# Focused Crawling for both Topical Relevance and Quality of Medical Information

Thanh Tin Tang
Department of Computer Science, ANU
Canberra, Australia
tim@cs.anu.edu.au

David Hawking
CSIRO ICT Centre
Canberra, Australia
david.hawking@acm.org

Nick Craswell
Microsoft Research
Cambridge, UK
nickcr@microsoft.com

Kathy Griffiths
Centre for Mental Health Research
ANU, Australia
kathy.griffiths@anu.edu.au

## ABSTRACT

Subject-specific search facilities on health sites are usually built using manual inclusion and exclusion rules. These can be expensive to maintain and often provide incomplete coverage of Web resources. On the other hand, health information obtained through whole-of-Web search may not be scientifically based and can be potentially harmful.

To address problems of cost, coverage and quality, we built a focused crawler for the mental health topic of depression, which was able to selectively fetch higher quality relevant information. We found that the relevance of unfetched pages can be predicted based on link anchor context, but the quality cannot. We therefore estimated quality of the entire linking page, using a learned IR-style query of weighted single words and word pairs, and used this to predict the quality of its links. The overall crawler priority was determined by the product of link relevance and source quality.

We evaluated our crawler against baseline crawls using both relevance judgments and objective site quality scores obtained using an evidence-based rating scale. Both a relevance focused crawler and the quality focused crawler retrieved twice as many relevant pages as a breadth-first control. The quality focused crawler was quite effective in reducing the amount of low quality material fetched while crawling more high quality content, relative to the relevance focused crawler.

Analysis suggests that quality of content might be improved by post-filtering a very big breadth-first crawl, at the cost of substantially increased network traffic.


*CIKM'05,* October 31–November 5, 2005, Bremen, Germany.


## Categories and Subject Descriptors

H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval—*information filtering, retrieval models*

## General Terms

experimentation, performance, measurement

## Keywords

quality health search, focused crawling, domain-specific search

## 1. INTRODUCTION

A survey of US Internet users found that forty percent of respondents used the Internet to find advice or information about health or health care [2]. However, other studies have shown that medical information on the Internet can be fraudulent, of dubious quality and potentially dangerous [18, 23].

It is desirable that a search service over health web sites should return results which are not only relevant to the query but in accord with evidence-based medical guidelines. Health experts, based on either scientific evidence or accountability criteria, have developed protocols for manual assessment of medical web site quality [12, 8]. However, there is very little prior work on using automated quality assessments, either in determining what to index or how to rank potential search results. One exception, due to Price and Hersh [21], reranks results from general search engines based on automated ratings of relevance, credibility, absence of bias, content currency and value of links.

ANU's Centre for Mental Health Research operates a web site[1] which publishes evidence-based information on depressive illness and also provides integrated search of over 200 depression sites. Currently, the set of indexed sites is manually maintained, using a seed list and URL-based inclusion rules that determine which parts of each site are indexed. Here we report our experiences in developing a fully automatic alternative, using a focused-crawler that takes into account relevance and quality.

[1] `bluepages.anu.edu.au/`

## 2. BACKGROUND AND RELATED WORK

### 2.1 Assessment of the Quality of Information on Medical Sites

The ultimate measure of the quality of a health web site is its effect on health outcomes but it is not usually feasible for website publishers or visitors to obtain that information. Next best would be an assessment of the extent to which the content of the site is consistent with the best available scientific evidence — evidence-based medicine — but determining this requires expert raters.

Therefore in the present study, experts rate our crawled sites on a 21-point scale derived by Griffiths and Christensen [13]. These ratings are based on a set of evidence-based depression guidelines published by the Centre for Evidence Based Mental Health (CEBMH) [5].

There are also rating schemes for non-experts such as Silberg [27] and DISCERN [8]. They focus on accountability criteria which could be measured by people without extensive medical expertise, such as whether the author is identified and whether the site has been recently updated. However, a study of depression web sites by Griffiths and Christensen [12] found no correlation between Silberg scores and expert evidence-based ratings. The latter was found to be correlated with DISCERN scores [14], but carrying out such manual assessments is a lengthy process.

In the Web search literature, link graph measures such as PageRank [4] have been promoted as indicators of quality, but how this type of quality might correlate with a medical definition has been little studied. A very recent study by Griffiths and Christensen [14] found only a moderate correlation between Google-reported PageRank and the 21-point rating scale. In this study we follow a content-based approach.

### 2.2 Relevance Feedback

Relevance feedback (RF) is a well-known IR approach of 'query by example'. Given example sets of relevant documents, the goal is to find more of the same. In this paper, we use this both to identify depression-relevant pages and high-quality depression-relevant pages. Our specific application of RF is described in more detail in Section 3.1.

