2008

Quality information retrieval for the World Wide Web

Milly W. Kc
University of Wollongong, millykc@uow.edu.au

Markus Hagenbuchner
University of Wollongong, markus@uow.edu.au

Ah Chung Tsoi
Hong Kong Baptist University, act@uow.edu.au

Quality information retrieval for the World Wide Web

Abstract

The World Wide Web is an unregulated communication medium which exhibits very limited means of quality control. Quality assurance has become a key issue for many information retrieval services on the Internet, e.g. web search engines. This paper introduces some quality evaluation and assessment methods to assess the quality of web pages. The proposed quality evaluation mechanisms are based on a set of quality criteria which were extracted from a targeted user survey. A weighted algorithmic interpretation of the most significant user quoted quality criteria is proposed. In addition, the paper utilizes machine learning methods to produce a prediction of quality for web pages before they are downloaded. The set of quality criteria allows us to implement a web search engine with quality ranking schemes, leading to web crawlers which can crawl directly quality web pages. The proposed approaches produce some very promising results on a sizable web repository.

Disciplines

Physical Sciences and Mathematics

Publication Details

Kc, M. W., Hagenbuchner, M. & Tsoi, A. (2008). Quality information retrieval for the World Wide Web. IEEE/WIC/ACM international Conference on Web Intelligence and Intelligent Agent Technology (pp. 655-661). Australia: IEEE.

2008 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology

Quality Information Retrieval for the World Wide Web

Milly Kc
University of Wollongong
Wollongong, NSW 2522
millykc@uow.edu.au

Markus Hagenbuchner
University of Wollongong
Wollongong, NSW 2522
markus@uow.edu.au

Ah Chung Tsoi
Hong Kong Baptist University
Hong Kong
act@hkbu.edu.hk

Abstract

The World Wide Web is an unregulated communication medium which exhibits very limited means of quality control. Quality assurance has become a key issue for many information retrieval services on the Internet, e.g. web search engines. This paper introduces some quality evaluation and assessment methods to assess the quality of web pages. The proposed quality evaluation mechanisms are based on a set of quality criteria which were extracted from a targeted user survey. A weighted algorithmic interpretation of the most significant user quoted quality criteria is proposed. In addition, the paper utilizes machine learning methods to produce a prediction of quality for web pages before they are downloaded. The set of quality criteria allows us to implement a web search engine with quality ranking schemes, leading to web crawlers which can crawl directly quality web pages. The proposed approaches produce some very promising results on a sizable web repository.

1 Introduction

The amount and variety of materials on the World Wide Web has made it a popular medium for information retrieval activities. The Web is largely unregulated which renders its contents vary in degrees of quality and reliability; consequently, retrieval of quality information has become a research focus in recent years [1, 3, 8, 11]. Moreover, current popular information retrieval systems, e.g. Google, Yahoo! are reported to only index a small portion of information from the Web, and the resources required to retrieve and index all the information from the Web cannot hope to keep up with its predicted rate of growth. To prevent the waste of resources it is desirable to ensure that the information that is retrieved is of some value and meets a certain level of quality standard. Manual filtering (curation) while usually delivers higher quality results should also be avoided due to its resource intensiveness and its slow rate of identifying high quality web pages.

This paper examines the issue of quality assessment of web pages on the Internet with an aim to assist with common information retrieval problems e.g. web search and web page retrieval. Web search requires quality assessment of known crawled pages whereas the task of quality web page retrieval requires some means of predicting the quality of the page before its retrieval. This ability to predict a web page's quality before its retrieval allows the realization of a focused crawler for quality information retrieval. A focused crawler is suitable for retrieving information from a large repository such as the Web with an ability to retrieve targeted pages that match a set of quality criteria without needing to crawl exhaustively. In order to achieve such a goal, the following questions need to be answered:

1. What are the criteria which lead users to consider a document or web page as a source of quality information[1]?
2. How to translate user perception of quality (e.g. as revealed from a user survey) into a machine understandable set of instructions?
3. How can such a measure on quality be predicted reasonably accurately before a web page is retrieved?

