Focused crawling in depression portal search: A feasibility study

Thanh Tin Tang
Department of Computer Science, ANU
ACT 0200, Australia
thanh.tang@cs.anu.edu.au

David Hawking
CSIRO ICT Centre
ACT 2601, Australia
david.hawking@csiro.au

Nick Craswell
Microsoft Research
CB3 0FB, UK
nickcr@microsoft.com

Ramesh S. Sankaranarayana
Department of Computer Science, ANU
ACT 0200, Australia
ramesh@cs.anu.edu.au

Abstract *Previous work on domain specific search services in the area of depressive illness has documented the significant human cost required to setup and maintain closed-crawl parameters. It also showed that domain coverage is much less than that of whole-of-web search engines. Here we report on the feasibility of techniques for achieving greater coverage at lower cost. We found that acceptably effective crawl parameters could be automatically derived from a DMOZ depression category list, with dramatic saving in effort. We also found evidence that focused crawling could be effective in this domain: relevant documents from diverse sources are extensively interlinked; many outgoing links from a constrained crawl based on DMOZ lead to additional relevant content; and we were able to achieve reasonable precision (88%) and recall (68%) using a J48-derived predictive classifier operating only on URL words, anchor text and text content adjacent to referring links. Future directions include implementing and evaluating a focused crawler. Furthermore, the quality of information in returned pages (measured in accordance with the evidence based medicine) is vital when searchers are consumers. Accordingly, automatic estimation of web site quality and its possible incorporation in a focused crawler is the subject of a separate concurrent study.*

Keywords focused crawler, hypertext classification, mental health, depression, domain-specific search.

1 Introduction

Depression is a major public health problem, being a leading cause of disease burden [13] and the leading risk factor for suicide. 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 [6]. Thus, the Web is a potentially valuable 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 [8, 10]. It is thus important that consumers can locate depression information which is both relevant and of high quality.

Recently, in [15], we compared examples of two types of search tool which can be used for locating depression information: whole-of-Web search engines such as Google, and domain-specific (portal) search services which include only selected sites. We found that coverage of depression information was much greater in Google than in portals devoted to depression or health.

BluePages Search (BPS)[1] is a depression-specific search service offered as part of the BluePages depression information site. Its index was built by manually identifying and crawling areas on 207 Web servers containing depression information. It took about two weeks of intensive human effort to identify these areas (seed URLs) and define their extent by means of include and exclude patterns. Similar effort would be required at regular intervals to maintain coverage and accuracy. Despite this human effort, only about 17% of relevant pages returned by Google were contained in the BPS crawl.

One might conclude from this that the best way to provide depression-portal search would be to add the word 'depression' to all queries and forward them to a general search engine such as Google. However, in other experiments in [15] relating to quality of information in search results, we showed that substantial amounts of the additional relevant information returned by Google was of low quality and not in accord with best available scientific evidence. The operators of the BluePages portal (ANU's Centre for Mental Health Research) were keen to know if it would be feasible to provide a portal search service featuring:

1. increased coverage of high-quality depression information,

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

Proceedings of the 9th Australasian Document Computing Symposium, Melbourne, Australia, December 13, 2004.

2. reduced coverage of dubious, misleading or unhelpful information, and

3. significantly reduced human cost to maintain the service.

We have attempted to answer the questions in two parts. Here we attempt to determine whether it is feasible to reduce human effort by using a directory of depression sites maintained by others as a seedlist and using focused crawling techniques to avoid the need to define include and exclude rules. We also investigate whether the content of a constrained crawl links to significant amounts of additional depression content and whether it is possible to tell which links lead to depression content.

A separate project is under way to determine whether it is feasible to evaluate the quality of depression sites using automatic means. It will be reported elsewhere. If the outcomes of both projects are favourable, the end-result may be a focused crawler capable of preferentially crawling relevant content *from high quality sites*.

2 Focused crawling - related work

Focused crawlers, first described by de Bra et al. [2], for crawling a topic-focused set of Web pages, have been frequently studied [3, 1, 5, 9, 12].

A focused crawler seeks, acquires, indexes, and maintains pages on a specific set of topics that represent a relatively small portion of the Web. Focused crawlers require much smaller investment in hardware and network resources but may achieve high coverage at a rapid rate.