We applied Robertson's approach to term selection [24]. In this approach, there are three ways to calculate the selection value for a term: using the probability of the term occurring in a relevant document ($r/R$), rewarding terms that occur frequently in relevant documents ($\sum_{reldocs} tf/R$), or the average of these. We used the third approach, computing the selection value of a term $Q_t$ as:

$$Q_t = w * \frac{r/R + \sum_{reldocs} tf/R}{2} \quad (1)$$

where $R$ is the number of known relevant documents, $r$ is the number of documents in $R$ that contain term $t$ and $tf$ is the frequency of occurrence of the term within a document.

The weight $w$ was calculated using the Robertson-Sparck Jones weight [25]:

$$w = log \frac{(r + 0.5)/(R - r + 0.5)}{(n - r + 0.5)/(N - n - R + r + 0.5)}$$

where $N$ is the number of documents in the collection, $n$ is the number of documents containing a specific term; and $R$ and $r$ are defined as above.

### 2.3 Focused Crawling

First introduced by de Bra et al. [3], and subsequently studied by many others [6, 9, 16], focused crawlers are designed to selectively fetch content relevant to a specified topic of interest using the Web's hyperlink structure.

A focused crawler starts from a seed list of topical URLs. It estimates the likelihood that each subsequent candidate link will lead to further relevant content, and may prioritise crawling order on that basis and/or reject low-likelihood links. Evidence such as link anchortext, URL words and source page relevance are typically exploited in estimating link value.

McCallum et al. [20] used Naive Bayes classifiers to categorise hyperlinks while Diligenti et al. [11] used the context-graph idea to guide a focused crawler. Rather than examining relevant nodes alone, both techniques trained a learner with features collected from paths leading up to the relevant nodes.

Chakrabarti et al. [6], on the other hand, used hypertext graphs including in-neighbours (documents citing the target document) and out-neighbours (documents that target document cites) as input to some classifiers. According to these authors, a focused crawler can acquire relevant pages steadily while a standard crawler quickly indexes a large number of irrelevant pages and loses its way, even though they started from the same seed list.

## 3. CRAWLERS

In this section we introduce three crawlers. First we describe our use of relevance feedback to estimate quality of pages, then our classifier to compute relevance scores for links. Finally we describe the crawlers: breadth-first, relevance-focused and quality-focused.

### 3.1 Relevance Feedback for Page Relevance and Quality

A quality-focused crawler needs some way of predicting the quality of uncrawled URLs, to set its priority. We tried various methods to predict this, using as training data quality-judged depression pages from a previous study [28]. We found it impossible to predict the quality of a link target based on its anchor context alone, so we abandoned attempts to score each link separately. Instead we scored the quality of the whole page and applied this equally to the page's outlinks.

We used relevance feedback to predict page quality. RF was a natural choice here, because a focused crawling framework needs to prioritise the crawling order, and RF gives us scores that can be used in ranking. We also made separate use of relevance feedback in scoring topic relevance for evaluation purposes only. Both quality RF and relevance RF are described in this section. Both use the term selection methods described in Section 2.2 to identify extra query words and phrases. Phrases usually include two adjacent words, but sometimes three words if the middle word is a preposition, for example 'treatment of depression'.

#### 3.1.1 Relevance Query

Using relevance judgments from a previous experiment [28], we selected 347 relevant and 9000 irrelevant documents. We applied the Robertson selection value formula to obtain weights for all the terms in relevant documents. Past re-

**Table 1: Examples of terms in the relevance query.**

| Term | Weight ($Q_t$) | Term | Weight ($Q_t$) |
|---|---|---|---|
| depression | 15 | anxiety | 2.6 |
| health | 6.9 | medication | 2.4 |
| mental | 5.4 | cognitive | 2.1 |
| treatment | 3.3 | patient | 1.8 |
| therapy | 2.7 | symptoms | 1.8 |

**Table 2: Examples of terms in the quality query.**

| Term | Weight | Term | Weight |
|---|---|---|---|
| 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 |

search has suggested that the number of terms that could be usefully added to expand a query might range from 20 to 40 [15]. We arbitrarily selected 20 top weighted single words and 20 top weighted phrases. See examples in Table 1.

#### 3.1.2 Quality Query

From the same previous experiment we identified 107 documents relevant to depression and of high quality, and another set of 3002 documents which were either irrelevant or relevant but not of high quality.

We used the same technique as for the relevance query to produce two candidate term lists: one containing single words and the other containing phrases. However, we used a more sophisticated procedure to choose a term selection cutoff.

We first derived a list of words and phrases representing effective depression treatments from [13]. Multi-word treatments were divided into two-word phrases (e.g. 'cognitive behaviour therapy' would become 'cognitive behaviour' and 'behaviour therapy') in order to match the two-word phrases in the relevance and quality queries described above. We then located these words and phrases in the candidate lists and cut off the lists just after the lowest-ranked occurrence of an effective treatment term. Surprisingly this gave us the same cutoff (20) for phrases and a similar cutoff for single words (29). Some example terms are shown in Table 2.

