2009

Building a prototype for quality information retrieval from the World Wide Web

Milly W. Kc
millykc@uow.edu.au

UNIVERSITY OF WOLLONGONG
COPYRIGHT WARNING

You may print or download ONE copy of this document for the purpose of your own research or study. The University does not authorise you to copy, communicate or otherwise make available electronically to any other person any copyright material contained on this site. You are reminded of the following:

This work is copyright. Apart from any use permitted under the Copyright Act 1968, no part of this work may be reproduced by any process, nor may any other exclusive right be exercised, without the permission of the author.

Copyright owners are entitled to take legal action against persons who infringe their copyright. A reproduction of material that is protected by copyright may be a copyright infringement. A court may impose penalties and award damages in relation to offences and infringements relating to copyright material. Higher penalties may apply, and higher damages may be awarded, for offences and infringements involving the conversion of material into digital or electronic form.

Unless otherwise indicated, the views expressed in this thesis are those of the author and do not necessarily represent the views of the University of Wollongong.

Recommended Citation

Research Online is the open access institutional repository for the University of Wollongong. For further information contact the UOW Library: research-pubs@uow.edu.au
NOTE

This online version of the thesis may have different page formatting and pagination from the paper copy held in the University of Wollongong Library.

UNIVERSITY OF WOLLONGONG

COPYRIGHT WARNING

You may print or download ONE copy of this document for the purpose of your own research or study. The University does not authorise you to copy, communicate or otherwise make available electronically to any other person any copyright material contained on this site. You are reminded of the following:

Copyright owners are entitled to take legal action against persons who infringe their copyright. A reproduction of material that is protected by copyright may be a copyright infringement. A court may impose penalties and award damages in relation to offences and infringements relating to copyright material. Higher penalties may apply, and higher damages may be awarded, for offences and infringements involving the conversion of material into digital or electronic form.
Building a Prototype for Quality Information Retrieval

from the World Wide Web

by

Milly Wei-Tsen Kc

A thesis submitted in partial fulfillment of the requirements for
the award of the degree Doctor of Philosophy
Faculty of Informatics
University of Wollongong

June 2009
This research project has been partially funded by Australian Research Council in the form of a Discovery Project Grant DP0452862 (2004-2006, extended to 2007), and a Linkage International Grant LX0454446 (2004, 2005).

This research was supervised by Dr. Markus Hagenbuchner and Prof. Ah Chung Tsoi
CERTIFICATION

I, Milly Wei-Tsen Kc, declare that this thesis, submitted in partial fulfilment of the requirements for the award of Doctor of Philosophy, in the Faculty of Informatics, University of Wollongong, is wholly my own work unless otherwise referenced or acknowledged. The document has not been submitted for qualifications at any other academic institution.

Milly Wei-Tsen Kc
22 June 2009
Abstract

Given the phenomenal rate by which the World Wide Web is changing, retrieval methods and quality assurance have become bottleneck issues for many information retrieval services on the Internet, e.g. Web search engine designs. In this thesis, approaches that increase the efficiency of information retrieval methods, and provide quality assurance of information obtained from the Web, are developed through the implementation of a quality-focused information retrieval system.

A novel approach to the retrieval of quality information from the Internet is introduced. Implemented as a component of a vertical search application, this results in a focused crawler which is capable of retrieving quality information from the Internet. The three main contributions of this research are: (1) An effective and flexible crawling application that is well-suited for information retrieving tasks on the dynamic World Wide Web (WWW) is implemented. The resulting crawling application (crawler) is designed after having observed the dynamics of the web evolution through regular monitoring of the WWW; it also addresses the shortcomings of some existing crawlers, therefore presenting itself as a practical implementation. (2) A mechanism that converts human quality judgement through user surveys into an algorithm is developed, so that user perceptions of a set of criteria which may lead to determination of the quality content on the web pages concerned, can be applied to a large number of Web documents with minimal manual effort. This was obtained through a relatively large user survey which was conducted in a collaborative research work with Dr Shirlee-Ann Knight of Edith Cowan University. The survey was conducted to determine what criteria Web documents are perceived to meet to qualify as a quality document. This results in an aggregate numeric score for each web page between 0 and 1 respectively indicating that it does not meet any quality criteria, or that it meets all quality criteria perfectly. (3) This research proposes an approach to predict the quality of a web page before it is retrieved by a crawler. The approach allows its incorporation into a vertical search application which focuses on the retrieval of quality information. Experimental results on real world data show that the proposed approach is more effective than any other brute force approaches which have been published so far.

