Quality and relevance of domain-specific search: a case study in mental health

Thanh Tin Tang

Department of Computer Science
CSIT Building, ANU
Canberra, ACT 0200, Australia

thanh.tang@cs.anu.edu.au

Nick Craswell and David Hawking

CSIRO ICT Centre
CSIT Building, ANU
Canberra, ACT 2601, Australia

nick.craswell@csiro.au, david.hawking@csiro.au

Kathy Griffiths and Helen Christensen

Centre for Mental Health Research, ANU
Canberra, ACT 0200, Australia

kathy.griffiths@anu.edu.au, helen.christensen@anu.edu.au

October 1, 2004

Abstract

When searching for health information, results quality can be judged against available scientific evidence: Do search engines return advice consistent with evidence based medicine? We compared the performance of domain-specific health and depression search engines against a general-purpose engine (Google) on both relevance of results and quality of advice. Over 101 queries, to which the term 'depression' was added if not already present, Google returned more relevant results than those of the domain-specific engines. However, over the 50 treatment-related queries, Google returned 70 pages recommending for or against a well studied treatment, of which 19 strongly disagreed with the scientific evidence. A domain-specific index of 4 sites selected by domain experts was only wrong in 5 of 50 recommendations. Analysis suggests a tension between relevance and quality. Indexing more pages can give a greater number of relevant results, but selective inclusion can give better quality.

Keywords Domain specific search, focused crawling, mental health, depression

1 Introduction

Searching for health information is a common activity on the Internet. Forty percent of respondents in a study of US Internet users reported using the Internet to look for advice or information about health or health care [4]. In Excite logs from 1997, 1999 and 2001, the proportion of queries relating to 'health or sciences' was 7–10 percent [24].

Two important avenues for health search are general-purpose search engines and domain-specific (portal) search services. General engines, such as Google, index pages from a general crawl of the Web. They index a very large number of pages from a very wide variety of sources, although the majority of pages are about health. A domain-specific engine, on the other hand, indexes documents relevant to a particular domain such as health or mental health.

This study evaluates domain-specific engines against the general-purpose engine Google, in order to better understand the relative merits of the two types of engine. It investigates whether the time, resources and effort required to operate a domain-specific search engine can be justified in terms of quality or quantity of search results.

We chose depressive illness (mental health) as the domain of interest, because it is among the most common reasons why people search for health information [8], and because it is known that some of the available information is of poor quality. [12]. Our objective was to evaluate both the relevance of search results, using well established IR (Information Retrieval) methodology, and also the quality of advice, in terms of evidence-based medicine. Evidence-based medicine (see [9]) relies upon systematic reviews of scientific research.

This study compares general and domain-specific engines in terms of both precision/recall and quality, using actual queries submitted to depression portal or general search services. It is likely that most of these queries would have been submitted by consumers or members of the general public rather than by health practitioners.

Section 2 provides background to the study and reviews past work in the area. Section 3 outlines our experimental methodology while Section 4 presents and discusses results obtained. Section 5 concludes and suggests directions for future work.

2 Related work

2.1 General-purpose and domain-specific engines

General-purpose search engines such as Google[1] and Yahoo[2] process queries over a very large number of Web pages. The pages are collected in a general crawl of the Web, so any available Web page may be included.

A domain-specific engine limits its index to pages corresponding to a particular subject area, publisher or purpose. These are often a subset of the pages available to a general engine. In practice, this subset is often chosen by including hand-picked Web sites. For example, the BPS engine studied here was built by manually identifying areas on 207 Web servers with information on depressive illness. Another evaluated engine is based on just 4 carefully chosen sites.

We identified two potential advantages of domain-specific search, both relating to the subset of pages searched. Domain-specific search might provide more relevant results, since it indexes a non-uniformly chosen, relevant subset. It might also provide higher quality results, if its subset includes high-quality information sources and avoids pages with false, harmful or misleading information.

Use of specific search engines for reasons of quality was observed in a study of knowledge workers [22]. It found that those whose jobs depend on accuracy, such as journalists, producers, marketing consultants and historians tend to search branded sites such as Encyclopedia Britannica, official societies and universities. Such users were found to employ general search engines only 19% of the time.

Manually identifying a domain-specific set of Web sites for indexing requires significant and ongoing human effort. To improve the situation, McCallum et. al. [18] suggested automating many aspects of creating and maintaining domain-specific search engines by using machine learning techniques. Focused crawlers, for crawling a topic-focused set of Web pages, have been frequently studied [1, 6, 7, 14, 19, 20].

2.2 Depression information on the Web

Depression is a major public health problem, being a leading cause of disease burden [21] and the leading risk factor for suicide. However, many people with depression receive no professional help [2]. Recent research has demonstrated that high quality web-based depression information can improve public knowledge about depression and is associated with a reduction in depressive symptoms [8]. Thus, the web is a potentially important resource for people with depression. However, a great deal of depression information on the Web is of poor quality when judged against the best available scientific evidence [11, 12]. It is therefore important that users can locate depression

[1] http://www.google.com

[2] http://search.yahoo.com/

information which is both relevant and of high quality.

Eysenbach and Kohler [10] studied how consumers search for and appraise health information on the world wide web. They found that most of the time people used general search engines as a starting point instead of medical search engines. As for the queries entered into search engines, users tended not to enter a combination of words but only a single word. Participants usually looked at results on the first page (top 10), and if they coudn't find the information, tended to rephrase the query rather than exploring the second page of the results. Very few internet users later remembered from which websites they retrieved information or who published the sites.

