Towards higher quality health search results: Automated quality rating of depression websites.

David Hawking
CSIRO ICT Centre
GPO Box 664
Canberra, Australia 2601
david.hawking@acm.org

Thanh Tang
Computer Science Department
Australian National University
Canberra, Australia 0200
ttintang@gmail.com

Ramesh Sankaranarayana
Computer Science Department
Australian National University
Canberra, Australia 0200
ramesh@cs.anu.edu.au

Kathleen M. Griffiths
Centre for Mental Health Research
Australian National University
Canberra, Australia 0200
kathy.griffiths@anu.edu.au

Nick Craswell
Microsoft Research
JJ Thomson Ave
Cambridge UK
nickcr@microsoft.com

Peter Bailey
Microsoft Research
GPO Box 664
Canberra, Australia 2601
peter.bailey@csiro.au

ABSTRACT

Consumers are increasingly reliant on the web for health information and advice, and increasingly reliant on search engines to locate health resources. A search engine able to bias its results against sites offering dubious or even harmful health advice, would obviously be of value to consumers.

We have developed an Automated Quality Assessment procedure, which learns complex information retrieval queries from training sets of high and low quality depression websites. Processing these queries on test sets yielded site scores which correlated 0.85 with human expert ratings based on evidence-based guidelines. We have subsequently used the AQA technique to guide a depression-focused web crawler and to filter results from a major web search engine with very encouraging results.

Currently, we are investigating whether the AQA technique will generalise to other health domains, starting with obesity.

1. INTRODUCTION

This paper provides an overview of progress achieved to date in a four-way informal collaboration addressing the quality of health information on websites and in search engine results. The collaboration involves researchers from the CSIRO ICT Centre[1], the Centre for Mental Health Research (CMHR)[2], the Australian National University Computer Science Department[3] and Microsoft Research[4]. Much of the work was carried out by a doctoral student (Tang) who graduated in 2006.

As participants in Medinfo are well aware, a large proportion of Internet users obtain health information via the Web and use search engines to locate it [2]. Web technology can deliver information very cost effectively. Christensen et al.[1] have shown that interaction with the BluePages depression information site[5] can reduce depressive symptoms.

Not all health web sites provide unbiased, high-quality, comprehensive information. Griffiths et al.[3] pioneered the use of evidence-based guidelines in rating the quality of health sites on the Web. In that study and in [4] they found that many sites do not accord well with scientific evidence.

Two types of web search engine are of interest to health search:

- Global search engines typified by Google, Yahoo! and Microsoft Live.
- Specialised health search engines such as the search facilities provided by `healthfinder.gov` and BluePages. The Healthfinder search facility (HFS) indexes health websites selected with a view to ensuring reliable information on a broad range of health topics. The BluePages search facility (BS) indexes more than 200 sites, restricted to those providing information about depression.

In 2004 [8], we assessed the results returned in response to 100 depression related queries by representatives of both types of search facilities, including Google, HFS[6], and BPS. By adding the word depression to Google queries which didn't already contain it (GoogleD condition), we were able to constrain Google search results to the depression domain. We found that GoogleD returned more relevant documents than either BPS or HFS but achieved a significantly lower quality score than BPS. In particular, GoogleD results tended to include recommendations for depression treatments of unproven or dubious effectiveness.

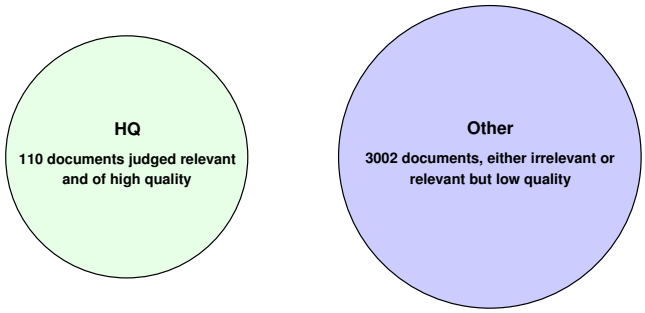
In [5] we used expert ratings of 29 sites evaluated in earlier studies as training data for an automated quality rating procedure (AQA). We showed that AQA-derived sites scores correlated very highly with evidence based ratings by human experts.

