Quality-oriented search for depression portals.

Thanh Tang[1], David Hawking[2], Ramesh Sankaranarayana[1], Kathleen M. Griffiths[3], and Nick Craswell[4]

[1] Department of Computer Science, Australian National University, Canberra, Australia, ttintang@gmail.com, ramesh@cs.anu.edu.au
[2] Funnelback Pty Ltd, Canberra, Australia, david.hawking@acm.org
[3] Centre For Mental Health Research, Australian National University, Canberra, Australia, kathy.griffiths@anu.edu.au
[4] Microsoft Research, Cambridge UK, nickcr@microsoft.com

Abstract. The problem of low-quality information on the Web is nowhere more important than in the domain of health, where unsound information and misleading advice can have serious consequences. The quality of health web sites can be rated by subject experts against evidence-based guidelines. We previously developed an automated quality rating technique (AQA) for depression websites and showed that it correlated 0.85 with such expert ratings.
In this paper, we use AQA to filter or rerank Google results returned in response to queries relating to depression. We compare this to an unrestricted quality-oriented (AQA based) focused crawl starting from an Open Directory category and a conventional crawl with manually constructed seedlist and inclusion rules. The results show that post-processed Google outperforms other forms of search engine restricted to the domain of depressive illness on both relevance and quality.

Key words: Health search on the Web, Health portals

1 Introduction

Large numbers of people seek and access health information via the Web [6]. Web-delivery of information and interventions relating to depression and other mental health issues helps overcome reluctance to seek help induced by stigma.

Unfortunately, information on many depression web sites does not accord well with scientific evidence [7]. For commercial, religious or other motives, sites may promote unproven or even harmful treatments while failing to mention alternatives which have been proven to be effective.

It is desirable that search engines covering depression (or other health) content should bias results in favour of sites providing quality information. In the present study, we examine whether a previously published automated method for rating depression website quality (AQA, [8]) can be used to improve the quality of depression search results.

Our first aim was to determine whether using AQA to filter or rerank results from a global search engine, Google, would improve the quality of results for queries relating to depression. Our second aim addressed much smaller scale search facilities typically associated with health information portals. Can an unrestricted focused crawler starting with an Open Directory depression category seed list and using AQA quality ratings outperform a manually defined portal search facility?

1.1 Contributions

To the best of our knowledge, the present paper is the first to:

1. Evaluate the three principal techniques for constructing subject portal search services: manual seedlist and rules, focused crawling, and filtering/reranking of a general search engine. Our evaluation is extensive (100 queries) and uses independent human assessors to rate both the relevance and the quality of information returned.
2. Show that the quality of health search results returned by a highly-performing general search engine (Google) can be improved by post-filtering or reranking, using automatically derived quality scores.

2 Previous Work

A number of previous studies (e.g. Ilic et al. [9] and Bin and Lun [2]) have compared the effectiveness of general search engines with that of medical-specific engines in locating medical information. Our own previous study [17] included relevance and quality evaluations of health search engines but none of those engines were based on focused crawling or on quality filtering.

In the domain of health, the quality of information (e.g. whether a treatment for a condition is effective or not) is of vital concern. Various authors have studied the quality of online health information, e.g. [1, 5, 7]

Evidence-based medicine defines rigorous procedures [4] for systematically reviewing scientific studies, assessing the level of evidence they provide and, from them, synthesizing guidelines for clinical practice. The Cochrane Collaboration[5]) "maintains a collection of evidence-based medicine databases including a database of systematic reviews".

In the area of depressive illness, Jorm et al [11, 12] have applied the evidence-based approach in rating a large number of conventional and alternative treatments while the Oxford Centre for Evidence Based Mental Health (CEBMH) have produced evidence-based clinical guidelines [3].

One might expect that the static scores used by Web search engines would predict evidence-based site ratings, but we found [8] only a moderate correlation ($r = .61, p = 0.002$) between expert ratings and PageRanks reported by the Google toolbar (and then only when PageRanks reported to be zero were

[5] `www.cochrane.org`

excluded). By contrast, we found a high correlation ($r = .85, p < 0.001$) between our AQA method (see below) and the expert ratings.

3 The present study

3.1 Overview of the AQA method

Space does not permit a full description of the AQA rating method. The reader is referred to [8] for full details. In essence, a weighted *relevance query* is learned by contrasting term probabilities in a set of documents relevant to depression and a set with low probability of relevance. The query consists of the 20 words and 20 phrases with highest Robertson term selection values (TSVs) [13]. A *quality query* is learned using a similar method contrasting high quality documents with a set with low probability of being high quality.