Note that the two queries include many terms in common, because both are on the topic of depression. High-quality depression-relevant documents are a subset of depression-relevant documents. The quality query contains more words relating to effective treatment methods such as 'cognitive therapy' or antidepressant medications like 'zoloft' and 'paxil'.

#### 3.1.3 Document Scoring Based on Relevance Feedback

We used the Okapi BM25 weighting function [26] to score documents against the two weighted queries:

$$w_t = tf_d * \frac{log\frac{N-n_t+0.5}{n_t+0.5}}{2*(0.25+0.75*dl/avdl)+tf_d} \quad (2)$$

where $tf_d$ is the number of times term $t$ occurs in document $d$, $N$ is the number of documents in the collection, $n_t$ is the number of documents containing $t$, $dl$ is the length of the document and $avdl$ is the average document length.

Scores calculated with BM25 are collection dependent. Rather than assuming a collection of the documents crawled thus far, we chose to assume a more general web context and used values for the collection parameters ($N = 2,376,673$, $avdl = 15,036$ and $n_t$) which were derived from a large general crawl of Australian educational websites. The values for $n_t$ varied depending on what term $t$ was used.

The final score was computed using the following equation:

$$DScore = \sum_{i=1}^{num\_of\_terms} Q_t * w_t \quad (3)$$

where $Q_t$ is obtained from equation 1 and $w_t$ from equation 2. These scores represented either quality or relevance depending on the query.

### 3.2 Decision Tree for Link Relevance

In our previous work we developed a classifier for predicting the relevance of a link target, based on features in the link's source page [29]. We evaluated a number of learning algorithms provided by the Weka package [30], such as k-nearest neighbor, Naive Bayes, and C4.5. Since then we also evaluated Perceptron. The C4.5 decision tree [22] was the best amongst those evaluated.

The classifier 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). If we found multiple links to the same URL, we included all available anchor contexts. This is a relatively standard approach [1, 9, 7].

To produce a confidence score at each leaf node of the decision tree we used a Laplace correction formula [19]:

$$confidence\_level_k = \frac{N_k + \lambda_k}{N + \sum_{k=1}^{K} \lambda_k} \quad (4)$$

where $N$ is the total number of training examples that reach the leaf; $N_k$ is the number of training examples from class $k$ reaching the leaf; $K$ is the number of classes and $\lambda_k$ is the prior for class $k$ and is usually set to be 1. In our case, $K$ is 2 because we only had two classes, relevance and irrelevance.

### 3.3 Combining Quality and Relevance Scores

We used the quality score of a page (computed using relevance feedback) to predict the quality of its outlinks. If more than one known page linked to the same URL, we took the mean quality score of the linking pages. Relevance scores computed from the decision tree were already aggregated across links.

To order the crawl queue for the quality crawler, we combined the quality and relevance scores. The overall score for a URL was given by:

$$URLScore = confidence\_level_{rel} * \frac{\sum_{i=1}^{m} DScore_i}{m} \quad (5)$$

where $confidence\_level_{rel}$ is the URL's relevance score (equation 4), $DScore_i$ using the quality query is a linking page's quality score (equation 3), and $m$ is the number of pages linking to the URL.

A side effect of taking the product is that if one of the two scores is zero, the overall priority score is zero.

The decision to multiply the scores was taken arbitrarily. We plan to investigate different options for balancing relevance and quality in future work, including dispensing with the relevance component altogether. It is possible that this may lead to significant improvement.

### 3.4 Our Three Crawlers

We evaluated three crawlers: the breadth-first (BF) crawler, the relevance crawler, and the quality crawler. When a crawler encounters a new URL that URL is added to a crawl queue, and the crawler proceeds by taking URLs from that queue. The crawlers differ in how their crawl queues are prioritised.

The BF crawler serves as a baseline for comparison. It traverses the link graph in a breadth-first fashion, placing each newly discovered URL in a FIFO queue. This crawler is likely to find some depression pages since we start it from depression-relevant seeds, but we would expect the relevance of its crawl to fall as the crawl progresses.

The relevance crawler is designed to prefer domain-relevant pages, ordering its crawl queue using the relevance decision tree discussed in Section 3.2. The relevance RF score is not used, we reserve it for use in evaluation. By crawling the highest-scoring URLs first, we would expect the relevance crawler to maintain its overall relevance more successfully than the BF crawler.

The quality crawler is designed to prefer higher-quality domain-relevant pages. Each URL is given a score that was computed using equation 5. A major focus of this paper is to evaluate whether the quality crawler can successfully prioritise its queue to maintain the overall quality of its crawl and avoid pages with low quality, potentially harmful advice (with respect to depressive illness).