[1] `www.ict.csiro.au/`
[2] `www.anu.edu.au/cmhr/`
[3] `www.anu.edu.au/cmhr/`
[4] `research.microsoft.com/`
[5] `bluepages.anu.edu.au`
[6] The HFS search facility now (06 August 2007) uses Google appliance technology, but at the time of the study it did not.

This paper explains how AQA works, reports its application in building health-specific search engines which are biased toward high-quality information sites, and discusses generalisation of AQA to health topics other than depression.

2. AQA PROCEDURE

Of necessity, this is a simplified explanation of AQA. Readers are referred to [5] for a fuller description.



depression	10.3	ECT	2.4
treatment	5.7	antidepressants	1.9
disorder	3.3	zoloft	1.5
patient	3.3	mental health	1.2
medication	3.0	cognitive therapy	0.84

Figure 1: By contrasting occurrence probabilities of words and phrases between the HQ and Other sets, we can make a list of words and phrases which most effectively discriminate between high quality depression documents and other documents.

The AQA procedure uses a variant of the *relevance feedback* technique from the field of information retrieval. This procedure is illustrated in Figure 1. For each term (word or phrase) occurring within the HQ set, we estimated the occurrence probability of that term in high quality relevant documents, by dividing the number of HQ documents containing the term by the total number of HQ documents. We estimated a corresponding occurrence probability for documents which are not high quality relevant, in similar fashion. Robertson term selection values (TSVs)[7], based on the differences between these occurrence probabilities, were calculated for all the terms.

We ordered words and phrases by their TSV weights, choosing 20 words and 29 phrases with weights, for inclusion in a *quality query*. When a document is scored against this query using an information retrieval system its score can be used as a measure of quality.

We used an exactly analogous process to "learn" a *relevance query* to distinguish between documents which are relevant to the topic of depression and those which are not.

To score a site using AQA, we first fetched all its pages using a web crawler. We then scored each page using the quality and relevance queries and aggregated the page scores into site scores using a formula which linearly combined average page score with number of pages from the site. We then linearly combined the site relevance and site quality scores into a single site score and scaled the result into the range 0–20 to match the rating scale used in expert human assessments.

3. EFFECTIVENESS OF AQA

In [5] we showed that AQA scores obtained for a set of 29 test sites (not overlapping with sites used for training) correlated 0.85 with human expert ratings. This was highly significant. ($p < 0.001$) By contrast, only a moderate correlation was found between the expert ratings and the PageRank scores reported by the Google toolbar. (PageRank is Google's proprietary, topic-independent measurement of a page's importance [6]. Some people think that it may measure or predict the quality of web information.)

4. APPLICATIONS OF AQA

AQA can be used to improve the quality of depression information returned to users by search engines — either global search engines or specialised depression search facilities.

In [9] we showed that AQA scores could be used to guide a *quality focused crawler* (QFC) which selectively crawled the relevant high quality parts of the web. Starting the QFC from the Open Directory[7] depression category can provide an effective, low-maintenance method for constructing a specialised depression search facility.

In [10] we report the use of AQA scores in post-processing global search engine results in order to improve the quality of the results presented to searchers. We compare the relevance and quality scores of BPS, QFC, GoogleD, and three post-processed variants. Relevance scores are based on a four point relevance scale and reflect the ability of a search engine to return relevant content regardless of quality. Quality scores were computed for the 50 search queries which represented conventional and alternative treatments for depression. These treatments had been rated using a systematic review of scientific evidence as Very Effective, Effective, OK, Unsure or Not Effective. Non-expert judges were asked to assess whether each result page recommended or advised against the treatment. A search engine returning results recommending effective treatments and advising against ineffective treatments receives a high quality score.

We show that best overall quality score performance is achieved by a condition in which normal Google result ranks are averaged with rank in AQA order. We also show that the quality and relevance results for QFC are significantly better than those for BPS, which seems to have deteriorated over the two year period since it was last measured, due to lack of maintenance of the list of included sites. Engineering limitations in the QFC caused the crawl to be truncated early but analysis of its behaviour suggest that if those limitations were able to be overcome, QFC performance might approach that of post-processed global search results.