To answer these questions, the paper is structured as follows: Section 2 gives an overview of the information on quality retrieval which had been obtained through a user survey. Section 3 proposes algorithmic descriptions of some of the most significant quality criteria extracted from the user survey. Section 4 presents a machine learning approach to the weighting and prediction of the set of quality criteria, Some experimental results are presented in Section 5. Finally, Section 6 draws some conclusions.

2 Quality assessment

The realization of a focused crawler for quality information retrieval requires two components: (1) a mechanism which predicts the quality of a target page, and (2) a verification mechanism which assesses the quality of the page that has been retrieved. An answer to (1) necessitates a set of criteria (either implicitly or explicitly perceived by humans) to retrieve quality information. One way in which such a set of perceived quality criteria may be obtained is

[1]Note that in this paper the term *quality* is independent of any specific topic or specific search query. Quality of documents is to be assessed on the basis of features which a web page exhibits.

978-0-7695-3496-1/08 $25.00 © 2008 IEEE
DOI 10.1109/WIIAT.2008.378

through a qualitative user survey. There will be a need to "translate" this set of perceived quality criteria used by human into ways which can be implemented automatically on computers. One answer to (2) would be to compare the web pages retrieved by the proposed method with those obtained using common search engines. In this section we will describe the design of the user survey and its results while the verification process is reported in Section 5.

A total of 132 participants were asked to address a total of 127 questions which were presented in an electronic form [8]. Some of these questions were targeted at obtaining an overview of information seeking behaviour, to develop an understanding of the users' social background, to understand what type of information may be sought by a user, and to understand how a user's view changes during the course of the survey [8]. For the purpose of the survey, the participating users all had to be currently active academics and/or post-graduate students who engaged Internet surfing regularly for the purpose of retrieving high quality information related to their research/work. It was assumed that the user group would have a relatively high cognitive awareness regarding how they interact with target information and make choices about its quality.

The page limits for this paper does not allow the presentation of all survey questions and to list the responses of users to each of the questions (the details are contained in [8]). Table 1 gives an overview of a number of questions (only a relatively small subset of questions asked during the survey is shown to avoid cluttering of information), and a summary of user responses. It can be observed that a machine implementable algorithm may be quite difficult for questions such as by how much a page contains "Information that is incorrect". One contribution of this paper is to propose computationally efficient methods allowing the implementation of quality assessment criteria (as given by a group of users), and to propose methods to combine these individual quality scores so as to obtain an overall quality measure of a given web page.

The survey [8] provides some "ground truth" data which help to develop an understanding as what criteria affect the perceived quality of a web document. From Table 1 it can be observed that users agreed strongly that some criteria do not affect the quality of a page. On the other hand, some criteria such as those based on "erroneous content" or "spelling errors" are commonly regarded as indicators that influence the perception of the quality of a page strongly.

Table 1. Some web information characteristics and their relationship to perceived quality

Question 17: Indicate how your perception of information quality of a visited web page/website changes when the following characteristics are encountered on those pages.			
	Does not affect	Marginally decreases	Greatly decreases
- Information that lacks an attributed author	2.1%	54.2%	43.8%
- Information that seems unreliable	0.0%	16.7%	83.3%
- Pages that contain numerous spelling errors	4.2%	18.8%	77.1%
- Information that is incorrect	0.0%	14.6%	85.4%
- Pages that contain out-of-date/broken hyperlinks	25.0%	54.2%	20.8%
- Out-of-date information	4.2%	58.3%	37.5%
- Too much information	72.9%	25.0%	2.1%
- Too little information	14.6%	50.0%	35.4%
- Irrelevant Information	27.1%	43.8%	29.2%
- Web pages that are difficult to navigate	35.4%	31.3%	33.3%
- Information that is hard to find	33.3%	35.4%	31.3%
- Information that is bias in nature	10.4%	52.1%	37.5%
- Poorly written information	2.1%	27.1%	70.8%
- "Under Construction" or "Coming Soon" statements	22.9%	31.3%	45.8%
- Information that probably breaches copyright laws	39.6%	29.2%	31.3%
- Information that contains poor grammar	4.2%	25.0%	70.8%
- Information that is clearly erroneous	0.0%	8.3%	91.7%
- Information that lacks credibility	2.1%	16.7%	81.3%
- Information that doesn't meet your information needs	56.3%	18.8%	25.0%

3 Algorithmic interpretation

In the process of developing possible criteria to evaluate the quality of a web page, many features were considered. Here we present a selection of 12 high impact quality criteria. These 12 criteria are categorized into two groups. The first group lists a selection of criteria which can be applied to a given page, where the page content is available for analysis; and the second group can be applied to the hyperlink leading to a page, which is based on the limited information available about the target page. While most of the criteria considered are based on the user survey, we also include some criteria which are the result of independent work with implications to document quality.