The queries are run against a collection including sites to be rated and BM25 scores are calculated and normalised relative to the highest possible BM25 score for that collection. The relevance score for a site (S_r) is computed from the mean score of pages in that site ($\bar{r}$) and the number of non-zero pages.

$$S_r = \alpha\bar{r} + (1 - \alpha)|R| \tag{1}$$

A site quality score (S_q) is calculated in similar fashion and the two are linearly combined. Combining coefficients are learned by simple optimisation over a training set of 29 sites.

3.2 Engines

Six "engines" were involved in this experiment, as shown in Table 1[6].

BluePages Search (BPS): BPS is fully described in [17]. It is a crawled index hosted on `bluepages.anu.edu.au`. The BPS crawler's behaviour is controlled by a seed list of more than 200 sites plus corresponding URL-based include/exclude patterns. The seed list and patterns were manually constructed in 2004. Unfortunately, they have not been updated since, due to heavy time cost and limited resources.

GoogleD: GoogleD converts Google into a depression-specific search facility by adding the word *depression* to queries which don't already include it. Previous work [17] showed that GoogleD achieved good coverage and good precision.

Coverage of GoogleD is very large. At the time of writing, Google estimated 15.8 million results[7] for the query `depression "mental health"`. GoogleD results were obtained using the Google API.

[6] BPS and QFC used the *BM25* [14] formula as implemented by the Panoptic search engine developed in a CSIRO/Australian National University research project and subsequently commercialised. GoogleD, FGD1 and FGD2 results rely on Google's proprietary algorithm.

[7] Reported result counts are known to be estimates only.

Table 1. The search engines included in the experiment. Note that Google no longer reports its index size.

Engine	URL	Pages in index	Notes
BPS	`bluepages.anu.edu.au/search.html`	2.07×10^4	Depression specific
GoogleD	`google.com`	$\approx 10^{10}$	Google with "depression" added to queries
FGD1	not publicly available	$\approx 10^{10}$	Filtered out 75% of the bottom-quality sites from GoogleD results
FGD2	not publicly available	$\approx 10^{10}$	Filtered out 95% of the bottom-quality sites from GoogleD results
QFC	not publicly available	4.18×10^4	Quality focused crawler
AvgRankGD	not publicly available	$\approx 10^{10}$	Averaging Google Rank and quality rank

Quality Focused Crawl (QFC): The QFC engine was built around the quality focused crawler described in [18], but working with an enlarged seed set comprising 301 depression category sites taken from the Open Directory [8].

Briefly, the focused crawler maintains a list of unvisited candidate pages (the frontier) ordered by the product of estimated relevance and quality scores. The relevance score is a confidence level associated with a decision tree node. The relevance decision tree is based on words in the anchor text, words in the target URL and words in the 50 characters before and after the link (link context). The quality score of a linking page is computed using the AQA quality query and it is propagated equally along all of its outlinks.

The quality score of pages in the frontier is the average of incoming quality scores. Consequently, the ordering of the frontier changes as further incoming links are processed.

The QFC crawler was not engineered to the necessary degree, nor were machine and network resources available to support a crawl to even approach Google coverage. Accordingly, the QFC crawl was stopped after 41,823 pages, about twice the size of BPS. At this stage, the crawler had already become trapped in a site with dynamically generated URLs.

Quality filtering of Google results(FGD1 and FGD2): FGD1, FGD2 and AvgRankGD (see below) rely on obtaining a deep results list from GoogleD and deriving a corresponding *sorted site list* as follows:

– The Google API was used to collect the top 1,000 ranked results for each of the 100 test queries. 97,445 distinct results were collected and organised according to the Google rank of each URL for each query, forming a list called the *Google ranked list*.

[8] `dmoz.org`

- From the Google ranked lists aggregated across queries, a *site list* was derived using host name prefixes and other heuristics.
- The AQA procedure described in [8] was applied to the *site list*, and the sites were then sorted, resulting in a list of websites with quality scores arranging in descending order of quality, called the *sorted site list*.

The *Google ranked list* was then filtered, selecting for each query the top ten URLs belonging to sites whose quality scores reach a threshold, while preserving their original Google rank order.

In the absence of any prior experience or published literature, we arbitrarily set two thresholds designed to include approximately the top 25% of sites (FGD1) and approximately the top 5% of sites (FGD2). This resulted in 360 and 58 sites respectively.

Quality re-ranking of Google results(AvgRankGD): Another engine, AvgRankGD, was created by combining the GoogleD rank and the rank of its site based on the quality score of that site. Each URL in the *Google ranked list* was re-ordered according to the mean of these two ranks. The ten top ranked URLs for each query were selected as the AvgRankGD result set.

