Discovering Web Page Communities for Web-Based Data Management

A Dissertation submitted to
The Department of Mathematics and Computing
Faculty of Sciences
The University of Southern Queensland
Australia

for the degree of
Doctor of Philosophy

by

Jingyu Hou
PhD and BSc, Shanghai

December 2002
Abstract

The World Wide Web is a rich source of information and continues to expand in size and complexity. Mainly because the data on the web is lack of rigid and uniform data models or schemas, how to effectively and efficiently manage web data and retrieve information is becoming a challenge problem. Discovering web page communities, which capture the features of the web and web-based data to find intrinsic relationships among the data, is one of the effective ways to solve this problem.

A web page community is a set of web pages that has its own logical and semantic structures. In this work, we concentrate on the web data in web page format and exploit hyperlink information to discover (construct) web page communities. Three main web page communities are studied in this work: the first one is consisted of hub and authority pages, the second one is composed of relevant web pages with respect to a given page (URL), and the last one is the community with hierarchical cluster structures.

For analysing hyperlinks, we establish a mathematical framework, especially the matrix-based framework, to model hyperlinks. Within this mathematical framework, hyperlink analysis is placed on a solid mathematic base and the results are reliable.

For the web page community that is consisted of hub and authority pages, we focus on eliminating noise pages from the concerned page source to obtain another good quality page source, and in turn improve the quality of web page communities. We propose an innovative noise page elimination algorithm based on the hyperlink matrix model and mathematic operations, especially the singular value decomposition (SVD) of matrix. The proposed algorithm exploits hyperlink information among the web pages, reveals page relationships at a deeper level, and numerically defines thresholds for noise page elimination. The experiment results show the effectiveness and feasibility of the algorithm. This algorithm could also be used solely for web-based data management systems to filter unnecessary web pages and reduce the management cost.

In order to construct a web page community that is consisted of relevant pages with respect to a given page (URL), we propose two hyperlink based relevant page finding algorithms. The first algorithm comes from the extended co-citation analysis of web pages. It is intuitive and easy to be implemented. The second one takes advantage of linear algebra theories to reveal deeper relationships among the web pages and identify relevant pages more precisely and effectively. The corresponding page source construction for these two algorithms can prevent the results from being affected by malicious hyperlinks on the web. The experiment results show the feasibility and effectiveness of the algorithms. The research results could be used to enhance web search by caching the relevant pages for certain searched pages.

For the purpose of clustering web pages to construct a community with its hierarchical cluster structures, we propose an innovative web page similarity measurement that incorporates hyperlink transitivity and page importance (weight).
Based on this similarity measurement, two types of hierarchical web page clustering algorithms are proposed. The first one is the improvement of the conventional K-mean algorithms. It is effective in improving page clustering, but is sensitive to the predefined similarity thresholds for clustering. Another type is the matrix-based hierarchical algorithm. Two algorithms of this type are proposed in this work. One takes cluster-overlapping into consideration, another one does not. The matrix-based algorithms do not require predefined similarity thresholds for clustering, are independent of the order in which the pages are presented, and produce stable clustering results. The matrix-based algorithms exploit intrinsic relationships among web pages within a uniform matrix framework, avoid much influence of human interference in the clustering procedure, and are easy to be implemented for applications. The experiments show the effectiveness of the new similarity measurement and the proposed algorithms in web page clustering improvement.

For applying above mathematical algorithms better in practice, we generalize the web page discovering as a special case of information retrieval and present a visualization system prototype, as well as technical details on visualization algorithm design, to support information retrieval based on linear algebra. The visualization algorithms could be smoothly applied to web applications.

XML is a new standard for data representation and exchange on the Internet. In order to extend our research to cover this important web data, we propose an object representation model (ORM) for XML data. A set of transformation rules and algorithms are established to transform XML data (DTD and XML documents with DTD or without DTD) into this model. This model capsulizes elements of XML data and data manipulation methods. DTD-Tree is also defined to describe the logical structure of DTD. It also can be used as an application program interface (API) for processing DTD, such as transforming a DTD document into the ORM. With this data model, semantic meanings of the tags (elements) in XML data can be used for further research in XML data management and information retrieval, such as community construction for XML data.
Certification of Dissertation

I certify that the ideas, results, analyses, and conclusions reported in this dissertation are entirely my own effort, except where otherwise acknowledged. I also certify that the work is original and has not been previously submitted either in whole or in part for a degree at this or any other universities.

---------------------------------------------------------------
Signature of Candidate                  Date (DD/MM/YYYY)

ENDORSEMENT

---------------------------------------------------------------
Signature of Supervisor(s)               Date (DD/MM/YYYY)

---------------------------------------------------------------
Acknowledgements

I am deeply indebted to my supervisor, Associate Professor Yanchun Zhang, for his help, guidance and encouragement throughout the course of my doctoral program at the University of Southern Queensland, and his criticisms and constructive suggestions on the draft of the dissertation. His patience, insights, research style and the ability to draw results out of his students have been integral to the success of this work and to my education as a researcher. Without his professional guidance and help, this work would not have been possible. I am also grateful to him for providing me with various supports to conduct this study and many invaluable suggestions for my future academic career.