We believe that AQA could potentially be used to provide site quality ratings for web users. Rather than providing each interested web user with their own copy of AQA rating software, it would make sense to offer a Web directory service for AQA rated sites. Consumers or other interested parties wishing to ascertain the quality of information on a website could consult the AQA online directory. If the site had already been rated, the rating would be available immediately, otherwise the site would be scheduled for rating and the rating emailed after some hours. Over time the site would become more and more comprehensive. Ratings would need to be confirmed at intervals to confirm the con-

[7] dmoz.org

tinued existence of the site and to accommodate changes in its content.

One potential problem in using AQA in "production" is that some sites providing depression information will feel a strong motivation to optimise their AQA scores so as to maintain or increase the prominence of their site in search results and the acceptability of their site to consumers. If they did this by making their site more comprehensive and more in accord with scientific evidence, this would be a beneficial outcome. However, less scrupulous publishers may add words and phrases favoured by AQA, while continuing to deliver inaccurate information and misleading advice. In the such cases, normal search engine spam suppression techniques, such as black-listing and detection of abnormal patterns, would need to be invoked.

5. GENERALISATION OF AQA TO OTHER HEALTH TOPICS

Thus far, all our AQA work has been in the area of depression. We want to confirm that the method will generalise to other topics and have already commenced a project in the area of obesity, with some financial assistance from Microsoft Research Asia.

The main cost in applying the AQA to other topics, such as obesity, is the expert time required to judge a large number of training sites. The Centre for Mental Health Research has recently completed assessments of over 60 obesity sites, chosen to constitute a random sample of sites listed in the weight loss and obesity categories of the Open Directory. Judging took of the order of four months equivalent full-time effort. We plan to use half of the sites as training data and half to test the accuracy of AQA ratings in this new domain.

6. REFERENCES

[1] H. Christensen, K. Griffiths, and A. Jorm. Delivering interventions for depression by using the internet: randomised controlled trial. *British Medical Journal*, 328(7434):265–0, 2004.

[2] S. Fox. Health information online. PEW Internet & American Life Project, May 2005. `/www.pewinternet.org/PPF/r/156/report_display.asp`.

[3] K. Griffiths and H. Christensen. Quality of web based information on treatment of depression: cross sectional survey. *British Medical Journal*, 321(7275):1511–1515, 2000.

[4] K. Griffiths and H. Christensen. The quality and accessibility of australian depression sites on the world wide web. *Medical Journal of Australia*, 176:S97–S104, 2002.

[5] K. Griffiths, T. Tang, D. Hawking, and H. Christensen. Automated assessment of the quality of depression websites. *Journal of Medical Internet Research*, 7(5), 2005. `http://es.csiro.au/pubs/griffiths_jmir.pdf` and `http://www.jmir.org/2005/5/e59/`.

[6] L. Page, S. Brin, R. Motwani, and T. Winograd. The pagerank citation ranking: Bringing order to the web. Technical report, Stanford, Santa Barbara, CA 93106, January 1998. `dbpubs.stanford.edu:8090/pub/1999-66`.

[7] S. E. Robertson. On term selection for query expansion. *Journal of Documentation*, 46(4):359–364, 1990.

[8] T. Tang, N. Craswell, D. Hawking, K. Griffiths, and H. Christensen. Quality and relevance of domain-specific search: A case study in mental health. *Information Retrieval*, 9(2):207–225, 2006. `http://es.csiro.au/pubs/tang_domainspec.pdf`.

[9] T. Tang, D. Hawking, N. Craswell, and K. Griffiths. Focused crawling for both relevance and quality of medical information. In *Proceedings of CIKM 2005*, pages 147–154, 2005. `http://es.csiro.au/pubs/tang_cikm05.pdf`.

[10] T. Tang, D. Hawking, R. Sankaranarayana, K. Griffiths, and N. Craswell. Biasing health domain search results in favour of evidence-based sites. In submission.