2009

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

Milly W. Kc
millykc@uow.edu.au

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 responce to a search query.

The following list of publications were a direct result of research performed in this thesis [1].

1. M. Kc, M. Hagenbuchner, and A.C. Tsoi. Quality Information Retrieval for the World Wide Web. In *International Conference on Web Intelligence*, Vol.1, pp. 655-661. Sydney, Australia, 9-12 December 2008.

2. M. Kc, M. Hagenbuchner, and A.C. Tsoi. A scalable lightweight distributed crawler for crawling with limited resources. In *International Conference on Web Intelligence*, Vol3., pp. 663-666. Sydney, Australia, 9-12 December 2008.

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

3. M. Hagenbuchner, S. Sperduti, A.C. Tsoi, and M. Kc. Self-organizing maps for cyclic and unbound graphs. In *European Symposium on Artificial Neural Networks*, 203-208 April 2008.

4. M. Hagenbuchner, A.C. Tsoi, A. Sperduti, and M. Kc. Efficient clustering of structured documents using graph self-organizing maps. In N. Fuhr et al., editor, *LNCS 4862, Lecture Notes in Computer Science*, pages 207–221. Springer-Verlag Berlin Heidelberg, 2008.

5. M. Kc, Markus Hagenbuchner, Ah Chung Tsoi, Franco Scarselli, Alessandro Sperduti, and Marco Gori. Xml document mining using contextual self-organizing maps for structures. In *INEX*, pages 510–524, 2006.

6. W.M. Chiang, M. Hagenbuchner, and A.C. Tsoi. The WT10G dataset and the evolution of the web. In *14th International World Wide Web conference*, Alternate track papers and posters, pages 938–939, Chiba city, Japan, May 2005.

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 $F1 = \frac{2 \times precision \times recall}{precision + recall}$

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 α. 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., $\boldsymbol{\mathcal{I}}$. 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 $\mathbf{v}_i$. Letters when used in combination with brackets such as in $f(x, y)$ denote functions. A few examples are given below:

$n = \|\mathbf{x}\|$	n is the dimension of vector $\mathbf{x}$
$\mathbf{x} = (x_1, \ldots, x_n)$	Vector $\mathbf{x}$ consisting of n elements.
$F(\mathbf{x})$	A function taking a vector as argument.
$\mathbf{C} = \mathbf{AI}$	$\mathbf{C}$ is the result of a matrix multiplication.
$\mathbf{W}_{ij}$	refers to the ij-th element of the matrix $\mathbf{W}$.
S = {0,1,2}	A set with three elements.
$\mathbf{m}_i = \alpha\mathbf{m}_i$	Recursive update of the i-th element of vector $\mathbf{m}$
$\alpha(t)$	The parameter α depends on time t.

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

List of Figures

List of Tables