The proposed methods produce a numerical quality score for any text based Web document. This thesis will show that such a score can also be used as a web page ranking criterion for horizontal search engines. As part of this research project, this ranking scheme has been implemented and embedded into a working search engine. The observed user feedback confirms that search
results when ranked by quality scores satisfy user needs more satisfactorily than when ranked by other popular ranking schemes such as PageRank or relevancy ranking. It is also investigated whether the combination of quality score with existing ranking schemes can further enhance the user experience with search engines.
Contribution of this thesis

The contribution of this thesis is multi-fold. This is due to the fact that research on quality information retrieval mechanisms for the World Wide Web is only just evolving, and hence, datasets, domain knowledge, and suitable approaches had to be examined or realized. A successful investigation into quality retrieval methods required access to reliable testbeds. An analysis into existing testbeds revealed that they were incomplete or out-dated, and hence, were no longer reflecting WWW properties. As a result, we developed a distributed crawler which enabled us to retrieve accurate snapshots of a portion of the WWW at regular intervals. In addition, the work for this thesis required a good understanding of the behaviour of web page creation, evolution on the Internet. Existing literature analysed the properties of the WWW as was valid at the time of the examination. We examined the WWW properties on our snapshots in order to verify claims made by others, and in order to understand the WWW as it evolves over time, and detect their trends. The afore-mentioned tasks enabled us to address the quality information retrieval aspect of this thesis. As a result, the contributions of this thesis can be split into several parts as follows:

A.) Development of a scalable and accurate distributed crawler for the WWW: All crawlers known at the commencement of this project implement approximations or exhibit other limitations so as to maximize the throughput of the crawl, and hence, maximize the number of pages that can be retrieved within a given time frame. As a consequence, it is known that existing crawlers are not capable of obtaining accurate snapshots of the Internet. For the purpose of this research, it is essential to have access to an accurate and reliable testbed on which development and experiments can be based. As a consequence, we realized a distributed crawling concept which is designed to avoid such approximations, to reduce the network overhead, and runs on relatively inexpensive hardware. This allowed us to generate regular snapshots of portions of the Internet containing over 27 million web pages in each snapshot.

B.) The analysis of WWW properties, WWW dynamics, and trends: The Internet is continuously changing. It is known that the degree of change in the WWW follows an exponentially increasing curve. Hence, existing literature on WWW properties may no longer reliably reflect properties of the current Internet. This motivated us to verify statements made in the literature through an analysis of the snapshots of the WWW which we obtained at regular intervals. The analysis revealed up-to-date properties of the WWW, enabled us to understand its dynamics, and to detect its trends. The development of quality information retrieval methods benefits from such an analysis in that the awareness of actual changes in
the WWW is taken into account when addressing quality assessment criteria of web pages.

C.) A novel mechanism for predicting web page quality: The aim of any quality information retrieval system is to retrieve documents of high quality without having had prior access to these documents (i.e. to allow the evaluation of the quality of the document). It is thus required that a prediction mechanism to produce a recommendation regarding the order by which documents are presented from within a set of possible candidates. In other words, a mechanism is required which can estimate or predict the quality of a document before it is retrieved such that it becomes possible to decide on which of the possible documents should be retrieved next. This research deployed a machine learning approach to learn to predict document quality on the basis of knowledge about the document and its surroundings. More specifically, parent pages, the links, and the link structure are analysed for indications towards the quality of a target page.

D.) A novel ranking scheme for WWW documents: The method of producing a prediction for web page quality can be readily applied to assess the quality of pages in a web page repository. This associates a numeric value or vector to a document to indicate its quality. As a result, it becomes possible to sort the documents such that high quality documents are listed first whereas documents of lower quality are listed later. In practice, the ordering of web documents according to some criteria is known as web page ranking. Existing criteria are popularity which orders web documents by using link analysis techniques, and relevancy in which pages are ordered with respect to relevancy to a search criterion. This project produced a new web-page ranking criterion based on document quality. The process can be readily applied to realize Internet search engines which will return documents of high quality in response to a search query.

