A system to automatically assess credibility of web sites is described. The system uses information about credentials, advertisements, and web design to produce a credibility measure. The system was tested on several web sites in the medical domain. It selected the same three domains as most credible as did two manual rankings from the Wall Street Journal and a Stanford University research study.

Introduction

Tools like search engines and web portals do an exceptional job of directing users to websites with information relevant to what they are looking for. Unfortunately, users of these services receive no indication of how reliable the information contained within those websites actually is. For example, a search on Google for "Multiple Sclerosis Association" in July 2004 gave the Multiple Sclerosis Association of King County website as the third result. This website sounds and looks official. However, it contains a section that makes a convincing argument for MS patients to try bee sting therapy. This therapy has been studied and there is no scientific evidence to suggest that it is effective, and some to show that it may be detrimental (National Institute of Nursing Research, 2003).

Clearly, if unwary Internet users can be lured into unsafe or untested medical practices by credible-looking websites, then there is a significant problem to be solved. To combat the problem of determining the credibility of online sources we propose an automated system. Recent work in retrieval has moved beyond simply estimating relevance to include the examination of link structures to identify authoritative and important documents (Google, 2005; Haveliwala 2002). The problem with these algorithms is the underlying assumption that high relevance implies credibility. This is not always a bad assumption. For example, the American Diabetes Association is a perfectly reputable source of diabetes information and shows up as the most relevant on Google, Yahoo, and Altavista. However, in a number of cases this assumption can be flawed and may lead to harm.

The first step to creating an automated system for determining credibility is to identify the specific elements of a web page that affect its actual and perceived credibility. Unfortunately, the only absolute measure of credibility, accuracy of information, is unsuited to an automated system at this time. Instead, we can use measures that are indicative of a site's credibility. These indicators include credentials, domain assertions, amount and specificity of advertising, and psychological heuristics. A set of surveys done by the Stanford Persuasive Technology Lab have identified the top ten heuristics searchers use to make quick decisions on credibility. These are the design of the page, the information structure, information focus, company motive, usefulness, accuracy of information, name recognition and reputation, advertising, bias, tone, identity of site sponsor, functionality, and customer service (Fogg et al., 2002; Fogg et al., 2003).

Credibility Assessment in the Medical Domain

We use the medical domain in this preliminary research because it has immediate importance to users and an existing structure for credentials. The current implementation uses credentials, advertising, and design (the first of the Stanford heuristics) to carry out three separate analysis phases to compute a credibility weighting on a scale from 0 to 10. The weightings used are 50% for credentials, 30% for advertising, and 20% for the design.

First is the credentials phase. Each page is searched for a reference to any of the AMA accredited schools, the National Institutes of Health, the American Medical Association, the Centers for Disease Control and Prevention, the Food and Drug Administration, or any of their acronyms. The number of credential references to pages analyzed is then computed.

Next is an advertising analysis. This phase presents challenges. The checker searches the pages for references to doubleclick, links to adv.thecurrentdomain, or ad.thecurrentdomain. For example, WebMD uses adv.webmd.com as an ad server. These are common methods of placing advertisements in the HTML. Unfortunately, this system is not perfect and may still miss or double-count advertisements.
Finally, the most prominent of the psychological heuristics, design, is analyzed using the W3C's online page validator and capturing the number, if any, of violations. The information structure could be analyzed, but further research would have to be done in figuring out what a “good” or “bad” graph would look like.

Comparison with Human Judgment

The system was tested on nine websites that had been ranked by experts in a Wall Street Journal article (Forster, 2002). In the article these pages were ranked from most to least credible in this order: National Institutes of Health, MayoClinic.com, WebMD, InteliHealth, MDChoice, DrKoop, HealthWorld, Oxygen Health and Fitness, and Health Bulletin. In test, the system ranked them on a 10 point scale as shown in Table 1:

<table>
<thead>
<tr>
<th>Medical Web Site</th>
<th>Ranking</th>
</tr>
</thead>
<tbody>
<tr>
<td>National Institutes of Health</td>
<td>10.00</td>
</tr>
<tr>
<td>MayoClinic.com</td>
<td>7.97</td>
</tr>
<tr>
<td>InteliHealth</td>
<td>7.40</td>
</tr>
<tr>
<td>HealthBulletin</td>
<td>5.15</td>
</tr>
<tr>
<td>HealthWorld</td>
<td>4.68</td>
</tr>
<tr>
<td>MDChoice</td>
<td>4.59</td>
</tr>
</tbody>
</table>

These results are very similar except for two major deviations, the placements of WebMD and DrKoop. Actual examination of the DrKoop site reveals roughly 15 advertisements per page and very few credible citations on any of the articles. Its low rating may be fully justified, and the discrepancy may be due to the fact that the Journal study was done in 2002. On the other hand, WebMD gets a falsely low rating for two reasons. First, many of its articles are on multiple pages, with the credentials and sources only on the last page. Also, each WebMD writer has one single page on which they state their credentials. The effect of this is a much lower credential per page count.

An additional comparison can be made to the rankings determined by the Stanford study. The users in this study ranked several medical web sites for credibility. The highest ranked cluster contained MayoClinic.com, InteliHealth, and National Institutes of Health. The automatic ranking examined here identified the same three sites as most credible. The bottom cluster in the Stanford study contained HealthWorld Online, Dr. Weil, Health Bulletin, and Oxygen.com; the middle cluster contained MDChoice, Dr. Koop, and WebMD.

Conclusions

In conclusion, an automated system for determining online credibility is needed for users for the cases where the relevancy = credibility assumption breaks down. While credibility cannot be precisely determined without exhaustive fact checking, 'a good enough' system can be produced by use of a few heuristics as has been shown. The experimental system developed produced credibility assessments comparable to those produced by human searchers and analysts. Further work in this area would involve further consideration of the metrics and weightings being used in the analysis, generalizing the heuristics used in this project to all domains, further research in natural language processing for the addition of assertions into the algorithm, and research into the impact the semantic web will have on this type of analysis.

ACKNOWLEDGMENTS

This research was supported in part by the National Science Foundation under grant CNS-0353637.

REFERENCES