## 4. EXPERIMENTS AND MEASURES

### 4.1 Relevance Experiment

We used our RF relevance score (applying the relevance query in equation 3), and a score threshold to evaluate the overall relevance of our three crawls. The threshold was found using 1000 relevant and 1000 irrelevant pages from our previous study (these were separate from those used to generate the relevance query). A threshold at 25% of the theoretical maximum BM25 score (of 502.88[2]) minimised the total number of false positives and false negatives, so in our crawls we labeled pages with RF relevance score greater than this threshold as RF-relevant.

Using RF scores rather than real relevance judgments allows us to get some idea of relevance without extensive relevance judging. However, to validate the accuracy of our RF-based 'judgments', we employed two lay relevance assessors[3] to judge the relevance of 300 RF-relevant and 120 RF-irrelevant pages. These pages were randomly selected from all the RF results of all the crawled pages. As for the judging criterion, any page about the mental illness 'depression' was considered relevant.

The level of agreement between the two assessors was high (91.2%) indicating that judging for such a simple topic is easy. The RF-judgments had an accuracy of 89.3%, a 90.9% success rate in predicting the relevance category, and a 84.6% success rate in predicting the irrelevance category. We concluded that these levels were high enough to present some RF-judgment-based results.

Note that this RF classifier was only used in evaluating the relevance of sets of pages returned by the various crawlers. None of these three crawlers used this classifier in deciding priorities of links for crawling.

We evaluated relevance of the three crawlers, each starting from a seed set of 160 URLs taken from the DMOZ depression directory[4]. We evaluated the first 10,000 pages from each crawler according to RF-relevance.

### 4.2 Quality Experiments

Most of the models for assessing the quality of depression content on the Web refer to the entire sites, not individual pages [8, 17]. We therefore grouped all the pages in each crawl into sites. Pages originated from the same host names were considered to be from the same sites.

The quality of the sites was evaluated by a research assistant from the Centre for Mental Health Research using a rating scale derived by Griffiths and Christensen [13] from the CEBMH evidence-based clinical guidelines. Each site was assigned a quality score in the range 0 to 20.

Since judging took 4 hours per site on average, we could not use the full 160 page seed list. If we did, a large amount of effort would be needed just to judge seeds, and these are uninformative with respect to crawl strategy. Therefore we randomly selected 18 URLs from the 160 to use as our quality experiment seeds. We cut off each of our three crawls at 3,000 pages. For this small crawl size, we were able to judge the quality of any site with six or more crawled pages in all the crawls.

We propose three measures to compare crawl quality. Note that, in our measures, the quality score of a page is assigned the quality score of the site containing it.

- Quality score using all crawled pages: We first computed the mean value of the quality scores of all the judged sites. We then transformed the site scores by subtracting the mean, giving negative scores to sites with below-the-mean ratings. The score of a crawl was given by the sum of quality scores of all its judged pages (all pages from quality-judged sites). This means that the quality score captures both the quality of the pages and the size of the crawl.

- Quality score using RF-relevant pages: Not all sites with quality judgments are dedicated to depression, and many contain a large number of irrelevant pages. We used our RF-relevance classifier to identify the relevant pages in each crawl, then calculated the total quality score as above using just those pages.

- AAQ and BAQ comparison: We grouped judged sites into three categories: above average quality (denoted as AAQ, the top 25% of the judged sites), average quality (denoted as AQ, the middle 50%) and below average quality (denoted as BAQ, the bottom 25%). In some tests we focused on the number of crawled pages from the 'extreme' AAQ and BAQ categories.

[2]Corresponding to a hypothetical zero-length document containing infinite numbers of each of the query terms.