A focused crawler starts with a seed list which contains URLs that are relevant to the topic of interest, it crawls these URLs and then follows the links from these pages to identify the most promising links based on both the content of the source pages and the link structure of the web [3]. Several studies have used simple string matching of these features to decide if the next link is worth following [1, 5, 9]. Others used reinforcement learning to build domain-specific search engines from similar features. For example, McCallum et al. [11] used Naive Bayes classifiers to classify hyperlinks based on both the full text of the sources and anchor text on the links pointing to the targets.

A focused crawler should be able to decide if a page is worth visiting before actually visiting it. This raises the general problem of hypertext classification.

In traditional text classification, the classifier looks only at the text in each document when deciding what class should be assigned.

Hypertext classification is different because it tries to classify documents without the need for the content of the document itself. Instead, it uses link information. Chakrabati et al. [3] used the hypertext graph including in-neighbours (documents citing the target document) and out-neighbours (documents that target document cites) as input to some classifiers.

Our work also used link information. We tried to predict the relevance of uncrawled URLs using three features: anchor text, text around the link and URL words.

3 Resources

This section describes the resources used in our experiments: the BluePages search service; the data from our previous domain-specific search experiments; the DMOZ depression directory listing and the WEKA machine learning toolkit.

3.1 BluePages Search

BluePages Search (BPS) is a search service offered as part of the existing BluePages depression information site. Crawling, indexing and search were performed by CSIRO's Panoptic search engine[2].

The list of web sites that made up the BPS was manually identified from the Yahoo! Directory and from querying general search engines using the query term 'depression'. Each URL from this list was then examined to find out if it was relevant to depression before it was selected. The fencing of web site boundaries was a much bigger issue. A lot of human effort was needed to examine all the links in each web site to decide which links should be included and excluded. Areas of 207 web sites were selected. These areas sometimes included a whole web server, sometimes a subtree of a web server and sometimes only some individual pages. Newspaper articles (which tend to be archived after a short time), potentially distressing, offensive or destructive materials and dead links were excluded during the construction of the BPS index.

A simple example of seeds and boundaries is:

- $seed$ = `www.counselingdepression.com/`, and
- $include\_patterns$ = `www.counselingdepression.com`.

In this case, every link within this web site is included. In complicated cases, however, some areas should be included while others are excluded. For instance, examining `www.drada.org` would result in the following seed and boundaries:

- $seed$ = `www.drada.org/`
- $include\_patterns$ = `www.drada.org`
- $exclude\_patterns$ = `www.drada.org/facts/bipolar.html`, `www.drada.org/facts/bipolar_nih.html`, `www.drada.org/Store/bookreviews_`.

The above boundaries mean that everything within the web site should be crawled except for pages about bipolar depression and book reviews.

[2] `http://www.panopticsearch.com/`

3.2 Data from our previous work

In our previous work, we conducted a standard information retrieval experiment, running 101 'depression' queries against six engines of different types: two health portals, two depression-specific search engines, one general search engine and one general search engine where the word 'depression' was added to each query if not already present (GoogleD). We then pooled the results for each query and employed research assistants to judge them. We obtained 2778 judged URLs and 1575 relevant URLs from all the engines. We used these URLs as a base in the present work to estimate relevance.

We found that, over 101 queries, GoogleD returned more relevant results than those of the domain-specific engines. 621 relevant URLs were returned by BPS while 683 relevant results were retrieved by GoogleD. As GoogleD was the best performer in obtaining the most relevant results, we also used it as a base engine to compare with other collections in the present work.

3.3 DMOZ

DMOZ[3] is the Open Directory Project which is "the largest, most comprehensive human-edited directory of the Web. It is constructed and maintained by a vast, global community of volunteer editors"[4].

We started with the Depression directory[5] which contains documents and directories purportedly relevant to depressive disorder.

3.4 Weka

Weka[6] was developed at the University of Waikato in New Zealand [16]. It is a data mining package which contains machine learning algorithms. Weka provides tools for data pre-processing, classification, regression, clustering, association rules, and visualization. Weka was used in our experiments for the prediction of URL relevance using hypertext features. It was used because it provided many classifiers, was easy to use and served our purposes well in predicting URL relevance.