3.1 Computing the Quality Score of pages

The following criteria can be applied to any given web page:

Spelling accuracy: 77% of participants from the quality perception survey agreed that pages with numerous spelling errors *greatly* decrease their perception of the page quality [8]. A score can be computed quite simply by the use of general-purpose spell checkers, but there is the possibility that special terminologies or uncommon proper nouns are incorrectly labelled as mis-spelled. To perform this task we used *Aspell*, a publicly available and widely used spell checker, which includes a feature that allowed us the addition of terminologies into its dictionary. The score is expressed as a percentage of correctly spelled words in the document, relative to the total number of unique words in the document. Note that for web documents which do not contain any text but are a composition of images, multimedia content, or others receive a score of 1.

Document size: This aims to identify documents that are

too short to contain sufficient information, which many survey users consider a factor that decreases their perception of the document's quality [8]. Document size component also directly evaluates the amount of actual textual data in a document. A score is computed by counting the number of words in the document, and thus, web documents which contain no text receive a score of 0. More text results in a higher score, but the score will no longer increase after reaching a predetermined maximum threshold of 800.

Explicit indication of authorship: Information about the author is identified to assist users to evaluate the reputability of a web page. The survey showed that 93.8% of participants would have a marginally or greatly decreased perception of the page quality if the page lacks an attributed author [8]. The identification of an attributed author is quite a challenging task due to the lack of standards. Some documents contain the author's name at the top of the document, some at the end of the document, where sometimes they are provided within the body of the document content without any identifiable keyword. The score for this component is calculated using the following steps: (1) if found author metatag, score is 1, otherwise continue with the following steps; (2) extract the body of the document, remove HTML tags; (3) search in the first two lines and the last line for signs of name (2 or 3 consecutive terms with capital letters). The score is 1 if there is evidence of an attributed author, and 0 otherwise. The approach is refined by explicitly looking for supporting evidence such as the keywords *author*, or an email address near suspected author names.

Existence of references: This aims to identify documents that provide referencing information to support the claims and information contained in the document. Referencing information is identified as one of the components to confirm the reliability of the page content which is a quality indicator according to [9]. More than a half of the survey participants also recognize the importance of references by stating that the lack of references to sustain the information greatly decreases the perception of information quality of a web page [8]. The score is calculated using the following steps (1) if found keywords "bibliography" or "references" towards the end of the document, score = 1; (2) otherwise, count the total number of links in the document (t); (3) count the number of links located within the bottom third of the document (s); $s_{ref} = \begin{cases} s/10 & \text{if } s/t >= 0.5 \\ 0 & \text{for } s/t < 0.5 \end{cases}$.

Non-spam probability: This aims to differentiate non-spamming documents from those that could be spam. The computation of this component is based on the research performed by Ntoulas [12], where it is indicated that some web documents attempt to include numerous keywords in the header section of the web page in order to be included in as many query results as possible. The score for this component is the probability of the web page not being a spam by calculating the average word length in the header. According to [12], if the average length exceeds 8, it has a 50% chance of being a spam. The score is calculated as follows: $s_{spam} = \begin{cases} 0.5 & \text{if } k > 8 \\ 1 & \text{if } k <= 8 \end{cases}$, where k is the average length of the words in the header section.

Grammatical correctness: Survey results [8] showed that 63.8% of participants believe that their perception of the page quality would greatly decrease if the page contains grammatical errors. The score for this component is computed based on the number of grammatical errors or ambiguities identified by the grammar checker [14]. The score is $s_{gram} = w - \frac{g}{kw}$, where w is the document length score, g the amount of grammatical errors and ambiguity warnings returned by Queequeg grammar checker, and k the average word length. Queequeg grammar checker may not be the most accurate grammar checker but it is one of very few open source grammar checkers that can be executed through command lines in a UNIX environment.