[3]university research assistants

[4]`http://www.dmoz.org/Health/Mental_Health/Disorders/Mood/Depression/`

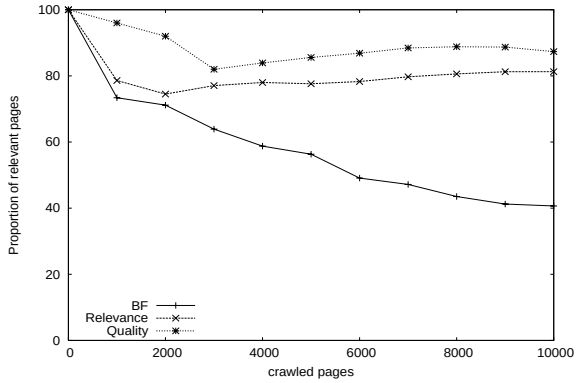


**Figure 1: Comparison of the BF, relevance and quality crawlers for relevance using the RF classifier.**

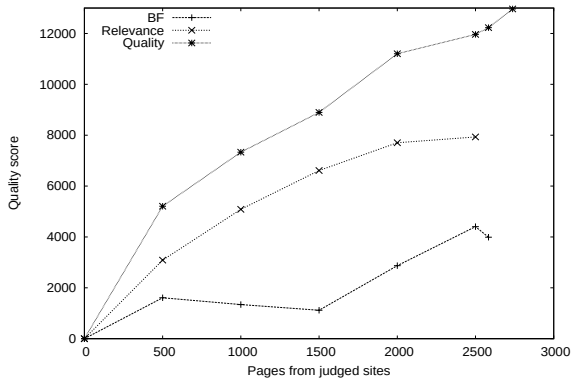


**Figure 2: Quality score for each crawl based on all pages from judged sites.**

# 5. RESULTS AND DISCUSSION

## 5.1 Relevance Results

Figure 1 depicts the relevance levels throughout each of our three crawls, based on RF relevance judgments. The relevance and quality crawls each stabilised after 3,000 pages, at about 80% and 88% relevant respectively. The breadth first crawler continued to degrade over time as it got further from the DMOZ depression seeds. At 10,000 pages it was down to 40% relevant and had not yet stabilised.

The quality crawler outperformed the relevance crawler, and this must be due to the incorporation of the quality RF score. Noticing this, we performed an additional crawl using relevance RF in place of quality RF, and achieved comparable results to the quality crawler. This indicates that RF scores can offer a small improvement in crawl relevance, on top of our relevance decision tree, with the caveat that, in this case only, we used RF techniques both to predict which links to follow and to evaluate relevance of crawled pages.

Our overall conclusion on relevance is simply that our focused crawlers succeed in maintaining relevance as crawls progress.

## 5.2 Quality Results

The quality scores based on all pages from judged sites are shown in Figure 2. All three crawlers achieved positive quality scores. This means they crawled more pages from higher-quality sites than lower-quality ones. Although this is surprising in the case of the breadth first crawler, it may be because higher-quality sites are simply larger. To explore this, we fully crawled ten AAQ sites and ten BAQ sites, all of which were randomly selected. We found that, on average, a BAQ site had 56.6 pages while an AAQ site had 450.2 pages, about eight times higher.

The main finding is that the quality crawler, using the quality RF scores of known link sources to predict the quality of the target, was able to significantly outperform the relevance crawler. Towards the end of the crawls its total quality was over 50% better than that of the relevance crawl.

Figure 3 shows the same total quality scores, but this time only counting pages judged relevant by our RF classifier. The results were similar to the previous figure, particularly for the quality crawler, so we concluded that the presence of irrelevant pages was not a major factor in quality evaluation. The relevance and quality crawlers suffered a little with the elimination of some irrelevant pages from higher-quality sites, whereas the breadth-first crawler benefited from the elimination of irrelevants from lower-quality sites.

Now we focus on the AAQ and BAQ categories.

An interesting set of pages are those that are from AAQ sites and are RF-judged to be relevant. These are the pages we would expect to be most useful in our domain-specific engine. Figure 4 shows the number of these pages in each crawl over time. The quality crawler performed very well, with more than 50% of its pages being AAQ and relevant. The other two performed well too, with over 25% of their pages in that category.

Figure 5 shows the number of pages from BAQ sites, regardless of relevance. The breadth first crawler was much worse on this count than the other two, with two or three times more BAQ pages than the other two. In the quality crawl, only about 5% of the pages were from BAQ sites, and this in combination with the 50% AAQ result underlines the success of the crawler.

Note that the number of AAQ pages was higher than the number of BAQ pages even in the BF crawl. The BF crawler benefited from the seed list in its early stages — we found that the seed list has 4 BAQ but 7 AAQ URLs — and also from the relative sizes of AAQ and BAQ sites. However, in larger crawls the influence of the seed list would become less, and focus would become increasingly important.

# 6. FURTHER QUALITY ANALYSIS

We ran two additional experiments using our quality judgments. One measured the 'quality locality' of linkage between judged sites. The other considered what happens if we post-filter our crawls using our quality scoring formula (equation 3) on the text of the crawled pages, dropping low-quality pages from the system.

## 6.1 Quality Locality Analysis

Topic locality experiments described in [10] indicated that pages typically link to pages with similar content. For a quality-focused crawler to function effectively we hope there is also 'quality locality'. More specifically it would be helpful if higher-quality sites tend to link to each other, making it easier for the crawler to identify more of the same.

