. In such cases, we found that the internal PageRank feature, PR_{int}, becomes predominant. On the other hand, link-based classifiers use the presence of contact link CL and privacy policy link PL as dominant features. But their discriminative power is limited as many unreliable pages also contain these links and many reliable pages do not.

In general, addition of more features usually resulted in a measurable performance improvement. This is to be expected as the features belonging to different sets are largely independent and unlikely to have a high mutual information. A notable exception is the drop in performance when adding features to word based SVM classifiers. In order to better understand this behaviour, we show the MAP values of different SVM classifiers in Table 1. In spite of the 5% accuracy drop between *Word* and *All Non-PageRank* feature sets, the MAP value continues to remain high, suggesting that additional features are leading to a number of near misses possibly due to low performing link-based features. Similarly, while percentage accuracy of classifier based on all features is nearly same as the one trained on only word features, a higher MAP value indicates that the ordering generated by the former is more accurate, making it more robust than the latter.

Features	Wtd. Accu. (%)			MAP		
$\lambda \Rightarrow$	1	2	3	1	2	3
Links	60.8	71.1	79.6	0.708	0.766	0.763
PageRank	72.5	77.6	**89.7**	0.856	0.846	0.866
Words	**80.6**	**83.9**	85.0	0.899	0.905	0.902
Links+Commercial	67.8	75.9	79.6	0.794	0.814	0.815
Links+Commercial+PageRank	76.4	**83.9**	86.5	0.876	0.868	0.888
All non-Word	77.2	82.4	84.6	0.873	0.863	0.881
All non-PageRank	75.8	80.6	83.5	0.886	0.890	0.893
All	80.0	83.2	86.8	**0.916**	**0.929**	**0.921**

Table 1. Weighted accuracy (Wtd. Accu.) and Mean Average Precision for different feature set combinations with SVM classifier



(a) (b) (c)

Fig. 1. Variation of weighted accuracy with percent training data used in the (a) heavily biased ($\lambda = 3$), (b) moderately biased ($\lambda = 2$), and (c) unbiased ($\lambda = 1$) cases.

7.2 Influence of training set size

Our next line of experiments was to measure the influence of training set size on performance of different feature sets. We experimented with four high performing feature set combinations using 5-fold cross validation. For calculating performance on $x\%$ of training data, we trained each fold with only the first $x \in \{25\%, 50\%, 75\%, 100\%\}$ of training examples and tested them on the entire test set. The variation of weighted accuracy with x for $\lambda \in \{1, 2, 3\}$ are shown in Fig. 1.

We observe that the classifier based on *all* features is the most robust and clearly outperforms other combinations. On the other hand, word-based features tend to perform poorly when the amount of training data is low, but their performance improves the fastest as we add more training examples. In general, both accuracy and MAP show an increasing trend with training set size, suggesting that increasing the amount of training data is likely to further improve performance. A surprising observation, however, is the fluctuation in the accuracy of PageRank based classifiers. We discuss this issue in detail below.

Issue of PageRank: PageRank is often regarded as a crude measure of reliability. To gain a deeper insight into the performance fluctuations of PageRank

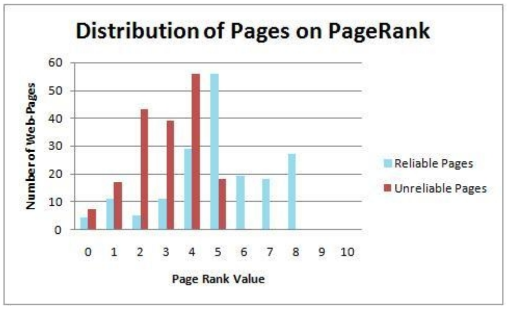


Fig. 2. Distribution of positive and negative webpages on PageRank values

features, we looked at the PageRank statistics of our dataset, shown in Fig. 2. The graph shows the distribution of all reliable and unreliable webpages present in the dataset based on their internal PageRank values (PR_{int}). Pages with high PageRank, in the band of $[6, 10]$, tend to be mostly reliable and, hence, easily separable. On the other hand, when the PageRank values are in $[0, 5]$, we find a mixture of reliable and unreliable pages that is hard to separate. Classifiers trained on PageRank features, tend to use $PR_{int} > \theta$ as their primary rule. Of the remaining five features, high values of $ExtPR_{min}$ (minimum $ExtPR$) and $ExtPR_{Q1}$ (first quartile of $ExtPR$) features are sometimes used for labeling pages as reliable when $PR_{int} < \theta$. The performance, therefore, mainly depends on learning an appropriate value of θ from the training examples. However, the narrow band between $(4, 6)$ contains a large number of both positive and negative examples. Thus, shifting θ by a single point on either side leads to high fluctuations in accuracy. For example, a simplistic classifier with only 1 rule: $PR_{int} > 4 \rightarrow Reliable$ would achieve an accuracy of 78.5% on our dataset. Raising or lowering the threshold by 1 results in a drop of 10% in accuracy. This is the reason for fluctuations in performance of PageRank classifiers. When we bias the classifier heavily, the learned classifier sets a high θ and completely disregards the remaining five PageRank features, resulting in a high reliability precision and, consequently, high weighted accuracy. We can therefore conclude that using PageRank alone as a measure of reliability is not sufficient.