Correctness of content: This component is the top most common dimension in the literature for information quality [9]. The importance of this component is confirmed with 86.3% of survey participants indicating that erroneous information greatly decreases the perceived quality of a web page [8]. Thus, content correctness is an essential index for assessing the quality of a web page. We propose to use a trusted source of information to help assessing the correctness of a given page. Here we propose the use of Wikipedia as a trusted source, as it is currently the largest free and open online encyclopedia, covering a huge range of topics with over 1.8 million articles as of July 2007 [7]. The Wikipedia content is updated by the Internet community collaboratively and regularly, therefore the articles may not be of uniform quality [6]. Nevertheless, it was found that the quality of scientific entries are comparable to an actual encyclopedia such as Encyclopedia Britannica [4]. Hence, it can be assumed that the information presented in the Wikipedia is information on which most users (experts in the field) agree that it is correct. For every document in the Wikipedia dataset, a word frequency vector is produced by using the well-known Bag of Words (BoW) approach [10]. This needs to be performed just once for each document in the Wikipedia dataset. We assess the correctness of the content of a particular web page by comparing its word frequency vector in the BoW approach with the best matching word frequency vector in the Wikipedia dataset. The greater the similarity, the higher the score. This may seem a rather crude approach. However, it turns out that this works quite well, judging from the results obtained in Section 5.

3.2 Computing the Quality Score of links

Some quality criteria can be computed based on properties of individual links to a page, and hence, is based on information that is available before a page is retrieved.

Anchor text: The anchor text is used to indicate the value

and relevance of a link to a descendent page. This approach has been found by [13] to be an effective indicator for executing high-precision focus crawling. The score is dependent on the degree of relevance of the anchor text with respect to the content of the document. The rationale behind this approach is to evaluate how relevant the linked page is to the parent page's topic area. The score is calculated as follows: s_{anc} = frequency of anchor keywords in the document. Where no anchor text could be detected, the score is 0. The scores are normalized to remain within $[0; 1]$.

Link location: This component identifies the location of the link within a web page. This criterion was not listed by user to be of relevance. However, we found that the location of a link can contribute significantly towards the computation of a (predicted) score of a target page. Reasons for this observation is that the link location complements the quality criterium *references*, and hence, we found that pages linked near the beginning of a document are more closely related in terms of quality. The score is calculated as follows: s_{loc} = amount of text after the link/total document size.

Timeliness: This component determines whether the web page is sufficiently up-to-date. Timeliness is the fourth common dimension of information quality according to [9]. This component is especially important for news articles, as news that has been out-of-date is of limited value to users. The score for this component is determined by the last-modified time stamp of the web page, returned by its server, as it is an indication of how up-to-date the information may be. Note that the time stamp is not actually a property of a link itself but can be obtained by requesting a target web page without actually downloading it. As a first approach, the timestamp is converted into the number of seconds since epoch (1/1/1970) and taken away from the current time. A score is computed based on the time difference dt, and a threshold value tt, which is set to the number of seconds in 10 years. Thus, $s_{time} = \begin{cases} 0 & \text{if dt} >= \text{tt} \\ 1 - (dt/tt) & \text{if dt} < \text{tt} \end{cases}$. The time threshold can be made dependent on the domain name to, for example, allow a smaller threshold value when assessing pages within known domains containing news.

Bias: This component aims to identify the probability of the web page's content being biased. Users from the survey [8] pointed out that all information is biased, and that that the generic top level domain (gTLD) influences the perception of bias greatest. For example, the user survey showed that users believe that information from a `.com` domain is most likely biased; however, the probability and amount of bias varies greatly. The score for this component is calculated by identifying the gTLD of the URL, then assigning a bias probability score based on the positive influence of bias directly derived from the survey result [8]. This gives the positive bias score. The probability score for a negative bias is calculated from a survey question that is worded differently so that users consider bias from a different perspective [8]. It should be noted that the positive and negative biases do not necessarily add up to 100% for the same gTLD.