The following list of publications were a direct result of research performed in this thesis.


1The list of publications is sorted by date of publication.


It should be noted that Wei-Tsen Milly Chiang changed her name to Milly Wei-Tsen Kc in 2006, and hence, there is a difference in name in the 2005 publication and subsequent publications.
Glossary

ANN  Artificial Neural Networks aim at emulating the behaviour of neurons or neural assemblies in the brain.

DAG  Directed acyclic graph.

DOAG  Directed ordered acyclic graph.

GraphSOM  A Self Organizing Map capable of processing many types of graphs.

HTML  This is a way to format a document using what is known as hypertext markup language, a special class of markup language for representing Internet documents.

INEX  This is an acronym for “INitiative for the Evaluation of XML Retrieval”, and refers to an international competition on XML structured document mining.

Internet  This refers to the large collection of online resources and services including the World Wide Web (WWW), email, file transfer and others.

Leaf node  is a node in a graph which has no outgoing links. This is sometimes called a frontier node.

Macro F1  A non-weighted performance measure. An average of $F1 = \frac{2 \times \text{precision} \times \text{recall}}{\text{precision} + \text{recall}}$

Micro F1  A weighted performance measure. Average F1 weighted by the number of documents in each class.

ML  Machine Learning.

MLP  Multilayer Perceptron is a neural network model based on artificial neurons that are arranged in layers.

MSE  Mean Squared Error.

Root node  is a node in a graph which has no incoming links.

SOM  Self Organizing Map, a neural network model where neurons are arranged on an n-dimensional grid, with $n = 2$ most commonly. This is used often used for the projection of high dimensional data to one with lower dimensions, with grid points being represented by neurons.

SOM-SD  Self Organizing Map for Structured Data. Similar to SOM but for the encoding of structured data.
CSOM-SD Contextual Self Organizing Map for Structured Data. This is a SOM-SD which includes the context of nodes to the learning process.

SSE Summed Squared Error.

TLD Top Level Domain, the end bit of a domain name. For example, “.de” is the TLD for the domain www.uni-ulm.de .

Tree A tree is a particular type of acyclic connected graphs where each node has at most one parent.

VQ Vector quantization.

Web A shortened form of World Wide Web, which generally refers to a system of documents accessible via the Internet.

Web document This refers to a document found on the World Wide Web which may be an HTML-formatted file, a plain text file or a binary file.

Web page This refers to a document which is formatted using the HTML convention.

WWW World Wide Web.
Notation

The following notations are used throughout this thesis. Scalars and constants are indicated by lowercase script letters e.g., \( c \). Parameters for dynamic processes are stated as lowercase Greek letters such as \( \alpha \). Vectors are denoted by lowercase bold letters, e.g., \( \mathbf{v} \). Sets and matrices are denoted by upper case letters, e.g., \( S \). Sometimes, in order to avoid confusion, we use uppercase bold letters e.g., \( \mathbf{M} \) to denote matrices. Calligraphic letters e.g., \( \mathcal{G} \) are used for representing graphs. Domains are indicated by bold calligraphic letters e.g., \( \mathcal{I} \). Lowercase script letters are used to access elements of a vector or matrix. As an example, in order to access the \( i \)-th element of a vector \( \mathbf{v} \) we use \( v_i \). Letters when used in combination with brackets such as in \( f(x, y) \) denote functions. A few examples are given below:

\[
\begin{align*}
n &= |\mathbf{x}| \\
\mathbf{x} &= (x_1, \ldots, x_n) \\
F(\mathbf{x}) &= \text{A function taking a vector as argument.} \\
\mathbf{C} &= \mathbf{A}\mathbf{I} \\
W_{ij} &= \text{refers to the } ij\text{-th element of the matrix } \mathbf{W}. \\
S &= \{0,1,2\} \\
\mathbf{m}_i &= \alpha \mathbf{m}_i \\
\alpha(t) &= \text{The parameter } \alpha \text{ depends on time } t.
\end{align*}
\]
Acknowledgment

I am grateful to Australian research Council (ARC) for the financial support provided in the form of an ARC Discovery Project Grant to Professor A. C. Tsoi which was subsequently transferred to Dr. M. Hagenbuchner, when Professor Tsoi became ineligible as he was employed by the ARC, which made this research project possible. My appreciation also goes to AC3 for providing high performance cluster computing facilities, staff at University of Sienna and University of Padova for providing much needed access to network and computing facilities, and staff at University of Wollongong for accommodating the special needs of this research. These all contributed to the successful execution of the experiments of this research.