4 Experiment 1 - Usefulness of a DMOZ category as a seed list

A focused crawler needs a good seed list of relevant URLs as a starting point for the crawl. These URLs should span a variety of web site types so that the crawler can explore the Web in many different directions. Instead of using a manually created list, we attempted to derive a seed list from a publicly available directory - DMOZ. Because depression sites on the web are widely scattered, the diversity of content in DMOZ is expected to improve coverage. Using DMOZ allows us to leverage off the categorisation work being done by volunteer editors.

4.1 DMOZ seed generation

We started from the 'depression' directory on the DMOZ web site, namely `http://www.dmoz.org/Health/Mental_Health/Disorders/Mood/Depression/`. This directory is intended to contain links to relevant sites and subsites about depression. The directory, however, also had a small list of 12 within-site links to other directories, which may or may not be relevant to depression. We, therefore, only needed to do some minor boundary selection for these links to include relevant directories. For example, the following directories were included because they are related to depression and they are links from the depression directory: `dmoz.org/Health/Mental_Health/Disorders/Child_and_Adolescent/Childhood_Depression/`, and `dmoz.org/Health/Pharmacy/Drugs_and_Medications/Antidepressants/`. These links were selected simply because their URLs contain the term 'depression' (such as `childhood_depression`) or 'antidepressants'. The seed URLs, as a result, included the above links and all the links to depression-related sites and subsites from this directory.

Include patterns corresponding to the seed URLs were generated automatically. In general, the include pattern was the same as the URL, except that default page suffixes such as `index.htm` were removed. Thus, if the URL referenced the default page of a server or web directory, the whole server or whole directory was included. If the link was to an individual page, only that page was included.

The manual effort required to identify the seed URLs and define their extent varied greatly between BPS and DMOZ. While it took about two weeks of intensive effort in the BPS case, it only required about one hour's work for DMOZ.

4.2 Comparison of the DMOZ collection and the BPS collection

This experiment aimed to find out if a constrained crawl from the low-cost DMOZ seed list can lead to domain coverage comparable to that of the manually configured BPS.

After identifying the DMOZ seed list and include patterns as described above, we used the Panoptic crawler to build our DMOZ collection. We then ran the 101 queries from our previous study and obtained 779 results for DMOZ.

We attempted to judge the relevance of these results using the 1575 known relevant URLs (see Section 3.2) and to compare the DMOZ results with those of the BPS collection.

Table 1 shows that 186 out of 227 judged URLs (a pleasing 81%) from the DMOZ collection were relevant. However, the percentage of judged results (30%)

[3] `http://www.dmoz.org`

[4] `http://www.dmoz.org/about.html`

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

[6] `http://www.cs.waikato.ac.nz/~ml/weka/`

Table 1: Comparison of relevant URLs in DMOZ and BPS results of running 101 queries.

	URLs	judged URLs	relevant URLs
BPS	683	683	621
DMOZ	779	227	186

was too low to allow us to validly conclude that DMOZ was a good collection.

Since we no longer had access to the services of the judges from the original study we attempted to confirm that a reasonable proportion of the unjudged documents were relevant to the general topic of depression by sampling URLs and judging them ourselves.

We randomly selected 2 lists of 50 non-overlapped URLs among the unjudged results and made relevance judgments on these. In the first list, we obtained 35 relevant results and in the second list, 34 URLs were relevant. Because there was close agreement between the proportion relevant in each list we were confident that we could extrapolate the results to give a reasonable estimate of the total number of relevant pages returned.

Extrapolation suggests 381 relevant URLs for the unjudged DMOZ set. Hence, in total we might be able to obtain 567 (186 + 381) relevant URLs from the DMOZ set. This number was not as high as that of BPS, but it was relatively high (72% relevant URLs in DMOZ set compared to 91% of these in BPS). Therefore, we could conclude that the DMOZ list is an acceptably good, low-maintenance starting point for a focused crawl.

5 Experiments 2A-2C - Additional link-accessible relevant information

Although some focused crawlers can look a few links ahead to predict relevant links at some distance from the currently crawled URLs [7], the immediate outgoing links are of most immediate interest.