4 Quality prediction

We use a popular machine learning method, known as a multilayer perceptron (MLP) [5] to combine the contributions of each quality score in determining the overall quality of a web page, before its retrieval. We computed the score of a relatively small set of pages. This set is then used as the training set for the MLP where the input is initially a 12-dimensional vector representing the 12 individual quality scores of a source page with respect to a target page being pointed to by the source page, and the output is a 10-dimensional vector representing the quality scores of the target page[2]. In other words, the MLP learns to predict the quality score based on quality information available within a source page. The trained network can then be presented with a source page, and then produces a prediction of the score of a (possibly unknown) target page as output. We used a MLP model featuring a fully connected single layer, non-linear output nodes, and trained it until convergence to a lowest network error had occured. The analysis of the trained network parameters allows to draw conclusions to (a) how to weigh the input scores so as to maximise the correctness of the score prediction, and (b) identify score components which do not help the prediction of scores (i.e. if the associated network weight is close to zero).

5 Implementation and experiment results

For the experiments, a snapshot of a portion of the World Wide Web was taken. The snapshot contained $26,617,303$ HTML documents containing English text from over $5,600$ domains which were retrieved from $1,755$ unique sites. A neural-network simulator [16] is used for the machine learning task. The evaluation of the proposed approach can be categorized into a number of phases:

Phase 1: A first consideration is to analyze the impact of the 12 scoring components on the accuracy of the score prediction. This is achieved by analyzing the internal network parameters of a trained MLP. Training an MLP was done on a subset of $30,635$ web documents[3]. The analysis of network parameters is a common procedure used to identify possible problems with the learning domain. For example, network parameters which are close to zero can be pruned, and hence, the associated input can be made redundant. As another example, network parameters which are very large can indicate a conflict in the input such as contradictions. An alternative approach to achieving such anal-

[2]The output dimension of 10 is due to anchor text and link location which do not assess the quality of the target page but rather the associated links within the source page. Hence, these scores only form part of the input, and not the target.

[3]Care was taken that the training set contained pages that were reachable from a single seed. Hence, this produced a subset which is somewhat smaller than the small dataset containing $30,959$ pages.

ysis is through support vector machines [2]. MLPs is a very appropriate approach since MLP is proven to be a general approximator, and is known to work efficiently with large amounts of data. Hence, it is appropriate in this paper to restrict classification and prediction tasks to MLP.

The analysis revealed that the "authorship existence" component does not help the prediction of a score for target pages and was therefore removed from the list of suitable components. This may be attributed to the algorithm inaccurately extracting authorship information, and hence, can indicate that this module requires improvement in the future. Extra attention was given to the "link location" component, as it does not have as strong theoretical foundation as the other components. It was found that the "link location" alone improves the performance by approximately 2%. As a result of this cycle of experiments, the number of component to be used as score predicting inputs is set to 11.

Phase 2: The quality score is an 11-dimensional vector. It will be much more useful if this vector can be converted to a scalar value representing the overall document quality. We propose to compute the overall score by using a weighted sum over the vector components. Two weighting schemes are considered: one is based on the impact of a score component on the MLP training performance, and, the other is derived from the amount of agreement among survey users.

The first weighting scheme is obtained by investigating the contribution of each component on the ability of the MLP to learn the component scores by comparing the error associated with each target score. In the investigation, the 11 predicted component scores were used as the input, and one of the actual component scores was set as the output. The training would repeat with each of the actual component scores taking turn to be the output, one at a time. The errors from each actual component are compared to reveal that the errors are all similar with less than 1% of difference. As a result, the weights obtained from this approach indicated that each component can be weighed equally.

The weights from the second weighting scheme are derived from the strength of user agreements on each of the criteria. This is to ensure that the features that users consider to strongly affect their perception of a web document's quality are included, and that the amount of influences from the target component scores correspond to the importance of the features in judging the page quality. For example, if 60% of users agreed that correct spelling is an important property of quality documents, and from those, 10% of them strongly agree, then the weighted contribution of that score is $(0.5 + 0.1 \times 2)/2$. Thus, a strong user response is weighted double. The sum is then normalized so as to ensure that the quality score is $\in [0; 1]$.

Using the selected weighting scheme, a single target score is obtained. A different set of weight for the 11 component scores to arrive at a single quality score is then required, with the aim of ensuring that the quality score used to set crawling priorities are predicted as accurately to web pages' actual quality score as possible. The weights for the individual components were obtained after the network was trained using the best performing setting, and were implemented into the focused crawler. A learning performance of 0.00116 MSE on the training data set appears promising, as that indicates the best performance achievable. However, the practical performance can only be verified once the result from the focused crawler using the set of component weights is compared to the quality score of the actual pages.