Very importantly, I would like to thank my supervisors Dr. Markus Hagenbuchner and Prof. Ah Chung Tsoi for their support and guidance. I appreciate their friendship, valuable opinions and commitment for excellence. I would also like to thank the fantastic researchers I had the pleasure to collaborate with. They are Sweah Liang (Linus) Yong, Dr. Shirlee-Ann Knight, Prof. Franco Scarselli, Prof. Alessandro Sperduti and the many friendly researchers and staff at Edith Cowen University (WA, Australia) and University of Sienna (Italy) who made the collaboration a wonderful experience. My teachers in the past also need a mention for doing a great job passing the knowledge on, especially my honours thesis supervisor, Associate Prof. Carole Alcock, who lead me into the world of research with enthusiasm and provided much appreciated guidance.

On a more personal note, I would like to thank my husband Suresh and my son Ryan for being so understanding and helpful, not to mention always showering me with encouraging words. A big thank you also to my cool parents for providing me wonderful learning environments and the opportunities to explore anything and everything when I was younger. Finally, thanks to all who pitched in to help when I was away from home on research related trips. I am thankful for the many forms of support received during my research; they all contributed to the successful completion of this thesis.
# Contents

1 Introduction ............................................. 1
   1.1 Introduction to the topic area ......................... 1
   1.2 Research motivation ................................... 6
   1.3 Aims and objectives .................................. 7
   1.4 Research design .................................... 8
   1.5 Research scope .................................... 9
   1.6 Thesis outline .................................... 10

2 Background and motivation ................................. 11
   2.1 Introduction ..................................... 11
   2.2 Domain knowledge .................................. 12
      2.2.1 Web size estimation ............................ 14
      2.2.2 The dynamics of the Web ......................... 14
      2.2.3 Power-law distribution ........................... 15
   2.3 Search and retrieval on the World Wide Web .......... 17
      2.3.1 Search engines ............................... 18
      2.3.2 Meta-search engines ............................ 21
      2.3.3 Web directories ............................... 22
   2.4 Crawling models .................................. 24
      2.4.1 Basic crawling .................................. 24
      2.4.2 Focus crawling .................................. 25
      2.4.3 Parallel crawling ............................... 28
      2.4.4 Distributed crawling ......................... 28
   2.5 Scoring and ranking models ............................ 33
      2.5.1 Link-based algorithms ......................... 34
      2.5.2 Usage-based algorithms ....................... 37
2.5.3 Profile-based algorithms ........................................ 38
2.6 Index update approaches ........................................ 41
  2.6.1 Re-crawl strategies ........................................ 41
  2.6.2 Refresh approaches ........................................ 42
2.7 Quality assessment models in the literature ...................... 43
  2.7.1 General quality models .................................... 45
  2.7.2 Quality models for web documents ......................... 47
2.8 Machine learning approaches to quality estimation ................ 52
2.9 Conclusion .................................................... 55

3 Data collection 57
  3.1 Introduction .................................................. 57
  3.2 Existing testbeds ............................................. 58
    3.2.1 WT10G and WT100G benchmark testbeds .................. 58
    3.2.2 INEX 2005 movies collection ............................ 62
    3.2.3 INEX 2006 journal paper collection ...................... 63
  3.3 Data collection instrument .................................... 64
    3.3.1 The stand-alone crawling component .................... 70
    3.3.2 The central coordinating component .................... 75
    3.3.3 Communication strategy .................................. 80
    3.3.4 Realization of design goals ............................. 81
    3.3.5 Experiments ............................................. 82
  3.4 Properties of the snapshots ................................... 89
    3.4.1 Basic properties ......................................... 89
    3.4.2 Implications of the statistics ........................... 92
  3.5 Trends in Web dynamics .................................... 93
    3.5.1 The rate of page inaccessibility ....................... 93
    3.5.2 The variation in file extension usage ................... 95
    3.5.3 Composition of the general Top Level Domains .......... 96
    3.5.4 The type of changes in the page content ............... 99
  3.6 Discussion and conclusion .................................. 100