3.3 Query set

Queries were submitted to each primary engine and result sets obtained directly or after post-processing were pooled for each query and subjected to blind judging. Note that effectively, the test corpus is the entire public Web, and the performance of a retrieval system depends upon its coverage as well as its ranking function.

The query set comprises an equal mix of names of depression treatments (e.g. 'alcohol avoidance', 'cognitive behaviour therapy' and 'cipramil') and depression-related queries submitted to search engines. Depression treatment names are taken from a systematic review by Jorm et al. [11].

To avoid bias, the set of submitted queries comprised an equal mix of the most popular queries submitted to BPS (e.g. 'alcohol and fluoxetine metabolism' and 'anxiety') and the most popular depression-related queries for a general search engine, as listed by Overture (e.g. 'adolescent depression' and 'commit suicide'). Depression-related queries were identified using MeSH (Medical Subject Headings)[9].

3.4 Judgments

Relevance was measured for the full set of 100 queries. Judges were asked to assign one of four levels of relevance using the Sormunen scheme [16]).

Quality was assessed only for 50 treatment queries. Restricting quality judgments to treatment queries allowed the use of non-expert judges. They were asked to judge:

[9] `www.nlm.nih.gov/mesh/`

– Does this page recommend or support the treatment (*positive*), oppose the treatment (*negative*), or neither *neutral*)?

The treatments were classified on the basis of a systematic review of the effectiveness of depression treatments [11] into the categories listed in Table 3.

A quality score for engines was derived by asking two subject experts to assign appropriate rewards and penalties to each combination of treatment value and recommendation as shown in Table 3. The experts assigned the rewards and penalties without knowledge of experimental results.

4 Results

4.1 Relevance results

Significance tests were conducted using Wilcoxon Signed Rank tests [15], with confidence level of 95%.

Table 2 presents relevance performance of the six engines for two measures: mean *modified average precision* (MAP) and mean *normalised discounted cumulative gain* (NDCG) [10]. GoogleD returned best results on both relevance measures, followed by FGD2, FGD1, AvgRankGD, QFC, and BPS. The results were almost consistent for both measures, except for an order swap between FGD1 and FGD2.

There was no significant difference in mean MAP scores between the best performer, GoogleD, and any of the Google variants. However, GoogleD significantly outperformed QFC ($p < 0.001$) and QFC outperformed BPS ($p < 0.0001$).

4.2 Quality results

Table 4 lists the overall quality scores for each engine and the basis on which the scores were calculated. AvgRankGD achieved the highest overall quality score, followed by FGD2, QFC, FGD1, GoogleD, and BPS.

GoogleD returned a lot of correct advice but also the highest number of pages giving incorrect advice. BPS and GoogleD were the worst performers for this measure, achieving similar ratios of correct advice to all advice (74% and 76% respectively).

The numbers listed in Table 4 are dependent on the relevance of search results. For example, GoogleD retrieves more relevant pages than BPS, so it naturally retrieves more pages with advice. Note also that the domain-specific engines that use AQA (noted as "Y" in the *AQA* column) all achieved better quality ratios than those without AQA.

4.3 Further analysis of QFC performance

QFC retrieved significantly more relevant results than BPS, and also performed much better in terms of overall quality score as well as on overall proportion of good advice.

Table 2. Relevance results for the search engines. *MAP* refers to Modified Average Precision and *NDCG* means Normalised Discounted Cumulative Gain.

Engine	mean MAP	mean NDCG
GoogleD	0.554	0.709
BPS	0.256	0.469
QFC	0.400	0.566
FGD1	0.543	0.707
FGD2	0.548	0.683
AvgRankGD	0.513	0.657

Table 3. Quality Rating. *Positive* means that the treatment is recommended by the page being judged. *Negative* means that the treatment is recommended against.

Treatment rating	Positive	Negative
Very strong evidence (***)	4	-5
Strong evidence (**)	3	-4
Some evidence (*)	1	-2
No evidence (-)	-1	0
Not Effective (X)	-5	4

Table 4. Number of documents recommended for different treatment types. Effectiveness ratings (e.g. '***' and 'X') are defined in Table 3. Quality scores are the sum of the expert-assigned rewards and penalties listed in that table.