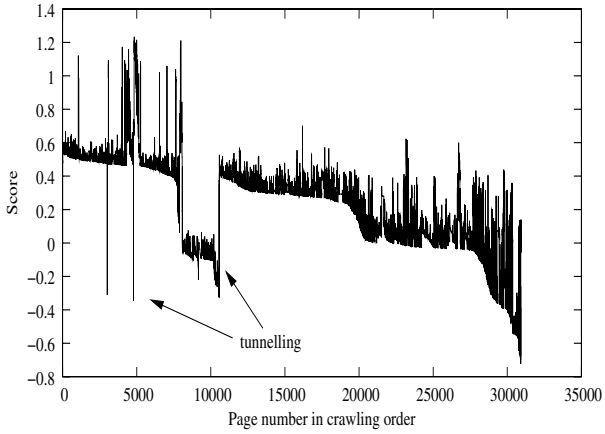

Figure 1. Scores of pages as they are crawled.

Phase 3: Retrieval of quality information: the quality prediction appears to perform well in theory, but as observed by [7, 17] incorporating some form of quality metrices generally improved the effectiveness of searching, therefore the performance also needs to be verified in a practical setting to observe the amount of improvement. This will provide an indication as to whether the performance improvement is significantly better than the general improvement observable with most metrices. In order to observe the practical performance level, the weights from the best theoretical performance in Phase 2 were incorporated into a focused crawler. Due to the fact that no hidden layer was involved in the machine learning process, the process of weight incorporation was straight forward. During the execution of the focused crawler, a weighted sum of the component scores is used as the predicted quality score, according to which the web pages are prioritized during crawling.

To analyze the performance, the predicted score of each retrieved web page is recorded. A first experiment was executed on a small subset of web pages containing $30,959$ pages from 12 hand selected domains. The domains were chosen so as to ensure the inclusion of domains which are known to regulate its content, contain frequently changing information (news), cover the 6 most common gTLDs, and are typical representatives in terms of the size of the do-

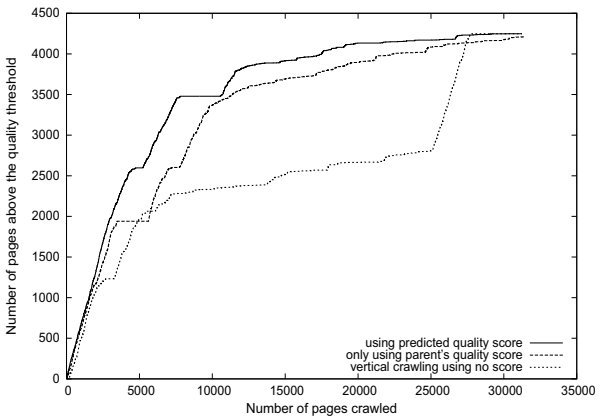

Figure 2. Retrieval rate of high quality web pages using different crawling methods.

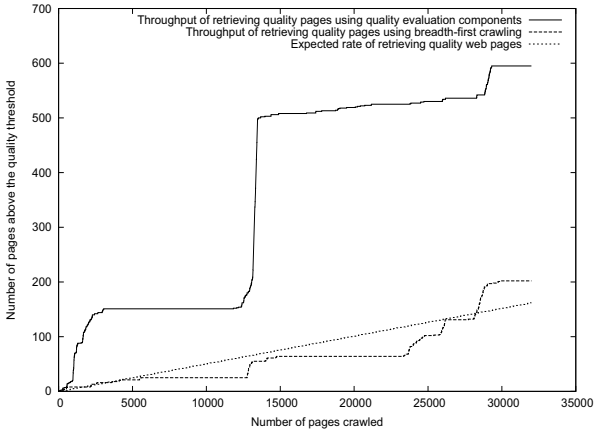

Figure 3. Early retrieval rate of high quality web pages using different crawling methods when applied to a set of 26.6 million pages

main. The seed pages for the crawler were the index pages of the 12 domains. The result is shown in Figure 1. It can be observed that the crawler is indeed predominantly retrieving pages which produce a higher quality score. The sudden drops in score values refer to *tunnelling* taking place (i.e. retrieving pages of relatively low quality score in order to reach pages of higher quality score). The term "tunnelling" refers to a process of finding a path between two distinct clusters of web pages which both meet a given criterion, and is a known issue associated with focused crawlers [15]. The quality scores drop with the continuation of the crawling process, indicating that most of the high quality pages have already been retrieved.