Thanks must also go to my associate supervisor, Dr Jinli Cao, for her help, encouragement and many constructive suggestions throughout my doctoral program. I would like to thank many anonymous referees for their comments on our papers, which are the basis of this dissertation. A special thank should be given to Associate Professor Chris Harman for checking the English of my papers and many other appreciated supports.

I am grateful to the Department of Mathematics and Computing for offering me a Postgraduate Research Scholarship, Tutor and Part-Time Lecturer positions to support my study throughout my PhD program. I am also grateful to the Faculty of Sciences and the Department for supplying good services and providing the finance to travel to several conferences during my time here. My gratitude also goes to the Head of the Department, Professor Tony Roberts, the Manager of Research and Higher Degrees, Ms Ruth Hilton, Ms Christine Bartlett, Mrs Carla Hamilton, all staffs in the Department and Faculty, as well as my friends for their help and supports, which enabled me to concentrate on my research.

Finally, I would like to express my gratitude to my wife Huiming and children Mingxi and Mingyi for their love, support, encouragement, as well as understanding and patience.
Publications Based on This Dissertation


5th Asia Pacific Web Conference (APWeb2003), Xi’an, China, 27-29 September, 2003.
# Table of Contents

1 **Introduction**  
1.1 Overview ................................................................. 1  
1.2 Motivation ............................................................... 6  
1.3 Claims of the Dissertation ........................................ 15  
1.4 Outline of the Dissertation ....................................... 20  

2 **Fundamentals of Hyperlink Analysis and Web Page Community**  
2.1 Introduction ........................................................... 25  
2.2 Hyperlink Models ..................................................... 27  
2.3 Matrix Expression of Hyperlinks .................................. 29  
2.4 HITS Algorithm ......................................................... 31  
2.5 Singular Value Decomposition of Matrix ......................... 34  
2.6 Hyperlink Analysis Applications ................................. 38  

3 **Eliminating Noise Pages for Good Quality Communities**  
3.1 Introduction ........................................................... 43  
3.2 Noise Pages Elimination Algorithm (NPEA) ..................... 47  
3.3 Experimental Results ................................................ 54  
3.4 Related Work and Discussions .................................... 63  
3.5 Conclusions .............................................................. 70  
Appendix: Noise Page Elimination Algorithm (NPEA) .............. 71  

4 **Finding Relevant Web Pages for a Given Page**  
4.1 Introduction ........................................................... 73  
4.2 Extended Co-Citation Algorithm ................................ 77  
4.2.1 Citation and Co-Citation Analysis ............................. 77  
4.2.2 Extended Co-Citation Algorithm ............................. 80  
4.3 Latent Linkage Information (LLI) Algorithm .................... 86  
4.4 Experimental Results ............................................... 92  
4.5 Related Work and Discussions .................................... 100  
4.6 Conclusions .............................................................. 104  
Appendix: Depiction of LLI (Latent Linkage Information) Algorithm ... 105
5 Visualization Support for Information Retrieval 107
5.1 Introduction ................................................................. 107
5.2 Visualization Examples & System Prototype .................. 112
5.3 SVD-based Information Retrieval ................................. 116
5.4 Visualization Algorithms for Information Retrieval ........ 121
  5.4.1 Visualization Algorithm for SVD-based Retrieval ...... 123
  5.4.2 Algorithm for Match Ratio ................................. 125
  5.4.3 Algorithm for Displaying ..................................... 125
5.5 Conclusions .............................................................. 127

6 Web Page Similarity Measurement and Clustering Improvement 129
6.1 Introduction ................................................................. 129
6.2 Web Page Similarity Measurement .............................. 131
  6.2.1 Page Source Construction ....................................... 132
  6.2.2 Page Weight Definition ......................................... 134
  6.2.3 Page Correlation Matrix ....................................... 136
  6.2.4 Page Similarity .................................................. 140
6.3 Hierarchical Web Page Clustering .................................... 145
6.4 Evaluations ............................................................... 148
6.5 Related Work and Discussions ..................................... 152
6.6 Conclusions .............................................................. 156

7 Matrix-Based Hierarchical Web Page Clustering 158
7.1 Introduction ................................................................. 158
7.2 Matrix-Based Clustering Algorithms ............................ 159
  7.2.1 Similarity Matrix Permutation ............................... 160
  7.2.2 Clustering Algorithm from Matrix Partition .............. 162
  7.2.3 Cluster-Overlapping Algorithm ............................. 165
7.3 Evaluations ............................................................... 169
7.4 Conclusions .............................................................. 174
Appendix ................................................................. 175

8 Object Representation Model for XML Data 177
8.1 Introduction ................................................................. 177
8.2 XML & Object Representation Model (ORM) ................. 181
8.3 Transformation Rules from DTD to ORM ..................... 191
8.4 DTD-Tree and Transformation Procedure .................... 198
8.5 Transformation Rules for XML Document with DTD to ORM ... 203
List of Figures