2.3 Effectiveness of domain-specific search

A study by Ilic et. al. [15] compared five medical search-engines with four general search-engines in sourcing consumer information about androgen deficiency (ADAM). The relevance results for all engines were very low. The highest precision was achieved by Google, but was only 4%. Two factors might explain this. First, all queries were judged against the criterion 'How is ADAM recognised as a medical condition and what treatment regimes are available?'. This was the case even for more specific queries like 'steroids' and 'low libido'. In such experiments it is more usual for the judging criteria and query to match. Second, the queries were not all specific to ADAM. It is not surprising if Google fails to return information on ADAM given the query 'steroids'. Depending upon what queries are actually submitted by those seeking information on steroid treatments for ADAM, a fairer test might have been to use the query 'steroids ADAM' or 'steroids androgen deficiency'.

Bin and Lun [5] studied the retrieval effectiveness of medical information on the Web using eight different search tools, among which three were Medical search engines and two were general search engines. The search types covered in their study included single keyword search and question answering. The results showed that, of all the pages returned for each type of search tools, there was no significant difference in the proportion of medically related pages. Overall, there was no trend to indicate that medical specific search tools were better than general search tools in searching for medical information.

3 Experimental methodology

We conducted a standard information retrieval experiment, running queries against engines, pooling the results for each query, and employing research assistants to judge them. The novel features of the experiment are its domain-specific nature and the judgment of results quality according to evidence based medicine.

To shed additional light, a domain-specific engine (HFS) was selected from the general health domain to complement those from the depression domain.

3.1 Engines

Table 1 lists the search engines included in our study[3]. BPS and HFS relate to the search functions of the BluePages and HealthFinder sites respectively. We have used the labels BPS and HFS rather than BluePages and HealthFinder to emphasise that search is only one of the functions of these portal sites.

For depression-specific engines, users can run a query such as 'chocolate' and expect to obtain results about how chocolate relates to depression. However, a more general search engine cannot be expected to infer the depression context. Accordingly, for the general search engine (Google), we added an additional condition (GoogleD) in which the query was augmented with the term 'depression', as in 'chocolate depression', if it was not already present. This ensured that engines which are not specific to depression have a chance of returning relevant results, even if the original query was not specific enough. All results are judged in the context of depression, even if the query term 'depression' is not present.

[3] **Declaration of interest:** At the time of the study, Griffiths and Christensen were operators of the BluePages portal and Hawking and Craswell were members of the team behind the commercially available search engine (Panoptic) which provides the BPS and 4sites search capability.

Table 1: The search engines included in the study. Note that HFSD was added after the main experiment, for completeness.

Engine	URL	Pages in index	Notes
BPS	`bluepages.anu.edu.au/search.html`	12,177	Depression specific
4sites	not publicly available	784	Index of four high quality depression sites.
HFS	`healthfinder.gov`	1,700+ sites	Health specific
HFSD	`healthfinder.gov`	1,700+ sites	HFS with 'depression' added to queries
Google	`google.com.au`	3,300,000,000	General Internet search
GoogleD	`google.com.au`	3,300,000,000	Google with 'depression' added to queries

Note that a general search engine such as Vivisimo[4] which clusters search results by sub-topic may be able to identify a depression sub-category for some queries such as 'paxil'. However, we observed that many queries, such as 'exercise', 'lemon balm' and 'chocolate' did not result in depression-specific clusters.

Google is a very popular and highly effective whole-of-Web search engine [13]. Because of its broad collection policy and huge size, it might be expected to have very good coverage of depression information. The Google crawling algorithm probably attempts to crawl high-quality pages first [7], so there may be some degree of quality filtering in effect.

BPS is a search service offered as part of the existing BluePages depression information site. Its index was built by manually identifying and crawling areas on 207 Web servers containing depression information. Areas sometimes included all of a server's pages, sometimes a URL subtree and sometimes only certain specific URLs. Crawling, indexing and search were performed by CSIRO's Panoptic search engine. BPS will inevitably miss some relevant pages, because it is based on a hand-made list rather than a 3 billion page crawl. When building BPS, no special measures were taken to exclude low-quality information, but some sites were excluded according to the rules listed in Table 2. Besides coverage, differences between BPS and Google could emerge due to use of different software and different ranking algorithms.

Griffiths and Christensen [12] systematically rated treatment advice on Australian web sites that pro-

Table 2: The types of site which were excluded during construction of the BPS search index.

1. Newspaper articles (as these are likely to be ephemeral).
2. Site unavailable/dead link.
3. Forbidden entry.
4. Not relevant to clinical depression.
5. Significant amount of information not relevant to depression. Three online pharmacies were excluded on this basis. Although they contained some specific pages on antidepressants, it was too cumbersome to enter the individual URLs.
6. Non-English.
7. Potentially distressing, offensive or destructive material (e.g. "depression is a punishment from God"), including material which might promote suicide (e.g. a site which featured a picture of a noose next to information about suicide).
8. Duplicates (e.g. one site simply redirecting to another) or very near duplicates (e.g. WebMD and MSN).

[4] `vivisimo.com`

vide information about depression. The consistently best scoring sites included two university-based sites (BluePages and CRUfAD[5]), the site of the National Depression Initiative (beyondblue[6]) and the privately owned site InfraPsych[7]. These sites had the best average ranks across the four main content measures, and achieved top scores on the evidence-based guideline scale and top ratings on at least three of the content measures. We therefore selected these 4 sites containing high quality information on depression and used Panoptic to crawl and index their 784 pages. We expected 4sites to return high quality advice, but perhaps not to include enough pages to answer all 101 queries well.

HealthFinder is a health portal sponsored by the U.S. government, designed to provide information for consumers from specially chosen health-related sites. The "About Us" and "selection policy" pages on the HealthFinder site provides detailed information about which sites are included in its search facility (HFS). In summary, there is a focus on government and nonprofit sources. We have no accurate information about what search technology is used but the "Search Tips" page on the site states:

> The search results are returned in order of relevance to your search terms. Relevancy is calculated based on several factors, such as whether the search term occurs in the name of the resource or organization and how many times in occurs in the description, and how closely search terms occur to each other.
>
> You don't have to choose the usual search options of Exact Match, All Words, or Any Words – the software uses all options and returns the best matches first. It also searches for common variations of the words you enter like plural forms and -ing endings.