The literature provides only few solutions towards efficient tunnelling. For instance, [15] proposed to probe the neighbourhood around a known group of pages by crawling pages which are not more distant than n links (the n-th neighbourhood) from the group of pages, then select the most promising direction based on an assessment of the neighbourhood. As is shown in [15], the approach is effective if two disjoint groups of relevant pages are not more distant than a distance of 3 links. This paper addresses tunnelling implicitly through an optimization procedure which predicts the score of pages, and hence, the best direction for the focus crawler can be determined without the requirement to probe around a group of known pages as in [15].

Although high scoring pages are retrieved early in the crawling process, showing that the focused crawler works as anticipated, the predicted score may not necessarily correspond to the actual quality score of the page. Therefore, a list of crawled URLs and their actual quality score is maintained so that the web pages above the quality threshold of 70% in the actual score can be identified. A graph is then plotted to compare the efficiency of retrieving high quality pages using different crawling approaches.

In Figure 2, the quality page retrieval rate from 3 different crawling methods is compared. The first method uses the quality index described in this paper; the second method assumes that pages have the same quality score as their source (parent) page, therefore only the un-weighted parent scores are used; and the third method uses vertical crawling with no score at all. As is shown in Figure 2, the proposed method is capable of retrieving a large portion of the quality web pages early in the crawling process, more so than the other methods. Note that the vertical crawler was expected to follow a diagonal line. The better than expected result for the vertical crawler is attributed to the properties of the dataset which contained only few domains, and a much larger percentage of quality domains than is typically observed on the Internet. When the same experiment was subsequently conducted by utilizing a much larger dataset containing over 26 million pages from over $1,300$ domains, it was observed that the improvement in the quality page retrieval rate becomes more significant, as is illustrated in Figure 3.

It can be seen in Figure 3 that the proposed crawler is able to retrieve substantially more quality pages than a standard vertical crawler. A plateau on the curve is caused by the crawler *tunneling* through relatively low quality pages so as to reach higher quality pages. Note that Figure 3 gives an early snapshot into the crawl. It is to highlight that the proposed crawler is effective in avoiding an exhaustive crawl to achieve the retrieval of many high quality pages.

Note that, the computed actual quality score may not necessarily refer to pages of high quality. What can be said is that the pages retrieved by the proposed method meet the quality criteria as were given by users. Whether the retrieved pages meet user expectations is an entirely different

question. In order to answer this question, we designed a search engine which ranks pages by their quality score. The search engine produces two lists of responses to a query, one ranked by quality score, and one using common relevancy meassures. The search engine is publicly accessible at "http://vault.uow.edu.au/searchcmp". Users can vote for the list which they believe contains pages with higher quality information. This is a fair double-blind voting procedure. Some early preliminary results already revealed tendencies that show that users vote about double as frequent for the list that was produced using the proposed quality score. The reader of this paper is encouraged to visit the search web site, obtain an impression, and to participate in the experiment.

6 Conclusions

In this paper we have presented a novel approach to retrieving quality pages from the World Wide Web. As "quality" may have different meaning to different users, here we anchor the concept of "quality" as perceived by a group of reasonably sophisticated and seasoned users when they retrieve information from the Internet as conducted in a user survey [8]. Then based on this user survey we were able to transform the criteria used by these users to machine implementable format. We have carried out experiments with this implementation embedded in a focused crawler. It is found that the focused crawler is capable of retrieving "quality" pages from a closed environment. Moreover it is found that as a side effect the focused crawler is capable of "tunneling" through a landscape of relative low quality web pages or domains to higher quality pages or domains.