We performed three experiments to gauge how much additional relevant information is accessible one link away from the existing crawled content. If no additional relevant content is linked to from pages in the original crawl, the prospects of successful focused crawling are very low. Figure 1 shows an illustration of the one-link-away set of URLs from the DMOZ crawl.

The first experiment (2A) involved testing if outgoing links from the BPS collection were relevant while the second (2B) compared the outgoing link sets of BPS and DMOZ to see if DMOZ was really a good place to lead a focused crawler to additional relevant content. The last experiment (2C) attempted to find out if URLs relevant to a particular topic linked to each other.

5.1 Experiment 2A: Outgoing links from the BPS collection

The data used for this experiment included:

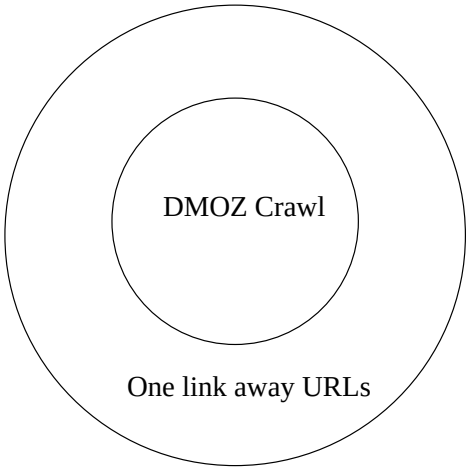

Figure 1: Illustration of one link away collection from the DMOZ crawl.

- the BPS index,
- the BPS outgoing link set containing all URLs linked to by BPS URLs, and
- 2 sets of judged-relevant URLs: BPS relevant and all relevant.

Our previous work concluded that BPS didn't retrieve as many relevant documents as GoogleD because of its small coverage of sites. We wanted to find out if focused crawling techniques have the potential to raise BPS performance by crawling one step away from BPS. Among 954 relevant pages retrieved by all engines except for BPS, BPS failed to index 775 pages. The extended crawl yielded 196 of these 775 pages or 25.3%.

In other words, an unrestricted crawler starting from the original BPS crawl would be able to reach an additional 25.3% of the known relevant pages, in only a single step from the existing pages. In fact, the true number of additional relevant pages is likely to be higher because of the large number of unjudged pages.

It is unclear whether the additional relevant content in the extended BPS crawl would enable more relevant documents to be retrieved than in the case of GoogleD. Retrieval performance depends upon the effectiveness of the ranking algorithm as well as on coverage.

5.2 Experiment 2B: Comparison of outgoing links between BPS and DMOZ

This experiment compared the out-going link sets of BPS and DMOZ to find out if the DMOZ seed list could be used instead of the BPS seed list to guide a focused crawler to relevant areas of the web. The following data were used:

- 2 sets of out-going links from the BPS and DMOZ collections, and
- 2 sets of all judged URLs and judged-relevant URLs.

Table 2: Comparison of relevant out-going link URLs for BPS and DMOZ.

Collection	size	judged	relevant
outgoing BPS	49,370	248	196
outgoing DMOZ	122,985	203	158

From our previous work, we obtained 2778 judged URLs which were used here as a base to compare relevance. Table 2 shows that even though the outgoing link collection of DMOZ was more than double the size of that of BPS, more outgoing BPS pages were judged. Among the judged pages, BPS and DMOZ had 196 and 158 relevant pages respectively in their outgoing link sets. Although DMOZ had less known relevant pages than BPS, the proportion of relevant pages versus judged pages were quite similar for both engines(78% for DMOZ and 79% for BPS). This result together with the size of each outgoing link collection implied that (1) The DMOZ outgoing link set contained quite a large number of relevant URLs which could potentially be accessed by a focused crawler, and (2) The DMOZ seed list could lead to much better coverage than the BPS seed list.

5.3 Experiment 2C: Linking patterns between relevant pages

We performed a very similar experiment to the experiment described in Section 5.1, with the purpose of finding out if relevant URLs on the same topic are linked to each other. Instead of using the whole BPS collection of 12,177 documents as the seed list, we only chose the 621 known relevant URLs. The following data were used:

- the BPS known relevant URLs,
- the BPS outgoing link set from the above, containing all URLs linked to by BPS known relevant URLs, and
- judged-relevant URLs from our previous work.