Wu and Li [25] stated that HealthFinder was one of the best sources of valuable and reliable consumer health information.

3.2 Queries

Our aim was to judge both relevance and quality of search results. We measured relevance over 101 queries, comprising 50 treatment queries and another 51 depression queries, relating to causes, symptoms and other depression topics.

Forty-five of the treatment queries were the names of depression treatments for which we have evidence based ratings, produced by domain experts at the Centre for Mental Health Research. These ratings are published in [17] and also on the BluePages depression information site. The rating system is:

- Very effective (8 treatments): These treatments are very useful. They are strongly supported as effective by scientific evidence.
- Effective (5 treatments): These treatments are useful. They are supported by scientific evidence as effective, but the evidence is not as strong.
- OK (17 treatments): These treatments are promising and may be useful. They have some evidence to support them, but more evidence is needed to be sure that they work.
- Unsure (12 treatments): These treatments have not been properly researched. It is not possible to say whether they are useful or not.
- Not Effective (8 treatments): On the available evidence, these treatments do not seem to be effective.

In rating the effectiveness of medical treatments, the following levels of evidence are recognized. (Quoted from the BluePages site[8].)

- Randomised controlled trials (RCTs, the best evidence): In an RCT, the people who volunteer to test out the treatment are randomly placed either in a treatment group (eg, given antidepressants) or a no treatment group (eg, given a sugar pill).
- Controlled trial, not randomised (the next best evidence): Sometimes scientists use controlled

[5] http://www.crufad.com/cru_index.htm

[6] http://www.beyondblue.org.au

[7] http://www.infrapsych.com/root/1033/default.htm

[8] http://bluepages.anu.edu.au/rating.html

trials where volunteers are not randomly placed in groups.

- Before and after group study: Another type of evidence involves measuring health before and after treatment.

- Little or no evidence: Sometimes people claim that a treatment works on the basis of their personal or professional experience.

Because of the availability of effectiveness ratings for these treatments, we were able to use the treatment queries to measure the quality of information returned by the search engines. We added five additional treatment queries, selected among specific antidepressants, to make up a set of 50 treatment queries. These antidepressants are known to be very effective.

The non-treatment queries came from two query log sources. To represent domain-specific queries, we used queries from BPS logs. To represent general-purpose queries, we used the Overture Search Terms Suggestion Tool[9], which covers queries submitted to general-purpose engines such as MSN and Yahoo. Rather than giving an overall query list, it presents a list of queries related to a specified search term in order of decreasing frequency of submission during the past month. For example, in March 2004, entering the word depression resulted in depression (279,696), great depression (86,836), manic depression (19,989), clinical depression (15,516), teen depression (13,073) and depression glass (11,379), etc.

We collected query suggestions from the Search Terms Suggestion Tool in November 2003, using depression related terms from the 2004 MeSH[10] database as input. MeSH, which is a well known, objective, independent classification system for medical information, was used to create an independent selection of terms relevant to depression. Terms selected were medical subject headings relevant to depression and their associated entry terms.

We took 28 queries from each source by arranging them in decreasing order of frequency and eliminating queries that were treatment queries and those which related to bipolar disorder (outside the scope of our experiment, and outside the scope of BPS and 4sites). Queries with spelling errors were corrected. Queries unlikely to be about mental health, such as ‘the great depression’ were also eliminated. Vetting was performed by two of the present authors (Griffiths and Christensen) with content expertise in depression.

The two lists contained 5 queries in common. The final list of 51 non-treatment queries comprised 23 from domain-specific query logs, 23 from general-purpose query logs and 5 which occurred in both.

Table 3 shows arbitrarily chosen queries from the the two lists and from the treatments list.

3.3 Result assessors and judging criteria

Our relevance judges were postgraduate students with no connection to any of the services studied. There was no need for them to be health professionals as they were not required to make quality assessments. Instead, they were asked to judge whether or not a page recommended the treatment. The judges were provided with instructions and training on a test query.

Relevance judging was applied to all 101 queries, while quality ratings were only applied to the 50 treatment queries. The few pages that were in foreign languages were assumed to be irrelevant. Pages were assessed based on content visible in a web browser, without following links.

Judging was blind. I.e. judges were not aware which engine or engines had returned the results they were evaluating.

3.3.1 Relevance judging

We used the four level relevance judging scheme developed at the University of Tampere [23]:

- 0 - The document does not contain any information about the topic

[9] `http://inventory.overture.com/d/searchinventory/suggestion`

[10] National Library of Medicine `http://www.nlm.nih.gov/mesh/meshhome.html`

Table 3: Examples of each type of query.

Overture/Mesh (OM)	BPS Logs (BP)	Treatments (TM)
adolescent depression	alcohol	acupuncture
anxiety depression	alcohol and fluoxetine metabolism	alcohol avoidance
beck depression inventory	anxiety	alcohol for relaxation
childhood depression	chemical imbalance in depression	antidepressants
chronic depresison	depression type	aromatherapy
clinical depression	domestic violence	avanza
committ suicide	dysthymia	bibliotherapy
major depression	famous	caffeine avoidance
depression help	gay	chocolate
depression quiz	genetic explanations for depression	cipramil
depression symptoms	hormones depression	cognitive behaviour therapy

- 1 - The document only points to the topic. It does not contain more or other information than the topic description. Typical extent: one sentence or fact.

- 2 - The document contains more information than the topic description but the presentation is not exhaustive. In case of a multi-faceted topic, only some of the sub-themes or viewpoints are covered. Typical extent: one text paragraph, 2-3 sentences or facts.

- 3 - The document discusses the themes of the topic exhaustively. In case of a multi-faceted topic, all or most sub-themes or viewpoints are covered. Typical extent: several text paragraphs, at least 4 sentences or facts.