We did a breadth first crawl of 100,000 pages starting from the 160 seed URLs on depression. Using these crawled

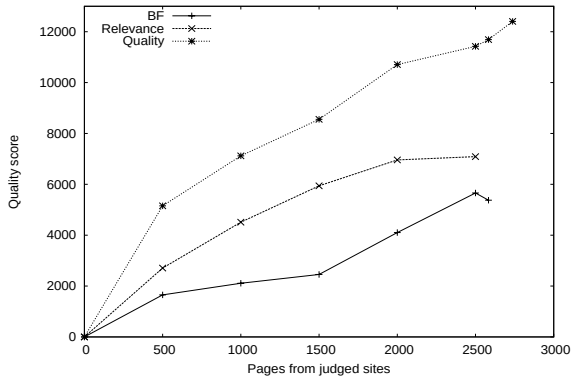


**Figure 3: Quality score for each crawl based on relevant pages from judged sites.**

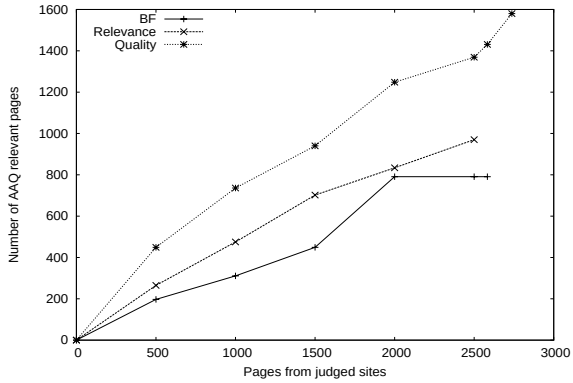


**Figure 4: Number of relevant and above-average-quality pages in each crawl.**

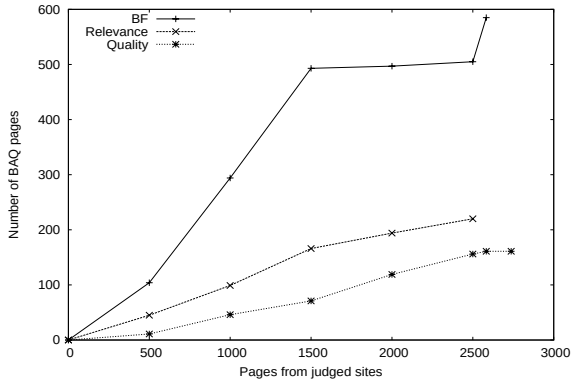


**Figure 5: Number of all below-average-quality pages in each crawl.**

**Table 3: Quality locality analysis according to the link structure between source sites and target sites for a 100,000 page BF crawl.**

| Target type | *Average number of sources* AAQ | AQ | BAQ |
|---|---|---|---|
| AAQ | 2.53 | 1.92 | 0.92 |
| AQ | 1.98 | 1.53 | 0.57 |
| BAQ | 1.46 | 0.83 | 0.36 |

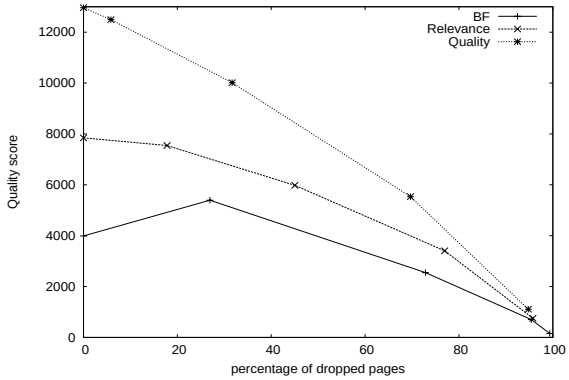


**Figure 6: Quality score for each crawl at different filtering points.**

**Table 4: A comparison of quality scores between the quality crawl and each of the post-filtering BF crawls of different sizes. The number of judged pages were set to 2737, which was the number of pages from judged sites in the quality crawl.**

| Crawl | Quality score | unjudged pages |
|---|---|---|
| BF10000 | 9995.4 | 2273 |
| BF15000 | 9708.4 | 1711 |
| BF20000 | 11454.4 | 1554 |
| **BF25000** | **13523.4** | **1311** |
| Quality | 12964.9 | 263 |

pages, we identified all links between sites, including links to URLs that were not yet crawled. We then analysed linkage between our 114 judged depression sites, in particular calculating the average number of sites of each type linking to sites of other types (Table 3). For example, on average each AAQ site had links from 2.53 AAQ sites, 1.92 AQ sites and 0.92 BAQ sites.

If quality locality were a direct analogue of topic locality, we might expect to see a cluster of AAQ sites linking to each other and another cluster of BAQ sites. What we observed in the linkage between judged sites was a tendency to link to AAQ sites, even amongst links from BAQ sites. This means that no matter which judged site is crawled, the crawler is most likely to find AAQ-site links. We also observed that higher-quality sites had more outlinks. We conclude that the observed link patterns are favourable for quality-focused crawling.

## 6.2 Post-filtering for Quality

We observed pages from BAQ sites in all three crawls (Figure 5). An alternate way of using our RF quality scores is to post-filter our crawls, removing pages with quality scores below some threshold. The question is whether filtering a crawl by RF quality score can improve its overall human-judged quality rating.