7.3 Applications

In this section, we evaluate our classifier for two potential applications. The first is webpage re-ranking where we re-rank the results generated by a search engine based on reliability scores. The second is website accreditation, where we automatically process websites to generate a site reliability score.

Webpage Re-ranking: For this task, we re-ranked Google's results for 22 medical queries. The queries were chosen randomly from the list of "Similar Queries" displayed by Google. For each query, we manually judged the top 10 results as reliable and unreliable. We then classified each of the results using an

	Query: `cure back pain`	
Rank	Google	Ours
1	cure-back-pain.org	**familydoctor.org**
2	**familydoctor.org**	**emedicinehealth.com**
3	**emedicinehealth.com**	ehow.com
4	health2us.com	**webmd.com**
5	**webmd.com**	**spineuniverse.com**
6	**spineuniverse.com**	losethebackpain.com
7	ehow.com	backpaindetails.com
8	losethebackpain.com	losethebackpain.com
9	backpaindetails.com	health2us.com
10	losethebackpain.com	cure-back-pain.org
MAP	0.608	0.888

Table 2. Sample re-ranking results for an example query. Pages judged reliable are in bold face. Only domain names are shown for brevity

Website	Rel	Unrel
mayoclinic.com	98%	2%
rxlist.com	91%	9%
medicinenet.com	87%	13%
cancer.gov	65%	35%
goldbamboo.com	57%	43%
healthy-newage.com	51%	49%
guide4living.com	45%	55%
mnwelldir.org	43%	57%
shirleys-wellness-cafe.com	9%	91%
northstarnutritionals.com	0%	100%

Table 3. Websites ordered based on percentage of reliable pages found (out of 100 webpages each)

unbiased SVM classifier ($\lambda = 1$) trained on all features. A re-ranked list was then generated based on the reliability scores. We assumed that the relevance values of all top 10 results were similar and hence our re-ranking would only slightly hurt relevance. Google's reliability MAP over 22 queries was found to be 0.753. After re-ranking, the reliability MAP improved to 0.817. The re-ranked results were found to be better in 15, worse in 5, and same in case of 2 queries. Using Wilcoxon's signed-rank test, the improvement was significant at 0.05-level. Table 2 shows results from Google and our re-ranking for a sample query "`cure back pain`". A main difference between the two ranked results is that the webpage *http://www.cure-back-pain.org/* was ranked the highest by Google, but our system ranked it at the bottom. When we looked at the page, we observed that it was actually a biased site which talked about the owner's own experiences and promoted a book. To summarize, these results show that even with a small training set of 360 examples, the trained classifier can already improve the quality of search results over rankings that ignore reliability. Given that our performance improves with training data, by adding more training examples, the automatic prediction method is expected to be even more useful.

Website Accreditation: For the website accreditation task we selected a set of 10 websites and classified their webpages. None of these websites were included in our original training set. For each website, 100 webpages selected in a breadth-first manner were classified, and the percentage of reliable and unreliable pages was calculated. Classification results using a moderately biased SVM classifier ($\lambda = 2$) trained on all feature sets except PageRank are as shown in Table 3. The websites are ordered based on percentage of reliable pages. We observe that more authoritative and trustworthy sites, such as *www.mayoclinic.com* or *www.cancer.gov*, are ranked high. On the other hand, websites like *www.northstarnutritionals.com*, which is purely a commercial site selling online medications, and *www.shirleys-wellness-cafe.com*, which is an alternative medicine website not conforming to most of the HON criteria and containing content strongly critical of modern medicine, are ranked lowest. Websites like *www.guide4living.com* and *www.healthy-newage.com*, which are not particularly authoritative, conform to only some of HON criteria and provide mostly unbiased non-commercial information are ranked in the middle. Thus, our system generated a reasonable overall ranking of websites. We did not use the PageRank features for these experiments, as PageRank values need to be requested from an external Google Web Service that does not serve the requisite high volume of requests generated for obtaining external PageRank features, $ExtPR$. Additional website accreditation experiments with upto 5000 pages per website returned similar results.

8 Conclusion

In this paper, we presented a study of automatically predicting reliability of webpages in the medical domain. We cast the problem in a supervised learning setup and created a publicly available test set to quantitatively evaluate the task. Experimental results on this dataset are very encouraging. We were able to achieve an overall accuracy of 80%, showing that it is indeed feasible to predict the reliability of medical webpages through automatic feature extraction and classification. Results further show that using all the types of proposed features works better than only some of them, and performance can generally be improved over the Google PageRank baseline. Due to the importance of reliability in medical domain, we believe that our study can potentially have an impact on helping users to better assess reliability of information on the Web in this very important domain.