	Recommend					Recommendagainst					
	***	**	*	-	X	***	**	*	-	X	**Quality score**
GoogleD	51	38	85	88	25	5	1	2	7	12	**205**
BPS	21	29	46	39	14	11	0	1	8	20	**131**
QFC	50	42	69	76	22	3	0	1	8	10	**233**
FGD1	46	39	74	84	19	3	0	0	7	11	**225**
FGD2	47	38	75	68	17	4	0	1	7	14	**258**
AvgRankGD	47	41	78	75	16	3	0	0	8	13	**271**

However, its quality result was not much better than that of GoogleD, and its relevance performance was worse. We hypothesize that QFC would have performed better if the crawl had been larger. The following analyses were carried out to investigate this.

Coverage analysis: The coverage analysis was conducted for the QFC, to find out how many of the relevant results in GoogleD did not exist in the QFC. Of the 837 relevant pages returned by GoogleD, only 279 were in the QFC index (33%). Conversely, an attempt to locate how many relevant URLs returned by the QFC are in the Google index showed that out of 785 relevant pages, at least 590 were in the Google index (75%).

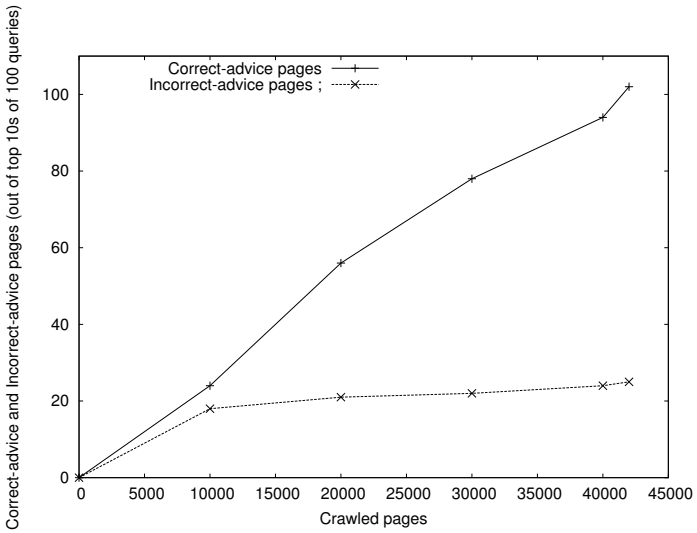

Fig. 1. The number of top ten correct-advice and incorrect-advice pages which were crawled according to the QFC's crawling progress.

Quality analysis: The QFC returned 102 webpages giving correct advice and 25 webpages recommending incorrect advice. Because the focused crawler acquires more information as it crawls, it may predict relevance and quality better during the later stages of the crawl. We hypothesized that, up to a point, correct-advice pages might be crawled at a steady rate while incorrect-advice pages would be more effectively rejected as the crawl progressed.

Figure 1 shows the rates of collecting good-advice and bad-advice pages. As may be seen, the number of good-advice pages rises sharply while the rate of fetching bad advice pages flattens off. Note that we only have quality judgments for pages retrieved in the first page of results for our test queries. Other good or bad advice pages may have been crawled without our knowledge.

5 Discussion

Our results suggest that if a global search engine operator wished to operate a vertical search portal within the domain of depression information, it could improve the quality of results returned by combining its normal ranking scores with quality scores derived from AQA. It could almost certainly do that more effectively than we could, because it has access to scores rather than ranks. To maintain query response at reasonable levels, AQA ratings would have to be computed at crawl/index time rather than on the fly.

If a health organization wishing to operate a depression portal search was unable or unwilling to engage in a contractual arrangement with a global search provider, our results suggest that high quality, highly relevant results would be obtainable from a more robustly engineered quality-focused crawler starting from the Open Directory. Although actual results for QFC were poor relative to post-processed Google, our follow-up analyses suggest that the gap could be substantially narrowed with a more extensive crawl.

5.1 Susceptibility to optimisation and spam

Because the AQA method relies on the presence of particular content words and phrases, sites could be artificially constructed to achieve very high AQA scores while at the same time delivering bad advice. It is however, not susceptible to link spam.

Spam rejection procedures are an inevitable fact of life for search engines; we see no reason why anti-spam measures cannot be adapted to deal with any targeting of AQA scores by spammers.

5.2 Comparison with 2004 study

BPS and GoogleD engines were both included in a comparitive evaluation conducted in 2004 [17], using the same queries and judging methodology. Although there are many uncontrolled factors (such as employment of different judges) affecting comparisons between the two studies, it is worth noting that over the intervening period, GoogleD has increased its quality performance substantially, while BPS has remained roughly the same. BPS now returns fewer relevant documents compared to the last experiment (MAP score declined from 0.319 to 0.256) while GoogleD returns more (MAP increased from 0.407 to 0.554).