3.3.2 Recommendation judging

As noted above, judging the quality of a page's advice would have been beyond the capabilities of our assessors, since they werre not medical experts. Therefore, for treatment queries, assessors judged whether the treatment was recommended:

- positive - The document supports or recommends the treatment for depression

- negative - The document opposes the treatment for depression

- neither - The document doesn't mention whether the treatment is good or bad for treating depression

We were thus able to judge the quality of advice based on scientific evidence. For example, if the treatment is strongly supported by scientific research, recommending it is good advice and recommending against it is bad advice. If a treatment has proven ineffective, such as taking tranquilisers or avoiding sugar, recommending it is bad advice.

3.4 Measures

We measured both relevance and quality.

3.4.1 Relevance measures

Two measures used for analysing the relevance of the results were: modified average precision (MAP) and normalized discounted cumulative gain ($NDCG$).

We used the standard formulation of average precision, but with a different denominator, to take into account the fact that a maximum of ten results were retrieved:

$$MAP = \frac{\sum_{i=1}^{num\_rel\_ret(n)} i/rank(i)}{R}$$

where $rank(i)$ is the rank of the ith relevant document and $num\_rel\_ret(n)$ is the number of relevant documents in the top n results (in our experiment $n = 10$), *num_rel_ret* is the total number of known relevant documents in collections being searched, and

$$R = \begin{cases} num\_rel\_ret & \text{if } num\_rel\_ret < 10 \\ 10 & \text{otherwise} \end{cases}$$

For computation of MAP, we converted four-point relevance into binary scores by classing scores of 2 and 3 as relevant, and 0 and 1 as irrelevant. This choice of threshold was somewhat arbitrary. We obtained very similar results when scores of 1 were also classed as relevant.

We report the mean of the MAP scores across the 101 queries.

We also measured mean normalised discounted cumulative gain [16]. Unlike MAP, $NDCG$ takes into account degrees of relevance. To calculate this measure it is necessary to go through a number of steps to determine the gain, the cumulative gain, the discounted cumulative gain and the ideal discounted cumulative gain.

The Gain (G) of each document is its relevance score, which in this case is 0, 1, 2 or 3. Thus,

$$G[n] = Score_{doc\_i}$$

where $Score_{doc\_i}$ is the score of the ith document in the retrieved list.

Cumulative gain (CG) is calculated as follows.

$$CG[1] = G[1]$$

$$CG[i] = CG[i-1] + G[i] \; if \; i > 1$$

Discounted cumulative gain (DCG) is similar to cumulative gain but a discount factor is applied to reflect the decreased utiltity of documents retrieved further down the ranking. Usually base $b = 2$ is used for the discount factor;

$$DCG[i] = CG[i] \; if \; i < b$$

$$DCG[i] = DCG[i-1] + G[i]/\log_b i + 1 \; if \; i >= b$$

Normalized discounted cumulative gain ($NDCG$) is the ratio of the discounted cumulative gain and the ideal discounted cumulative gain. The way to compute the ideal discounted cumulative gain ($IDCG$) is similar to calculating discounted cumulative gain, but using an ideal ranking. The ideal ranking arranged documents in descending order of relevant scores. It starts with all threes, followed by all twos and then all ones.

$$NDCG[i] = \frac{DCG[i]}{IDCG[i]}$$

We report mean $NDCG$ across the 101 queries.

3.4.2 Quality measures

We measured quality using a combination of recommendation judgments and evidence-based treatment ratings. We obtained recommendation judgments for each of the 50 treatment queries. Judgments specified whether the page was positive, negative or neither toward using the query treatment. We also used an evidence-based rating system for treatments: very effective, effective, OK, unsure and not effective. We used two quality measures, one based on a rating scale and one which only counted extreme examples of correct and incorrect advice.

The rating scale is outlined in Table 4. The scores were attached by consensus of two domain experts (authors KG and HC) in blind fashion, without looking at the experimental data. It was based on the domain experts' judgment of how good or bad a particular recommendation was considered to be and turned out to be non-symmetric. For example, for the same very effective treatment, there would be an award of four points (rating = 4) if the treatment is recommended but a penalty of five points (rating = -5) if it is recommended against.

For the first measure, the quality score for each engine was computed as follows.

$$QS = \sum_{all \;\; treatment \;\; ratings} (PP * PR + NP * NR)$$

where QS is quality score; PP (positive pages) and NP (negative pages) are the number of pages recommend for and against all treatments of the same rating respectively; PR (positive rating) and NR (negative rating) are the scores taken from the *Positive* and *Negative* columns in Table 4 respectively.

Table 4: 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 effective	4	-5
Effective	3	-4
OK	1	-2
Unsure	-1	0
Not Effective	-5	4

For the second measure, we defined 'good' and 'bad' treatments and judged 'correct' and 'incorrect' advice relative to them. A 'good' treatment was one which was rated as very effective or effective. A 'bad' treatment was one which was rated as not effective. We ignored treatments rated OK and unsure because they would be less significant in measuring system quality. We then defined 'correct' advice as recommending for good or against bad treatments. We defined 'incorrect' advice as recommending against good or for bad treatments. We then counted the number of instances of correct and incorrect advice.

4 Results and discussion

The 101 queries identified in section 3.2 were run on the 6 engines selected in section 3.1, taking a maximum of 10 results from each engine for each query. These 4325 results were then examined and judged by research assistants as described in Section 3.3.

4.1 Relevance results

Table 5 presents results for our two relevance measures. GoogleD returned the greatest number of relevant results, for both measures, followed by BPS, 4sites and Google. The results were consistent for both measures. HFSD returned the least number of relevant documents. We did t-tests (with a criterion of 95% confidence) on selected pairs of engines based on the mean *MAP* scores. These showed that GoogleD was better than BPS ($p < 0.001$) and that BPS outperformed both 4sites ($p < 0.0001$) and Google ($p < 0.0001$). However, there was no significant difference in modified average precision between 4sites and Google ($p = 0.363$).