In our first post-filtering experiment we progressively applied a stronger filter to our three main crawls (Figure 6). Because below-the-mean sites received negative scores in our scoring system, we expected an increase in total quality scores at certain thresholds where more low quality pages were filtered out. However, we were unable to improve the

quality crawl or the relevance crawl by post-filtering. These crawls already had good overall quality, and our RF quality score was not sufficient to improve on that. We observed some improvement in the breadth first crawl, but it did not overtake the other crawlers.

Since the BF crawler was able to be improved by post-filtering, our second experiment filtered successively larger breadth-first crawls, to see if the quality-focused crawl could be surpassed. The quality crawl contained 2,737 pages from judged sites, so for each BF crawl we set the filtering threshold to give us 2,737 pages from judged sites. Note, this threshold also gave us a large number of pages from unjudged sites, adding some uncertainty to the quality rating.

Table 4 shows the results of the experiment. To surpass the quality rating of the quality crawler we had to increase the breadth-first crawl size to 25,000 pages, compared to 3,000 pages for the quality-focused crawl. This means that if an appropriate threshold can be set and a massive increase in crawl traffic and server load is acceptable, a filtered breadth first crawler is an alternative to a quality-focused crawler. However, certainly at an Australian university that pays over AUD20 per gigabyte of traffic, some focus is desirable.

Finally, there are some experiments we did not perform. We did not consider how the quality score could be incorporated as a ranking feature, at query time. We do not have the necessary per-query relevance and quality judgments to do this. Also we did not consider post-filtering using the RF relevance score. Again, we do not have the necessary human judgments to carry out this experiment. Furthermore, standard IR systems are robust to having irrelevant documents in the crawl and the harm caused by retrieving one is low, so we believe quality filtering is the more important case.

## 7. CONCLUSIONS AND FUTURE WORK

Subject-specific search facilities on health sites are usually built using manual inclusion and exclusion rules, which require a lot of human effort in building and maintenance. We have designed and built a fully automatic quality focused crawler for a mental health topic of depression, which was able to selectively crawl higher quality and relevant content. Our work has resulted in four key findings.

First, domain relevance on depression could be well predicted using link anchor context. A relevance-focused crawler based on this information fetched twice as many relevant pages as a breadth-first control. A combination of link anchor context and source-page relevance feedback improved the prediction slightly further.

Second, link anchor context alone was not sufficient to predict quality of Web pages. Instead, relevance feedback technique proved useful. We used this technique to learn and derive a list of terms representing high quality content from a small set of training data, which was then scored against crawled source pages to predict the quality of the targets. Compared to the relevance and BF crawls, a quality crawl using this approach obtained a much higher total quality score, significantly more relevant pages from high quality sites and fewer pages from low quality sites.

Third, analysis on quality locality suggested that above average quality depression sites tended to have more incoming links and outgoing links compared to other types of site. This observed link pattern is favourable for quality focused crawling, explaining in part why it was able to succeed.

Fourth, quality of content might be improved by post-filtering a very big breadth-first crawl if an appropriate filtering threshold is set. This leads to a trade-off decision between cost and efficiency. The post-filtering approach could be adopted in cases where a massive increase in crawl traffic and server load is acceptable. Although we could not improve our other two crawlers by filtering, it might hypothetically be possible to do so in a larger-scale experiment, and this would be a less wasteful approach than all-out breadth first crawling.

Given the interesting results that we found, there is obvious follow-up work to be done on focused crawling. In particular, it would be interesting to compare our quality crawl with other depression-specific search portals and general search engines in terms of relevance and quality by running queries against these engines and measuring the results.

Another question would be whether we could improve our quality focused crawler. The current approach evaluated links on page basis. Possibly, another quality focused crawler working on site basis, (by accumulating the quality scores of all the crawled pages from the same sites, and crawling new pages according to the predicted quality score of the site containing them) could achieve even better results.

Investigation of whether our findings generalise to other health domains (characterised by an evidence-based notion of quality) or more generally is left for future work.

## 8. ACKNOWLEDGMENTS

We gratefully acknowledge the assistance of Alistair Rendell and Helen Christensen for seeking financial support for the project and the effort of our relevance and quality judges Sonya Welykyj, Michelle Banfield and Alison Neil.

## 9. REFERENCES

[1] C. C. Aggarwal, F. Al-Garawi, and P. S. Yu. On the design of a learning crawler for topical resource discovery. *ACM Trans. Inf. Syst.*, 19(3):286–309, 2001.

[2] L. Baker, T. H. Wagner, S. Singer, and M. K. Bundorf. Use of the internet and e-mail for health care information. *JAMA*, 289(18):2400–2406, 2003.

[3] P. D. Bra, G. Houben, Y. Kornatzky, and R. Post. Information retrieval in distributed hypertexts. In *Procs. of the 4th RIAO Conference*, pages 481–491, New York, 1994.

[4] S. Brin and L. Page. The anatomy of a large-scale hypertextual web search engine. In *WWW7*, pages 107–117, Brisbane, Australia, 1998.