4 Document grouping through clustering 103
  4.1 Introduction ................................................. 103
  4.2 Proposed approaches ....................................... 104
5 Document properties

5.1 Introduction ..................................... 127
5.2 Document structures ............................ 127
  5.2.1 Inter-document structure ................... 127
  5.2.2 Intra-document structure .................. 129
5.3 Document features ............................... 130
  5.3.1 Layout-based features ...................... 130
  5.3.2 Time-based features ........................ 131
  5.3.3 Usage-based features ...................... 131
  5.3.4 Text-based features ......................... 132
  5.3.5 Semantic-based features .................... 133
  5.3.6 Miscellaneous features ..................... 133
  5.3.7 Quality-based features ..................... 134
5.4 Criteria for quality analysis .................... 135
5.5 Suitability of extracted feature for quality analysis .......... 139
  5.5.1 Accuracy .................................. 140
  5.5.2 Consistency ................................ 141
  5.5.3 Believability .............................. 142
  5.5.4 Understandability ........................... 143
  5.5.5 Completeness .............................. 144
  5.5.6 Concise .................................. 145
  5.5.7 Timeliness ................................ 145
  5.5.8 Security .................................. 145
  5.5.9 Objectivity ................................ 146
6 Quality evaluation ........................................ 149
   6.1 Introduction ........................................... 149
   6.2 Foundation for quality assessment ....................... 150
   6.3 Quality criteria ....................................... 155
      6.3.1 Computing the Quality Score of links .................. 156
      6.3.2 Computing the Quality Score of a given document ........... 159
   6.4 Approaches ........................................... 161
      6.4.1 Quality assessment during crawling .................... 161
      6.4.2 Quality assessment during ranking ..................... 164
   6.5 Weight determination ................................... 165
   6.6 Score estimation ....................................... 166
   6.7 Experimental setting and results ......................... 171
      6.7.1 Phase 1: Scoring component evaluation .................. 171
      6.7.2 Phase 2: Achievable performance for score estimation ........... 174
      6.7.3 Phase 3: Retrieval of quality information ................. 178
   6.8 Conclusion ........................................... 181

7 Quality Information Retrieval ................................ 183
   7.1 Introduction ........................................... 183
   7.2 Implementation rationale .................................. 184
   7.3 Focus crawler with quality estimation feature .............. 186
      7.3.1 The stand-alone crawling component .................... 187
      7.3.2 Calculation of link-based component score ............... 191
      7.3.3 Calculation of page-based component score ............... 194
      7.3.4 Observations ...................................... 200
   7.4 Quality-based ranking algorithm .......................... 200
      7.4.1 Computing the quality score of links .................... 201
      7.4.2 Computing the quality score of a given document ........... 201
      7.4.3 Observations ...................................... 203
   7.5 Experimental setting and results ......................... 203
      7.5.1 Focus crawling performance .......................... 203
      7.5.2 Overall quality information retrieval system performance ....... 207
   7.6 Discussions and conclusions ................................ 215
8 Related work

8.1 Introduction ................................................................. 217
8.2 Changes in web search .................................................. 217
  8.2.1 The provision of search services ................................... 218
  8.2.2 Extensions on searching functionalities .......................... 220
8.3 Improvements in crawling efficiency .................................. 222
8.4 Web page ranking approaches ......................................... 225
  8.4.1 Link analysis based ranking approach ............................ 225
  8.4.2 Machine learning based quality evaluation ....................... 229
8.5 Conclusion ................................................................. 231

9 Discussion and conclusion

9.1 Introduction ................................................................. 233
9.2 Findings and implications ............................................... 234
  9.2.1 The need for automated quality information retrieval ........ 234
  9.2.2 The requirements of information retrieval experiments .......... 234
  9.2.3 The requirements of the quality evaluation process ............ 235
  9.2.4 The overall quality information retrieval system ............... 236
9.3 Contributions .............................................................. 236
9.4 Limitations ................................................................. 238
9.5 Future work ................................................................. 238