The outgoing link collection of the BPS known relevant URLs contained 5623 URLs. Of these, 158 were known relevant. This was a very high number compared to the 196 known relevant URLs obtained from the much bigger set of all outgoing link URLs (containing above 40,000 URLs) in the previous experiment. It is likely from this experiment that relevant pages tend to link to each other. This is good evidence supporting the feasibity of the focused crawling approach.

6 Experiment 3 - Hypertext classification

After downloading the content of the seed URLs and extracting links from them, a focused crawler needs to decide what links to follow and in what order based on the information it has available. We used hypertext classification for this purpose.

6.1 Collection of URLs for training and testing

For both BPS and DMOZ crawls, we collected all immediate outgoing URLs satisfying the following two conditions (1) known relevant or known irrelevant URLs and (2) the URLs pointing to each of these URLs were also relevant. We collected 295 relevant and 251 irrelevant URLs for our classification experiment.

6.2 Features

Several papers in the field used the content of crawled URLs, anchor text, URL structure and other link graph information to predict the relevance of the next unvisited URLs [1, 5, 9]. Instead of looking at the content of the whole document pointing to the target URL, Chakrabarti [4] used 50 characters before and after a link and suggested that this method was more effective. Our work was somewhat related to all of the above. We used the following features to predict the relevance of the target URL.

- anchor text on the source pages: all the text appearing on the links to the target page from the source pages,
- text around the link: 50 characters before and 50 characters after the link to the target page from the source pages[7], and
- URL words: words appearing in the URL of the target page.

We accumulated all words for each of these features to form 3 vocabularies where all stop words were eliminated. URL words separated by a comma, a full stop, a special character and a slash were parsed and treated as individual words. URL extensions such as `.html`, `.asp`, `.htm`, `.php` were also eliminated. The end result showed 1,774 distinct words in the anchor text vocabulary, 874 distinct words in the URL vocabulary, and 1103 distinct words in the content vocabulary.

For purposes of illustration , Table 3 shows the features extracted from each of six links to the same URL.

Assume that we would like to predict `www.ndmda.org` for its relevance to depression and that we have six already-crawled pages pointing to it from our crawled collection. From each of the pages, features are extracted in the form of anchor text words and the words within a range of a maximum of 50 characters before and after the link pointing to `www.ndmda.org`. There is no content around the link from `www.noondaydemon.com/patresources.html` to the target URL because that URL contains only stop words and/or numbers which have been stripped off. The URL words for the target URL after being parsed contains: `ndmda,org`.

[7] We first extracted the 50-character string and then eliminated markup and stopwords, sometimes leaving only a few words.

Table 3: Features for www.ndmda.org after removing stop words and numbers.

Target URL: www.ndmda.org URL words: ndmda, org		
source URL	**anchor text**	**content around the link**
www.paxil.com/depression/dp_sym.html	ndmda, org	depression, bipolar, support, alliance,american, psychiatric
www.healthyplace.com/communities/ depression/living/my_experience5.asp	depression, bipolar, support, aliance	support, group, affliated, highly, recommend
www.healthyplace.com/Communities/ Depression/nimh/suicide_5.asp	national, depressive, manic, depressive, association	
www.noondaydemon.com/ pat_resources.html	national, depressive, manic, depressive, association	ncwa, ndmda
www.paxil.com/depression/dp_ln.html	ndmda, org	depression, bipolar, support, alliance,american, psychiatric
www.emufarm.org/ cmbell/ depress/deplink.html	national, depressive, manic, depressive, association	organisations

Table 4: Algorithms used from Weka.

Classifier	**Description**
IBK	k-nearest neighbors.
ZeroR	Zero rule. Predicts the majority class. Used as a baseline.
NaiveBayes	Statistical method. Assumes independence of attributes. Uses conditional probability and Bayes rule.
Complement Naive Bayes	Class for building and using a Complement class Naive Bayes classifier.
J48	C4.5 algorithm. A decision tree learner with pruning.
Bagging	Class for bagging a classifier to reduce variance.
AdaBoostM1	Class for boosting a nominal class classifier using the Adaboost M1 method.

6.3 Classifiers

We compared a range of classification algorithms provided by Weka [16]. (See Table 4.)