A future challenge would be to compare the proposed approach with that of "live" retrieval from the Internet, and to gauge the feedback from users whether the retrieved pages are indeed perceived as of high quality. The beginning of such an experiment is described briefly towards the end of Section 5. Another interesting question to ask is whether there is a correspondence between the results of link analysis to document quality. For example, is it true that pages which feature many in-links are more likely to be pages containing quality information? This is currently being investigated through a comparison of PageRank with the rank of pages computed based on the quality meassure proposed in this paper.

Acknowledgement: This project has been funded in parts by the Australian Research Council in the form of a Discovery Project grant (DP0774168).

References

[1] C. Bizer, R. Cyganiak, O. Maresch, and T. Gauss. The wiqa - web information quality assessment framework. Online publication, 2006. http://sites.wiwiss.fu-berlin.de/suhl/bizer/wiqa/index.htm.

[2] C. J. C. Burges and B. Schölkopf. Improving the accuracy and speed of support vector machines. In M. C. Mozer, M. I. Jordan, and T. Petsche, editors, *Advances in Neural Information Processing Systems*, volume 9, page 375. The MIT Press, 1997.

[3] C. Christian. *Quality-Driven Information Filtering in the Context of Web-Based Information Systems*. PhD thesis, Freie Universität Berlin, Germany, http://sites.wiwiss.fu-berlin.de/suhl/bizer/pub/DisertationChrisBizer.pdf, 2007.

[4] J. Giles. Internet encyclopedias go head to head. *Nature*, 438(1):900–901, December 2005. [online] http://www.nature.com/news/2005/051212/full/438900a.html.

[5] S. Haykin. *Neural Networks, A Comprehensive Foundation*. Macmillan College Publishing Company, Inc., 866 Third Avenue, New York, New York 10022, 1994.

[6] M. Hu, E.-P. Lim, A. Sun, H. W. Lauw, and B.-Q. Vuong. Measuring article quality in wikipedia: Models and evaluation. In *CIKM '07: Proceedings of the 16th International ACM Conference on Information and Knowledge Management*, pages 243–252, Lisboa, Portugal, November 2007. ACM Press.

[7] M. Hu, E.-P. Lim, A. Sun, H. W. Lauw, and B.-Q. Vuong. On improving wikipedia search using article quality. In *WIDM '07 : Proceedings of the 9th International ACM Workshop on Web Information and Data Management*, pages 145–152, Lisboa, Portugal, November 2007. ACM Press.

[8] S. Knight. *The impact of user perceptions of Information Quality on World Wide Web Information Retrieval Strategies*. PhD thesis, School of Management Information Systems, Faculty of Business and Law, Edith Cowen University, Western Australia, Australia, 2007.

[9] S. Knight, J. Burn, and S. Bode. Developing a framework for assessing information quality on the world wide web. *Information Science Journal*, 2005.

[10] A. K. McCallum. Bow: A toolkit for statistical language modeling, text retrieval, classification and clustering. http://www.cs.cmu.edu/ mccallum/bow, 1996.

[11] F. Naumann. *Quality-Driven Query Answering for Integrated Information Systems*, volume 2261 of *Lecture Notes in Computer Science*. Springer, 2002.

[12] A. Ntoulas, M. Najork, M. Manasse, and D. Fetterly. Detecting spam web pages through content analysis. In *WWW '06: Proceedings of the 15th International Conference on World Wide Web*, Edinburgh, Scotland, May 2006.

[13] G. Pant and P. Srinivasan. Link contexts in classifier-guided topical crawlers. *Transactions on Knowledge and Data Engineering*, 18(1):107–122, Jan 2006.

[14] Y. Shinyama. Queequeg, an english grammar checker. [online] http://queequeg.sourceforge.net/index-e.html, 2003.

[15] A. C. Tsoi, D. Forsali, M. Gori, M. Hagenbuchner, and F. Scarselli. A novel focus crawler. In *WWW '03: Proceedings of the 12th International Conference on World Wide Web*, Budapest, Hungary, May 2003. ACM Press.

[16] A. Zell. Simulation of neural networks. [online] http://www-ra.informatik.uni-tuebingen.de/software/snns/, 2007.

[17] X. Zhu and S. Gauch. Incorporating quality metrics in centralized/distributed information retrieval on the world wide web. In *SIGIR '00: Proceedings of the 23rd Annual International ACM conference on research and development in Information Retrieval*, pages 288–295, Athens, Greece, 2000. ACM Press.