We hypothesize that the marked deterioration in BPS performance is due to the lack of maintenance of seed list and patterns[10]. If confirmed, this would validate our expectation that a manually constructed vertical search portal would require regular, time-consuming maintenance to retain high search effectiveness.

[10] The BluePages operator confirms that very little maintenance has been undertaken.

6 Conclusions

In this study, we have evaluated and compared representatives of three different techniques for constructing a domain specific search portal for the topic of depression: manual seedlist and rules, focused crawling, and filtering/reranking of a general search engine. Our evaluation considered not only the ability to retrieve relevant information, but the quality of the information retrieved, as measured against evidence-based guidelines.

We found that the best performance was obtained by pre-processing queries (adding the word depression) and filtering or re-ranking Google results using AQA scores. This approach achieved high relevance scores and returned pages with a high probability of providing correct advice. The rank averaging method appeared to outperform both filtering approaches on quality measures but was outperformed by them on relevance measures. Further work may suggest better combining methods than any of the three studied here.

Approaches available in the absence of cooperation with a major search provider failed to achieve the coverage of the Google variants. The apparent decline in relevance and quality scores for BPS since the initial study two years earlier almost certainly illustrates the effect of lack of maintenance. It highlights the amount of effort required to maintain a service based on manually defined seedlist and inclusion patterns.

Relevance and quality results for the quality-focused crawling approach, coupled with previous analysis of crawl progress, suggests that this approach to subject portal search is much more viable than manual definition. With additional engineering and a more extensive crawl, it is likely that relevance and quality scores could be improved over those reported here.

An obvious direction for future work is to confirm that the AQA is capable of generalisation to other health topics. A study in the domain of obesity is currently under way.

7 Acknowledgments

The authors gratefully acknowledge the diligence and skill of the assessors.

References

1. G. Berland, M. Elliott, L. Morales, J. Algazy, R. Kravitz, M. Broder, D. Kanouse, J. Munoz, J. Puyol, M. Lara, K. Watkins, H. Yang, and E. McGlynn. Health Information on the Internet: Accessibility, Quality, and Readability in English and Spanish. *Journal of the American Medical Association*, 285(20):2612–2621, 2001.
2. L. Bin and K. Lun. The retrieval effectiveness of medical information on the web. *International Journal of Medical Informatics*, 62:155–163, 2001.
3. Centre for Evidence Based Mental Health. A systematic guide to the management of depression in primary care: treatment. `www.psychiatry.ox.ac.uk/cebmh/guidelines/depression/depression1.html`. Accessed: 26 Oct 2005.

4. M. Clarke and A. Oxman. *Cochrane Reviewers' handbook 4.1.1.* The Cochrane Library, Oxford, 2001.
5. G. Eysenbach, J. Powell, O. Kuss, and E.-R. Sa. Empirical studies assessing the quality of health information for consumers on the world wide web. *Journal of the American Medical Association*, 287(20):2691–2700, 2002.
6. S. Fox. Health information online. PEW Internet & American Life Project, May 2005. `/www.pewinternet.org/PPF/r/156/report_display.asp`.
7. 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.
8. 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/`.
9. D. Ilic, T. Bessell, C. Silagy, and S. Green. Specialized medical search-engines are no better than general search-engines in sourcing consumer information about androgen deficiency. *Human Reproduction*, 18(3):557–561, 2003.
10. K. Järvelin and J. Kekäläinen. IR methods for retrieving highly relevant documents. In *Proceedings of SIGIR 2000*, pages 41–48, Athens, Greece, 2000.
11. A. Jorm, H. Christensen, K. Griffiths, A. Korten, and B. Rodgers. *Help for depression: What works (and what doesn't).* Centre for Mental Health Research, Canberra, Australia, 2001.
12. A. Jorm, H. Christensen, K. Griffiths, A. Korten, and B. Rodgers. Effectiveness of complementary and self-help treatments for depression. *Medical Journal of Australia*, 176:S84–S96, 2002.
13. S. E. Robertson. On term selection for query expansion. *Journal of Documentation*, 46(4):359–364, 1990.
14. S. E. Robertson, S. Walker, M. Hancock-Beaulieu, and M. Gatford. Okapi at TREC-3. In *Proceedings of TREC-3*, pages 109–126, November 1994. NIST special publication 500-225.
15. D. J. Sheskin. *Handbook of Parametric and Nonparametric Statistical Procedures.* Chapman and Hall/CRC, Boca Raton, Florida, USA, fourth edition edition, 2007.
16. E. Sormunen. *A method for measuring wide range performance of Boolean queries in full-text databases.* PhD thesis, University of Tampere, 2000. http://acta.uta.fi/teos.phtml?3786.
17. 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`.
18. 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`.