When training and testing the collection, we used a stratified cross-validation method, i.e. using 10-fold cross validation where one tenth of the collection was used for training and the rest was used for testing and the operation was repeated 10 times. The results were then averaged and a confusion matrix was drawn to find accuracy, precision and recall.

6.4 Input data

We treated the three vocabularies containing all features independently from each other. We computed term frequency and inverse document frequency ($tf.idf$) for each feature attached to each of the URLs specified in Section 6.1 using the following formula [14].

$$tf.idf = tf(t,d) * \log(n/df(t))$$

where t is a term, d is a document, $tf(t,d)$ is the frequency of t in d, n is total number of documents and $df(t)$ is the number of documents containing t.

By this means we obtained a list of URLs, each associated with the $tf.idf$s for all terms in the 3 vocabularies. A learning algorithm was then run in Weka to learn and predict if these URLs were relevant or irrelevant. We also used boosting and bagging algorithms to boost the performance of different classifiers.

6.5 Measures

We used three measures to analyse how a classifier performed in categorizing all the URLs. We denoted true positive and true negative for the relevant and irrelevant URLs that were correctly predicted by the classifier respectively. Similarly, false positive and false negative were used for irrelevant and relevant URLs that were incorrectly predicted respectively. The three measures are listed below.

- Accuracy: shows how accurately URLs are classified into correct categories.
  $accuracy = \frac{true\ positive + true\ negative}{all\ URLs}$
- Precision: shows the proportion of correctly relevant URLs out of all the URLs that were predicted as relevant.
  $precision = \frac{true\ positive}{true\ positive + false\ positive}$

- Recall: shows the proportion of relevant URLs that were correctly predicted out of all the relevant URLs in the collection.
  $recall = \frac{true\ positive}{true\ positive + false\ negative}$

Although accuracy is an important measure, a focused crawler would be more interested in following the links from the predicted relevant set to crawl other potentially relevant pages. Thus, precision and recall are better measures.

6.6 Results and discussion

The results of some representative classifiers are shown in Table 5. ZeroR represented a realistic performance "floor" as it classified all URLs into the largest category i.e relevant. As expected, it was the least accurate. Naive Bayes and J48 performed best. Naive Bayes was slightly better than J48 on recall but the latter was much better in obtaining higher accuracy and precision. Out of 228 URLs that J48 predicted as relevant, 201 were correct (88.15%). However, out of the 264 URLs predicted as relevant by Naive Bayes, only 206 (78.03%) were correct. Overall, the J48 algorithm was the best performer among all the classifers used.

We found that bagging did not improve the classification result while boosting showed some improvement for recall (from 64.74% to 68.13%) when the J48 algorithm was used.

We also performed other experiments where only one set of features or any combination of two sets of features were used. In all cases, we observed that the accuracy, precision and recall were all worse than when all three sets of features were combined.

Our best results, as detailed in Table 5, showed that a focused crawler starting from a set of relevant URLs, and using J48 in predicting future URLs, could obtain a precision of 88% and a recall of 68% using the features mentioned in Section 6.2.

We wished to compare these performance levels with the state of the art, but were unable to find in the literature any applicable results relating to the topic of depression. We therefore decided to compare our predictive classifier with a more conventional content classifier for the same topic.

We built a 'content classifier' for 'depression', using only the content of the target documents instead of the features being used in our experiment. The best accuracies obtained from the two classification systems were very similar, 78% for the content classifier and 77.8% for the predictive version. Content classification showed slightly worse precision but better recall.

We concluded from this comparison that hypertext classification is quite effective in predicting the relevance of uncrawled URLs. This is quite pleasing as a lot of unnecessary crawling can be avoided.

Finally we explored two variant methods for feature selection. We found that generating features using stemmed words caused a reduction in performance, as did reducing the feature set using a feature selection method.

7 Conclusions and future work

Weeks of human effort were required to set up the current BPS depression portal search service and considerable ongoing effort is needed to maintain its coverage and accuracy. Our investigations of the viability of a focused crawling alternative have resulted in three key findings.

First, web pages on the topic of depression are strongly interlinked despite the heterogeneity of the sources. This confirms previous findings in the literature for other topic domains and provides a good foundation for focused crawling in the depression domain. The one-link away extensions to the closed BPS and DMOZ crawls contained many relevant pages.