1.1 Logical architecture of a web-based data management system ............ 8

2.1 Construction of approximation matrix \( A_k \) ................................. 37

3.1 Getting new base set with less noise pages by applying the proposed Algorithms ................................................................. 47

3.2 Page measurement change trends for 20 arbitrary selected pages with different values of parameter \( \delta \) ........................................ 60

4.1 Page source \( S \) for the given \( u \) in the \( DH \) Algorithm ......................... 79

4.2 Page source structure for the Extended Co-Citation algorithm .......... 81

4.3 An example of intrinsic page treatment ........................................... 84

4.4 Comparison of \( bcp \), \( dd \), and \( sim \) values for the selected 10 pages ... 99

5.1 Visual selection for constructing a query type ................................. 113

5.2 Visual interface of information retrieval system ............................... 114

5.3 Visualization of the query and retrieved documents ........................ 114

5.4 Information of the mouse pointed document .................................. 115

5.5 Details of retrieved document from the database ............................. 116

5.6 Example of cosine threshold ....................................................... 120

6.1 Structure of the page source \( S \) .................................................. 133

6.2 Example of computing distance between pages ............................. 140

6.3 Example of the similarity ............................................................... 145

6.4 Hierarchical clustering diagram .................................................... 145

6.5 The average \textit{base} cluster accuracy with different clustering thresholds \( (T) \) ................................................................. 151

6.6 The average \textit{leaf} cluster accuracy with different clustering thresholds \( (T) \) ................................................................. 151

6.7 The \textit{overall} average cluster accuracy with different clustering thresholds \( (T) \) ................................................................. 151

6.8 Co-citation relationship between pages ........................................... 154

6.9 A special situation for similarity measurement ............................... 156
7.1 (a) A similarity matrix. (b) The permuted matrix of (a) .......................... 162
7.2 Matrix-based hierarchical clustering diagram ........................................ 165
7.3 Construction of new sub-matrix $SM'_{1,1}$ .............................................. 168
7.4 The average leaf cluster accuracies of the eight clustering algorithms .... 171
7.5 The comparison of $CA2(D)$, $CA1(D)$ and $WK01A$ on the average leaf cluster accuracy ................................................................. 172
7.6 The comparison among the leaf cluster accuracies of $CA2(D)$, $CA1(D)$ and the base cluster accuracies of $WK01A$ ..................................... 172
7.7 The comparison of $PCA2(D)$, $PCA1(D)$ and $WK01A$ on the average leaf cluster accuracy ................................................................. 173
7.8 The comparison among the leaf cluster accuracies of $PCA2(D)$, $PCA1(D)$ and the base cluster accuracies of $WK01A$ ............................... 173

8.1 Structures of two super classes: $XMLDoc$ and $Terminal$ ......................... 189
8.2 Object representation model (ORM) for XML data ................................. 190
8.3 Work description .................................................................................... 191
8.4 DTD-Tree of $bib.dtd$ ........................................................................... 199
8.5(a) The first result of the rule application ............................................... 202
8.5(b) The second result of the rule application .......................................... 203
8.5(c) The third result of the rule application .............................................. 203
8.5(d) The fourth result of the rule application ............................................ 203
8.5(e) The fifth result of the rule application ............................................... 203
8.6 Structure of an XML document in DOM ................................................ 213
List of Tables

3.1 Numerical results for three algorithms $maxAlgo$, $avgAlgo$ and $minAlgo$ .. 56
3.2 Ten arbitrary noise pages ................................................................. 59
3.3 Ten arbitrary topic-related pages ...................................................... 59
3.4 Page measurement changes of noise pages with different values of parameter $\delta$ .............................................................................. 60
3.5 Page measurement changes of topic-related pages with different values of parameter $\delta$ ............................................................... 60
3.6 Top five authorities and hubs for "Harvard" before noise pages are eliminated ................................................................. 61
3.7 Top five authorities and hubs for "Harvard" after noise pages are eliminated ................................................................. 61
3.8 Top five authorities and hubs for "Jaguar" before noise pages are eliminated ................................................................. 62
3.9 Top five authorities and hubs for "Jaguar" after noise pages are eliminated ................................................................. 62

4.1 Top 10 relevant pages returned by the $DH Algorithm$ ........................ 94
4.2 Top 10 relevant pages returned by the $Extended Co-Citation algorithm$ .. 94
4.3 Top 10 relevant pages returned by the $Companion algorithm$ .............. 95
4.4 Top 10 relevant pages returned by the $LLI algorithm$ .......................... 95
4.5 Top 10 relevant pages returned by the "Related Pages" service of $AltaVista$ .............................................................................. 95
4.6 Top 10 relevant pages returned by the "Similar Pages" service of $Google$ .............................................................................. 96
4.7 Randomly selected 10 pages from the page source $BS$ ....................... 99
4.8 Numerical results of $bcp$, $dd$ values and similarities of 10 selected pages in $BS$ .............................................................................. 99

6.1 Examples of some major clusters ..................................................... 152
6.2 Examples of one major cluster with hierarchical structure .................. 152

7.1 The algorithms used for evaluations ................................................. 170
7.2 Examples of some major clusters ..................................................... 174