Acknowledgments: This material is based in part upon work supported by the National Science Foundation under Grant Numbers CNS-0834709 and CNS 1027965, by MURI award FA9550-08-1-0265, by a Sloan Research Fellowship and by the Multimodal Information Access and Synthesis Center at UIUC, part of CCICADA, a DHS Science and Technology Center of Excellence.

References

1. R. Andersen, C. Borgs, J. Chayes, J. Hopcroft, K. Jain, V. Mirrokni, and S. Teng. Robust PageRank and Locally Computable Spam Detection Features. In *AIRWeb '08: Proceedings of the 4th Intl. Workshop on Adversarial Information Retrieval on the Web*, pages 69–76, 2008.
2. Y. Aphinyanaphongs and C. F. Aliferis. Text Categorization Models for Identifying Unproven Cancer Treatments on the Web. In *MedInfo*, pages 968–972, 2007.
3. L. Becchetti, C. Castillo, D. Donato, R. Baeza-Yates, and S. Leonardi. Link analysis for Web spam detection. *ACM Trans. Web*, 2(1):1–42, 2008.
4. A. Borodin, G. O. Roberts, J. S. Rosenthal, and P. Tsaparas. Link Analysis Ranking: Algorithms, Theory, and Experiments. *ACM TOIT*, 5(1):231–297, 2005.
5. S. Brin and L. Page. The Anatomy of a Large-Scale Hypertextual Web Search Engine. In *Proc. of WWW*, 1998.

6. A. Gaudinat, N. Grabar, and C. Boyer. Automatic Retrieval of Web Pages with Standards of Ethics and Trustworthiness Within a Medical Portal: What a Page Name Tells Us. In *Proc. of Conf. on Artificial Intelligence in Medicine (AIME)*, pages 185–189, 2007.
7. A. Gaudinat, N. Grabar, and C. Boyer. Machine Learning Approach for Automatic Quality Criteria Detection of Health Web Pages. In *Proc. of the World Congress on Health (Medical) Informatics – Building Sustainable Health Systems*, volume 129, pages 705–709, 2007.
8. J. Han and M. Kamber. *Data Mining: Concepts and Techniques.* Morgan Kaufmann Publ., 2006.
9. M. R. Henzinger. Link Analysis in Web Information Retrieval. In *Bulletin of the IEEE Computer Society Technical Committee on Data Engineering*, volume 23, pages 3–8, 2000.
10. T. Joachims. Making large-scale SVM Learning Practical. In B. Schlkopf, C. Burges, and A. Smola, editors, *Advances in Kernel Methods – Support Vector Learning.* MIT Press, 1998.
11. T. Joachims. Text Categorization with Support Vector Machines: Learning with Many Relevant Features. In *Proc. of ECML*, pages 137–142, 1998.
12. D. R. Lankes. *Trusting the Internet: New Approaches to Credibility Tools*, pages 101–122. MIT Press, 2008.
13. J. V. Marriott, P. Stec, T. El-Toukhy, Y. Khalaf, P. Braude, and A. Coomarasamy. Infertility information on the World Wide Web: a cross-sectional survey of quality of infertility information on the internet in the UK. In *Human Reproduction*, pages 1520–1525, Jul 2008.
14. M. J. Martin. *Reliability and verification of natural language text on the world wide web.* PhD thesis, Las Cruces, NM, USA, 2005. Chair-Hartley, Roger T.
15. S. C. Matthews, A. Camacho, P. J. Mills, and J. E. Dimsdale. The Internet for Medical Information About Cancer: Help or Hindrance? In *Psychosomatics*, volume 44, pages 100–103, Apr 2003.
16. S. L. Price and W. R. Hersh. Filtering Web pages for Quality Indicators: An Empirical Approach to Finding High Quality Consumer Health Information on the World Wide Web. In *Proceedings of AMIA Symposium*, pages 911–915, 1999.
17. V. L. Rubin and E. D. Liddy. Assessing credibility of weblogs. In *AAAI Symposium on Computational Approaches to Analysing Weblogs (AAAI-CAAW)*, pages 187–190, 2006.
18. T. T. Tang, N. Craswell, D. Hawking, K. Griffiths, and H. Christensen. Quality and relevance of domain-specific search: A case study in mental health. *Inf. Retr.*, 9(2):207–225, 2006.
19. V. N. Vapnik. The Nature of Statistical Learning Theory. In *Springer*, 1995.
20. V. Vydiswaran, C. Zhai, and D. Roth. Content-driven Trust Propagation Framework. In *Proceedings of the 17^{th} SIGKDD Conference on Knowledge Discovery and Data Mining (KDD)*, pages 974–982, 2011.
21. Y. Wang and R. Richard. Rule-based Automatic Criteria Detection for Assessing Quality of Online Health Information. *Journal on Information Technology in Healthcare*, 5(5):288–299, 2007.
22. L. Zhang, Y. Zhang, Y. Zhang, and X. Li. Exploring both Content and Link Quality for Anti-Spamming. In *Proceedings of the Sixth IEEE International Conference on Computer and Information Technology (CIT)*, page 37, 2006.