[5] CEBMH. A systematic guide for the management of depression in primary care: treatment. University of Oxford, UK, 1998. Available at `http://cebmh.warne.ox.ac.uk/cebmh/guidelines/depression/treatment.html`, Accessed 30 May 2005.

[6] S. Chakrabarti, M. Berg, and B. Dom. Focused crawling: A new approach to topic-specific web resource discovery. In *WWW8*, 1999.

[7] S. Chakrabarti, B. Dom, P. Raghavan, S. Rajagopalan, D. Gibson, and J. Kleinberg. Automatic resource compilation by analyzing hyperlink structure and associated text. In *Procs. of the WWW7*, pages 65–74, Brisbane, Australia, 1998. Elsevier Science Publishers B. V.

[8] D. Charnock, S. Shepperd, G. Needham, and R. Gann. Discern: an instrument for judging the quality of written consumer health information on treatment choices. *J. Epidemiol Community Health*, 53:105–111, 1999.

[9] J. Cho, H. Garcia-Molina, and L. Page. Efficient crawling through url ordering. In *WWW7*, 1998.

[10] B. D. Davison. Topical locality in the web. In *Procs. of the 23rd annual international ACM SIGIR conference on Research and development in information retrieval*, pages 272–279, New York, NY, USA, 2000. ACM Press.

[11] M. Diligenti, F. M. Coetzee, S. Lawrence, C. L. Giles, and M. Gori. Focused crawling using context graphs. In *Procs. of the 26th VLDB Conference*, Cairo, Egypt, 2000.

[12] K. Griffiths and H. Christensen. Quality of web based information on treatment of depression: cross sectional survey. *British Medical Journal*, 321:1511–1515, 2000. `bmj.bmjjournals.com/cgi/content/full/321/7275/1511`.

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

[14] K. Griffiths, H. Christensen, and S. Blomberg. Website quality indicators for consumers. In *Tromso Telemedicine and e-Health Conf.*, Tromso, Norway, 2004.

[15] D. Harman. Towards interactive query expansion. In *Procs. of the 11th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 321–331, New York, NY, USA, 1988. ACM Press.

[16] M. Hersovici, M. Jacovi, Y. S. Maarek, D. Pellegb, M. Shtalhaima, and S. Ura. The shark-search algorithm. an application: tailored web site mapping. In *WWW7*, 1998.

[17] A. R. Jadad and A. Gagliardi. Rating health information on the internet. *JAMA*, 279:611–614, 1998.

[18] R. Kiley. Quality of medical information on the internet. *J. Royal Soc. of Med.*, 91:369–370, 1998.

[19] D. D. Margineantu and T. G. Dietterich. Improved class probability estimates from decision tree models. In D. D. Denison, M. H. Hansen, C. C. Holmes, B. Mallick, and B. Yu, editors, *Lecture Notes in Statistics. Nonlinear Estimation and Classification*, volume 171, pages 169–184, New York, 2002. Springer-Verlag.

[20] A. McCallum, K. Nigam, J. Rennie, and K. Seymore. Building domain-specific search engines with machine learning technique. In *Procs. of AAAI Spring Symposium on Intelligents Engine in Cyberspace*, 1999.

[21] 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 *Procs. of the AMIA 1999 Annual Symposium*, pages 911–915, Washington DC, 1999.

[22] J. R. Quinlan. *C4.5: programs for machine learning*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 1993.

[23] A. Risk and J. Dzenowagis. Review of internet health information quality initiatives. *JMIR*, 3(4):e28, 2001.

[24] S. E. Robertson. On term selection for query expansion. *J. Doc.*, 46(4):359–364, 1990.

[25] S. E. Robertson and K. S. Jones. Relevance weighting of search terms. *Journal of the American Society for Information Science*, 27(3):129–146, 1976.

[26] S. E. Robertson, S. Walker, S. Jones, M. Hancock-Beaulieu, and M. Gatford. Okapi at trec-3. In *Procs. of the Third Text REtrieval Conference*, pages 109–126, USA, 1996.

[27] W. M. Silberg, G. D. Lundberg, and R. A. Musacchio. Assessing, controlling, and assuring the quality of medical information on the internet. *JAMA*, 277:1244–1245, 1997.

[28] T. T. Tang, N. Craswell, D. Hawking, K. M. Griffiths, and H. Christensen. Quality and relevance of domain-specific search: A case study in mental health. *To appear in the Journal of Information Retrieval - Special Issues*, 2005.

[29] T. T. Tang, D. Hawking, N. Craswell, and R. S. Sankaranarayana. Focused crawling in depression portal search: A feasibility study. In *Procs. of the Ninth ADCS*, pages 2–9, Australia, 2004.

[30] I. H. Witten and E. Frank. *Data Mining: Practical machine learning tools with Java implementations*. Morgan Kaufmann, San Francisco, 1999.