Table 5: Relevance scores for the search engines. * - Note that HFSD was run separately from the five others but results for each query were judged by the assessor who judged that query for the other engines. NCDG means Normalised Cumulative Discounted Gain and is explained in Section 3.4.1.

	average precision	NDCG
GoogleD	0.4074	0.6096
BPS	0.3192	0.5539
4sites	0.2250	0.4545
Google	0.1956	0.3498
HFS	0.0756	0.1888
HFSD*	0.0725	0.1679

GoogleD outperformed BPS, even though the latter is designed to have a high concentration of potentially relevant documents (and few off-topic documents). We considered two possible explanations:

Coverage hypothesis Google indexed a lot more relevant information than BPS. BPS failed to return relevant pages because they were not in its index.

Ranking hypothesis BPS indexed sufficient relevant pages, but failed to return them because its ranking algorithms were ineffective.

We carried out a relative coverage analysis to explore the relative contribution of these hypotheses. We found that, of the 456 relevant pages returned by GoogleD, only 76 were in the BPS collection (16.7%). Conversely, we attempted to locate all the relevant pages returned by BPS by querying the Google engine (keying the relevant URLs obtained from BPS into the Google search box). We found that 338 out of 377 relevant pages from BPS were indexed by Google (89.7%).

This suggests that the lower performance of BPS is mostly due to poor coverage, rather than poor ranking.

We were surprised by HFS's low relevance scores, as the HealthFinder site is a US government service which has previously been rated as a very useful health portal [25]. Once again, the explanation may lie in poor coverage, poor ranking or both. In the case of HFS we were unable to investigate coverage as thoroughly as for BPS since we had no access to the HFS crawl and lacked a practical means of determining by querying whether a particular URL was present. We were, however, able to obtain certain evidence indirectly.

We suspect that low depression coverage is also a problem for HFS as it returned no or few results for some queries. GoogleD returned 1009 results across all the queries (out of a possible 1010), while HFS returned only 485. Google and BPS returned similar numbers to GoogleD, while 849 results were returned for 4sites.

HFS's lack of coverage might be explained by its policy of concentrating on .gov and .org domains. An analysis of the full list of known relevant documents (Table 6) shows that almost 70% of relevant results are in .com and .au.

Table 6: Relevant results by domain, with HFS comparison. The top 7 top-level domains are shown.

	Overall Relevant	HF Relevant	HF Results
.com	402	0	4
.au	251	0	0
.org	104	13	181
.gov	59	25	288
.uk	51	0	0
.net	31	0	4
.edu	25	1	3

The fact that HFSD results are not better than those for HFS suggests that the ranking algorithm employed in HFS is not as sophisticated as suggested by the Search Tips quoted in Section 3.1. We observed that adding the word 'depression' to the query 'X' caused the return of some documents about 'depression' but not about 'X'. These documents sometimes were ranked even more important than those documents resulting when 'X' was the query.

4.2 Quality results

Table 7 shows the number of the results from each engine according to recommendation and treatment rating. The last column is the total quality score for each engine. The 4sites index had the highest overall quality score, followed by BPS and GoogleD. GoogleD falls down by returning 69 pages recommending treatments for which the scientific evidence is presently unsure. For example, unsure treatments 'pleasant activities' and 'lemon balm', were each recommended by 7 out of the 10 GoogleD result pages for those queries.

Table 8 shows correct and incorrect advice results, using the scoring system from Section 3.4.2. Again the 4sites index had the best performance, achieving a ratio of 90% correct. BPS also did well, with 85%. GoogleD didn't do as well because although it returned 51 documents with correct advice, it returned 19 with incorrect advice.

For both measures, 4sites was the most effective engine in retrieving documents with high quality treatment advice on depression. This probably stems from the fact that its four included sites were chosen very carefully, using evidence based criteria. BPS also performed well in returning high quality results, perhaps because it includes many sites and sub-sites dedicated to depression. By contrast, Google might return relevant results from non-depression sites, where the author had insufficient expertise in the area or may have included proportionately more sites that promoted a particular treatment for commercial gain or other reasons. GoogleD returned several results recommending the use of pets, an ineffective treatment, and recommending against St Johns Wort and Paxil, which are rated as effective and very effective respectively.

Note that the gold standard reference [17] we employed for judging the efficacy of treatments is published on the BluePages depression information site

Table 7: Number of documents recommended for different treatment types. Treatments are very effective (VE), effective (E), OK, unsure (U) or not effective (NE). The quality score is calculated according to Table 4.

	Recommend					Recommend against					
	VE	E	OK	U	NE	VE	E	OK	U	NE	**Quality score**
GoogleD	26	19	53	69	8	5	6	1	9	6	**78**
BPS	22	23	30	31	9	4	0	2	15	10	**127**
4sites	21	18	21	14	2	2	1	1	9	6	**143**
Google	17	9	6	15	3	7	3	0	0	1	**28**
HFS	2	1	0	0	1	0	2	0	0	0	**-2**
HFSD	0	3	0	1	1	0	1	0	0	0	**-1**

Table 8: Comparison of recommendations for extreme kinds of treatment. Correct means either recommending for good or against bad treatments. Incorrect means either recommending against good or for bad treatments.

	For		Against					
	Good	Bad	Good	Bad	Correct	Incorrect	Total	Ratio
GoogleD	45	8	11	6	51	19	70	0.73
BPS	45	9	4	10	55	13	68	0.85
4sites	39	2	3	6	45	5	50	0.90
Google	26	3	10	1	27	13	40	0.67
HFS	3	1	2	0	3	3	6	0.50
HFSD	3	1	1	0	3	2	5	0.60

and is therefore included in the indexes of 4sites, BPS and Google but not HealthFinder (all the engines, except for HFS and HFSD, contained the BluePages depression information site in their indexes). Because the BluePages treatment pages make up a higher proportion of the 4sites index, it is obviously easier for 4sites to achieve higher scores. One would expect a '1site' index containing only BluePages to achieve a correctness ratio of 100%.