Bibliography

A Report of an informal interview with a librarian
# List of Figures

<table>
<thead>
<tr>
<th>Figure</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.1</td>
<td>Illustration of the various elements of an URL</td>
<td>13</td>
</tr>
<tr>
<td>2.2</td>
<td>A visualization of a minute portion of the Web using TouchGraph, showing the connectivity that can be found on the Web</td>
<td>15</td>
</tr>
<tr>
<td>2.3</td>
<td>The documents on the Web displays a bow-tie structure according to the power-law distribution</td>
<td>16</td>
</tr>
<tr>
<td>2.4</td>
<td>A diagram of the architecture of a general search engine</td>
<td>18</td>
</tr>
<tr>
<td>2.5</td>
<td>A diagram of the architecture of a general Meta-search engine</td>
<td>21</td>
</tr>
<tr>
<td>2.6</td>
<td>A possible file structure on a website for illustrating breadth-first and depth-first crawling order</td>
<td>24</td>
</tr>
<tr>
<td>2.7</td>
<td>An illustration of centralized topology for distributed crawling system</td>
<td>29</td>
</tr>
<tr>
<td>2.8</td>
<td>An illustration of decentralized topologies for distributed crawling system</td>
<td>30</td>
</tr>
<tr>
<td>3.1</td>
<td>The vocabulary vector dot product score</td>
<td>62</td>
</tr>
<tr>
<td>3.2</td>
<td>An example of a symbolic link structure</td>
<td>65</td>
</tr>
<tr>
<td>3.3</td>
<td>The structure and components of the proposed distributed crawler</td>
<td>70</td>
</tr>
<tr>
<td>3.4</td>
<td>Illustration of the crawling throughput achieved when increasing number of crawlers are adopted in parallel</td>
<td>86</td>
</tr>
<tr>
<td>3.5</td>
<td>Illustration of the throughput over time during local and international crawl</td>
<td>87</td>
</tr>
<tr>
<td>3.6</td>
<td>The composition of various general TLDs in 1997 and 2004 respectively</td>
<td>97</td>
</tr>
<tr>
<td>4.1</td>
<td>The architecture of a simple Self-Organizing Map. Shown is a two-dimensional map with $8 \times 4$ neurons arranged on a hexagonal grid</td>
<td>105</td>
</tr>
<tr>
<td>4.2</td>
<td>Visualization of the mapping for documents of classes 1-6</td>
<td>117</td>
</tr>
<tr>
<td>4.3</td>
<td>Visualization of the mapping for documents of classes 7-12</td>
<td>118</td>
</tr>
<tr>
<td>4.4</td>
<td>Visualization of the mapping for documents of classes 13-18</td>
<td>118</td>
</tr>
<tr>
<td>4.5</td>
<td>Performance comparison between SOM-SD and CSOM-SD when utilizing both structure and content information</td>
<td>121</td>
</tr>
</tbody>
</table>
List of Tables

2.1 Crawler performance as measured by practical features .......................... 32
2.2 Quality criteria ordered according to their overall importance as recognized in literature ........................................... 51
3.1 Number of papers which employ the WT10G and/or WT100G in experimental settings ........................................... 58
3.2 A comparison of the basic properties of the 2 TREC datasets .................... 59
3.3 A comparing of the statistical properties of the 2 TREC testbeds .................. 60
3.4 Illustration of the effect of symbolic link on the link extraction process of crawling 66
3.5 The average throughput achieved by crawlers in various locations with various crawling approaches ........................................... 83
3.6 The crawling throughput achievable with 1 to 4 simultaneous processes ........ 85
3.7 The basic statistics of the retrieved web snapshots ................................... 90
3.8 The statistical summary of the filtered testbeds developed from web snapshots ... 91
3.9 Accessibility of pages and domains in the WT10G collection. Valid as of September 2004 ........................................... 94
3.10 The top 5 popular file extensions from all hyperlinks found in the web snapshots . 95
3.11 The top 5 popular unique file extensions per page in the web snapshots .......... 96
3.12 TDLs introduced after 2000 .................................................. 97
3.13 Properties of different types of domains ....................................... 98
3.14 Rate of change in different types of domains ..................................... 98
4.1 The topic area and types of journals in the INEX 2006 XML test-bed ........... 113
4.2 The training parameters used for the structure only learning task ............... 116
4.3 Training parameters used for the clustering of both structural and textual information ........................................... 120
4.4 A comparison of test results for structure only clustering ......................... 122