Second, although somewhat inferior to the expensively constructed BPS alternative, the DMOZ depression category features a diversity of sources and seems to provide a seed list of adequate quality for a focused crawl in the depression domain. This is very good news for the maintainability of the portal search because of the very considerable labour savings. Other DMOZ categories may provide good starting points for other domain-specific search services.

Third, predictive classification of outgoing links into relevant and irrelevant categories using source-page features such as anchor text, content around the link and URL words of the target pages, achieved very promising results. With the J48 decision-tree algorithm, as implemented by Weka, we obtained high accuracy, high precision and relatively high recall.

Given the promise of the approach, there is obvious follow-up work to be done on designing and building a domain-specific search portal using focused crawling techniques. In particular, it may be beneficial to rank the URLs classified as relevant in the order of degree of relevance so that a focused crawler can decide on visiting priorities. Also, appropriate data structures are needed to hold accumulated information for unvisited URLs (i.e. anchor text and nearby content for each referring link.) This information needs to be updated as additional links to the same target are encountered.

Another important question will be how to persuade Weka to output a classifier that can be easily plugged-in into the focused crawler's architecture. Since the best performing classifier in these trials was a decision tree, this may be easier than otherwise.

Once a focused crawler is constructed, it will be necessary to determine how to use it operationally. We envisage operating without any include or exclude rules but will need to decide on appropriate stopping conditions. If none of the outgoing links are classified as likely to lead to relevant content, should the crawl stop, or should some unpromising links be followed? And with what restrictions?

Table 5: Classification Results.

Classifier	Accuracy (%)	Precision (%)	Recall(%)
IBk	54.76	80	21.69
ZeroR	54.02	54.02	100
Complement Naive Bayes	71.06	77.51	65.42
Naive Bayes	73.07	78.03	69.83
J48	77.83	88.15	68.13

Because of the requirements of the depression portal operators site quality must be taken into account in building the portal search service. Ideally, the focused crawler should take site quality into account when deciding whether to follow an outgoing link, but this may or may not be feasible. Another more expensive alternative would be to crawl using relevance as the sole criterion and to filter the results based on quality.

Site quality estimation is the subject of a separate study, yet to be completed. In the meantime, it seems fairly clear from our experiments that it will be possible to increase coverage of the depression domain for dramatically lower cost by starting from a DMOZ category list and using a focused crawler.

Verifying whether techniques found useful in this project also extend to other domains is an obvious future step. Other health-related areas are the most likely candidates because of the focus on quality of information in those areas.

Acknowledgments

We gratefully acknowledge the contribution of Kathy Griffiths and Helen Christensen in providing expert input about the depression domain and about BluePages, and of John Lloyd and Eric McCreath for their advice on machine learning techniques.

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.*, Volume 19, Number 3, pages 286–309, 2001.

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

[3] 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.

[4] S. Chakrabarti, B. Dom, P. Raghavan, S. Rajagopalan, D. Gibson and J. Kleinberg. Automatic resource compilation by analyzing hyperlink structure and associated text. In *Proceedings of the seventh international conference on World Wide Web 7*, pages 65–74. Elsevier Science Publishers B. V., 1998.

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

[6] H. Christensen, K. M. Griffiths and A. F. Jorm. Delivering Interventions for Depression by Using the Internet: Randomised Controlled Trial. *British Medical Journal*, Volume 328, Number 7434, pages 265–0, 2004.

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

[8] Berland G, Elliott M, Morales L, Algazy J, Kravitz R, Broder M, Kanouse D, Munoz J, Puyol J, Lara M, Watkins K, Yang H and McGlynn E. Health Information on the Internet: Accessibility, Quality, and Readability in English and Spanish. *The Journal of the American Medical Association*, Volume 285, Number 20, pages 2612–2621, 2001.

[9] 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 *Proceeding of the Seventh World Wide Web Conference*, 1998.

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

[11] 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.

[12] 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.

[13] C. J. L. Murray and A. D. Lopez (editors). *The Global Burden of Disease and Injury Series*. Harvard University Press, Cambridge MA, 1996.

[14] G. Salton and C. Buckley. Term weighting approaches in automatic text retrieval. Technical report, 1987.

[15] 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*, 2004.

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