We attempted to remove potential bias from this cause by repeating the analysis after excluding all pages originating from the BluePages site. Table 9 shows that the quality scores for the two depression-specific search services remain above those of the general engines. HFS and HFSD results are unchanged (because they don't index any documents from the BluePages site) while the quality scores of the other engines all drop. 4sites is harmed most and drops below BPS but both these engines still score higher than GoogleD.

Figure 1 illustrates part of the reason why GoogleD returned lower quality results compared to BPS and 4sites. It returned many more pages which either support or oppose treatments whose effectiveness is ambivalent than did the other engines. It shows the number of ambivalent results out of the total of maximum 290 documents that an engine retrieved in the top 10s list. GoogleD returned more than 130 documents which discussed 'unsure' treatments out of 290 documents while 4sites only contained approximately 40 results in this category.

Figure 2 provides further illustration of the quality differences between search engine result sets in terms of correct and incorrect advice. BPS returned slightly more pages containing correct advice and substantially fewer pages containing incorrect advice than did GoogleD. 4sites returned almost as many pages with correct advice as did GoogleD but many fewer pages containing incorrect advice than either GoogleD or BPS.

Visual inspection of Figure 2 also shows that all engines which returned more than a few advice pages generated substantially more correct than incorrect advice. Among these, BPS and 4sites returned less incorrect advice than Google and GoogleD. The best engine was 4sites which returned the highest proportion of correct advice.

4.3 Meeting searcher needs

We have made no particular assumptions about how much information or what style of information is likely to be valued by people searching for depression information. We expect that there will be considerable variation across searchers.

However, reduced concentration and fatigue are among the diagnostic criteria for depression [3], hence it is important that the highest quality and most useful information is returned at the top of the results list. Searchers are unlikely to be able to judge for themselves the quality of information provided.

4.4 Bias in Query Selection

There were three sources of queries: our list of 50 treatments (TM), BPS logs (BP) and Overture keyword suggestions based on MeSH terms (OM). We compared the relevance effectiveness of all engines for the different query types. GoogleD performed best on all three query types, and particularly well on OM queries. Google, HFS and HFSD performed particularly badly for TM and BP queries, although GoogleD did better. Table 10 shows the average precision and the normalized discounted cumulative gain for each engine when queries from different sources were run. We were interested in whether queries obtained from the BPS query logs would favor depression-specific search engines (i.e. BPS and 4sites). The results showed that all the engines returned more relevant results for OM queries than for BP queries, but the effect was most pronounced for Google and GoogleD.

This may be due to the fact that people submitting queries to BPS know that they can rely on the implied context (depression resources) and submit queries which wouldn't be specific enough in general search, even with the addition of the word depression. Furthermore, the method of selecting the OM queries inevitably produced queries containing the word depression or a synonym, thus establishing an appropriate context.

Table 9: The same data as in Table 4 but excluding pages from the BluePages site. Note that, in this analysis, 4sites is actually 3sites!

	Recommend					Recommend against					
	VE	E	OK	U	NE	VE	E	OK	U	NE	**Quality score**
GoogleD	25	17	49	68	8	5	6	1	1	1	**45**
BPS	20	19	20	30	9	4	0	2	7	4	**74**
4sites	17	7	7	13	2	2	1	1	1	0	**57**
Google	17	9	5	14	3	7	3	0	0	0	**24**
HFS	2	1	0	0	1	0	2	0	0	0	**-2**
HFSD	0	3	0	1	1	0	1	0	0	0	**-1**

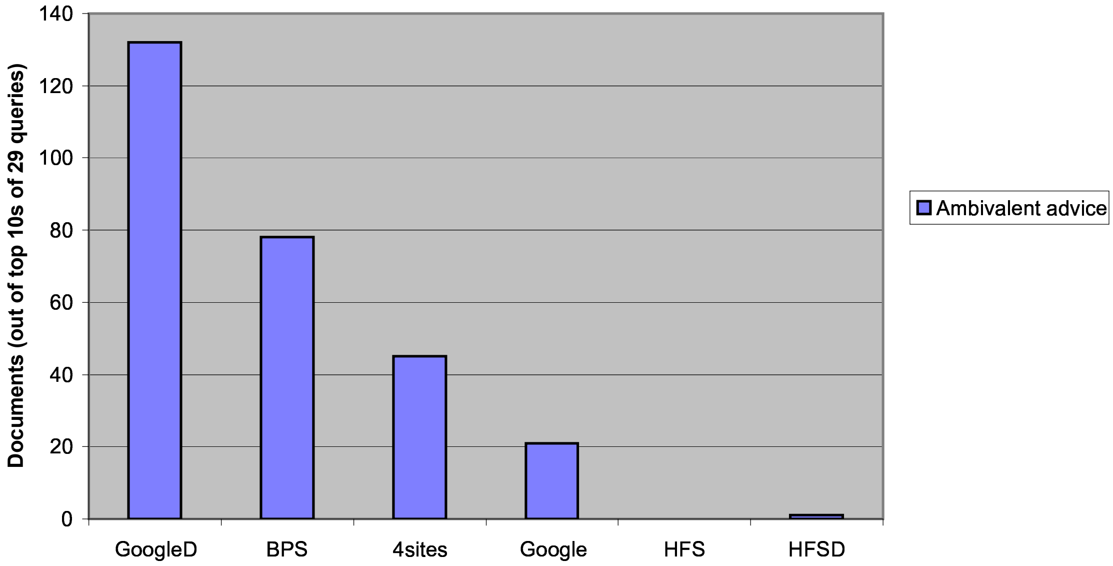

Figure 1: Quality Comparison – Ambivalent Advice is the total number of results that either support or oppose the use of treatments that there is no strong evidence (including results belong to OK and Unsure categories) for their effectiveness

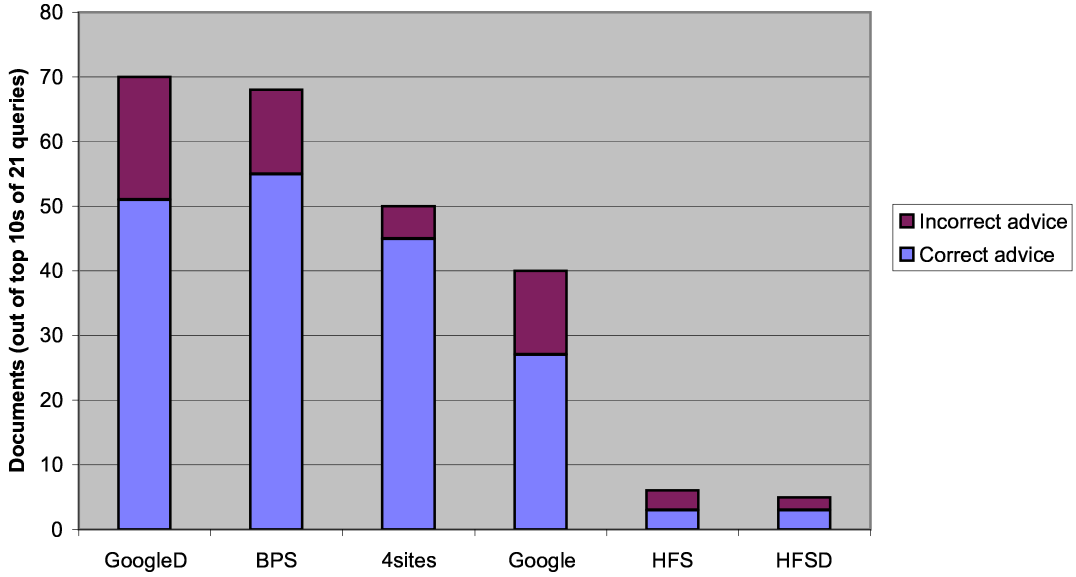

Figure 2: Quality Comparison – Incorrect Advice is total number of results that either support bad treatments (Recommend NE treatments) or oppose the use of good treatments (Recommend against VE and E treatments). Correct Advice is the total number of results that either support good treatments (Recommend VE and E treatments) or oppose the use of bad treatments (Recommend against NE treatments)

Table 10: Relevance scores on queries selected from different sources

	BP Queries		OM Queries		TM Queries		Ratio OM/BP	
	AP	NDCG	AP	NDCG	AP	NDCG	AP	NDCG
GoogleD	0.3036	0.5132	0.4438	0.6557	0.4546	0.6389	1.462	1.278
BPS	0.2557	0.5233	0.2944	0.5723	0.3704	0.5628	1.151	1.094
4sites	0.1692	0.3609	0.1845	0.3736	0.2691	0.5450	1.090	1.035
Google	0.1297	0.3090	0.4155	0.6155	0.1221	0.2357	3.204	1.992
HFS	0.0586	0.1887	0.1468	0.3607	0.0193	0.0798	2.505	1.911
HFSD	0.0477	0.1553	0.1477	0.2775	0.0830	0.0613	3.091	1.787

Google did very badly compared to GoogleD when running with BP queries and treatment queries, because those queries tended not to contain the word depression with consequent lack of specificity. Ten out of 23 BPS topics contained the word 'depression', compared with 18 out of 23 OM topics. Adding 'depression' into queries was quite effective in helping Google to find relevant results.

4.5 Relevance-quality tradeoff

Google was the best search engine for finding relevant results, provided the term 'depression' was added to queries. BPS achieved reasonable scores on both average precision and quality but the best performer in returning high quality results was 4sites.

We hypothesize that there is a trade-off between relevance and quality. In order to ensure that search results are of high quality, it may be necessary to exclude sources of relevant but low quality information.

From the point of view of health outcomes, information quality is far more important than relevance. It may be better to return no search results at all than to return relevant results which supply misleading or unhelpful information. The effect on consumers of provision of low-quality and misleading information has, to our knowledge, not been investigated—Indeed there would be serious ethical problems in setting up such an experiment!

In other non-health domains, the optimal trade-offs may be different, but investigation of this remains for future work.

Our findings somewhat agree with [5] in that medical specific search tools were less effective in finding relevant information than general search tools. However, our results suggest that better quality results can be obtained from medical search engines if the indexing is done properly. This quality aspect was not addressed in [5].

5 Conclusions and future work

High quality search in the area of mental health information is important because of the demonstrated positive effect of web-delivered information on mental health status. We have compared three general approaches to providing high quality search in the depression domain: general search engines exemplified by Google, health-specific engines represented by HFS and depression-specific engines exemplified by BPS.

Weeks of human effort was required to setup BPS and considerable ongoing effort would be needed to maintain its coverage and accuracy. It was built by manually selecting relevant web sites by filtering long lists of search results from major search engines, crawling them and indexing.

Our findings suggest that the effort of setting up and maintaining a portal search engine can best be justified in terms of focusing search context and filtering out low quality information. Searchers on BluePages would find more relevant documents if their queries were forwarded to Google with the addition of the word 'depression' but result quality would not be as high.

Anecdotally, another argument for portal search is that it can suppress harmful information such as "how to commit suicide" pages in response to queries from severely depressed consumers.

We found no support in terms of coverage for a depression-specific search service. In other domains, portal search might be justified on those grounds if it were able to index important content not indexable by general search engines.

Although the domain-specific engines 4sites and BPS returned pages with better depression treatment advice than the general-purpose engine Google, they retrieved fewer relevant results. This suggests that there may be a tradeoff between large scale coverage, which can increase the number of relevant documents returned, and selectivity, which can improve quality. Which particular tradeoff is optimal will depend upon the purpose of the search service and the needs of its intended users.

There is obvious follow-up work to be done on more effective and/or less labour intensive methods of creating domain-specific search engines. Different approaches could be evaluated using the methods described here, comparing their ability to find sites with potentially relevant and high-quality information. Could focused crawling methods be effective in the depression domain? If so, how should the initial seed list be created? Is it possible to automatically estimate the quality of web pages in health domains? If so, a focused crawler would be able to ignore low-value content.

In the future, we would like to extend our work to different domains. The present study only considers depression, which has a well-defined, evidence-based notion of quality. Its findings are likely to generalise to other health domains. In other domains, desirable attributes such as correctness, comprehensiveness or up-to-dateness might be harder to objectively measure.

The observed difference between Google and GoogleD results shows that restricting the domain of documents is necessary to achieve good results on a whole-of-Web search service. In the domain of depression and with conjunctive query semantics it was possible to improve search results substantially by adding a single query word[11]. For other subject domains such as 'chemistry' or 'trade unions' there may be no simple way of ensuring queries are sufficiently specific for effective use in general engines.

Hypothetically, a more sophisticated domain restriction could be applied by classifying all the pages in the Google index and restricting the search to pages with the relevant domain label, but this is likely to be very expensive and reliant upon the accuracy of the classifier.

Acknowledgments

We gratefully acknowledge the assistance of Francis Crimmins, Anthony Bennett, Simon Blomberg, Chloe Groves, Kelly Blewitt and that of our relevance judges Tania M. Colwell and Stephanie Liau.

References

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

[2] G Andrews, C Issakidis, and G Carter. Shortfall in mental health service utilisation. *British Journal of Psychiatry*, 179:417–25, November 2001.

[3] American Psychiatric Association. *Diagnostic and Statistical Manual of Mental Disorders.* Author, Washington, DC, 4th edition, 1994.

[4] L Baker, TH Wagner, S Singer, and KM Bundorf. Use of the internet and e-mail for health care information. *Journal of the American Medical Association*, 289(18):2400–2406, 2003.

[5] L Bin and KC Lun. The retrieval effectiveness of medical information on the web. *International Journal of Medical Informatics*, 62:155–163, 2001.

[6] S Chakrabarti, M Berg, and B Dom. Focused crawling: a new approach to topic-specific web resource discovery. In *Proceeding of the 8th International World Wide Web Conference (WWW8)*, 1999.

[7] J Cho, H Garcia-Molina, and L Page. Efficient crawling through url ordering. In *Proceeding of the Seventh World Wide Web Conference*, 1998.

[8] H Christensen, KM Griffiths, and AF Jorm. Delivering interventions for depression by using the internet: randomised controlled trial. *BMJ*, 328(7434):265–0, 2004.

[9] M Clarke and AD Oxman. *Cochrane Reviewers' handbook 4.1.1.* The Cochrane Library, Oxford, 2001.

[10] G Eysebach and C Kohler. How do consumers search for and appraise health information on the world wide web? qualitative study using focus groups, usability tests, and in-depth interviews. *BMJ*, 324:573–577, 2002. http:://bmj.com.

[11] BK Gretchen, MN Elliot, LS Morales, JI Algazy, RL Kravitz, MS Broder, DE Kanouse,

[11] Under the different query semantics implemented by HFS, adding the word depression led to worse results.

JA Munoz, JA Puyol, M Lara, KE Watkins, H Yang, and EA McGlynn. Health Information on the Internet: Accessibility, Quality, and Readability in English and Spanish. *JAMA*, 285(20):2612–2621, 2001.

[12] KM Griffiths and H Christensen. The quality and accessibility of australian depression sites on the world wide web. *MJA*, 176:S97–S104, 2002.

[13] D Hawking, N Craswell, P Bailey, and K Griffiths. Measuring search engine quality. *Information Retrieval*, 4(1):33–59, 2001. `http://es.cmis.csiro.au/pubs/hawking_ir01.pdf`.

[14] M Hersovicia, M Jacovia, YS Maareka, D Pellegb, M Shtalhaima, and S Ura. The shark-search algorithm. an application: tailored web site mapping. In *Proceeding of the Seventh World Wide Web Conference*, 1998.

[15] D Ilic, TL Bessell, CA 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.

[16] K Jarvelin and J Kekalainen. Cumulated gain-based evaluation of ir techniques. *ACM Trans. Inf. Syst.*, 20(4):422–446, 2002.

[17] AF Jorm, H Christensen, KM Griffiths, A Korten, and B Rodgers. *Help for depression: What works (and what doesn't).* Centre for Mental Health Research, Canberra, Australia, 2001.

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

[19] F Menczer, G Pant, and P Srinivasan. Evaluating topic-driven web crawlers. In *Proceeding of the 24th Annual Intl. ACM SIGIR Conf. On Research and Development in Information Retrieval*, 2001.

[20] S Mukherjea. Organizing topic-specific web information. In *Proceedings of the eleventh ACM on Hypertext and hypermedia*, pages 133–141. ACM Press, 2000.

[21] CJL Murray and AD Lopez, editors. *The global burden of disease and injury series.* Harvard University Press, Cambridge MA, 1996.

[22] AJ Sellen, R Murphy, and KL Shaw. How knowledge workers use the web. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 227–234. ACM Press, 2002.

[23] 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.

[24] A Spink, BJ Jansen, D Wolfram, and T Saracevic. From e-sex to e-commerce: Web search changes. *Computer*, 35(3):107–109, 2002.

[25] G Wu and J Li. Comparing web search engine performance in searching consumer health information: evaluation and recommendations. *Bull Med Libr Assoc*, 87(4):456–461, 1999.