



**UNIVERSITI PUTRA MALAYSIA**

***DOCUMENT RANKING USING INFORMATION QUALITY CRITERIA  
IN WEBLOG SEARCH ENGINE***

**FATEMEH AZIMZADEH**

**FK 2013 4**

**DOCUMENT RANKING USING INFORMATION QUALITY CRITERIA  
IN WEBLOG SEARCH ENGINE**

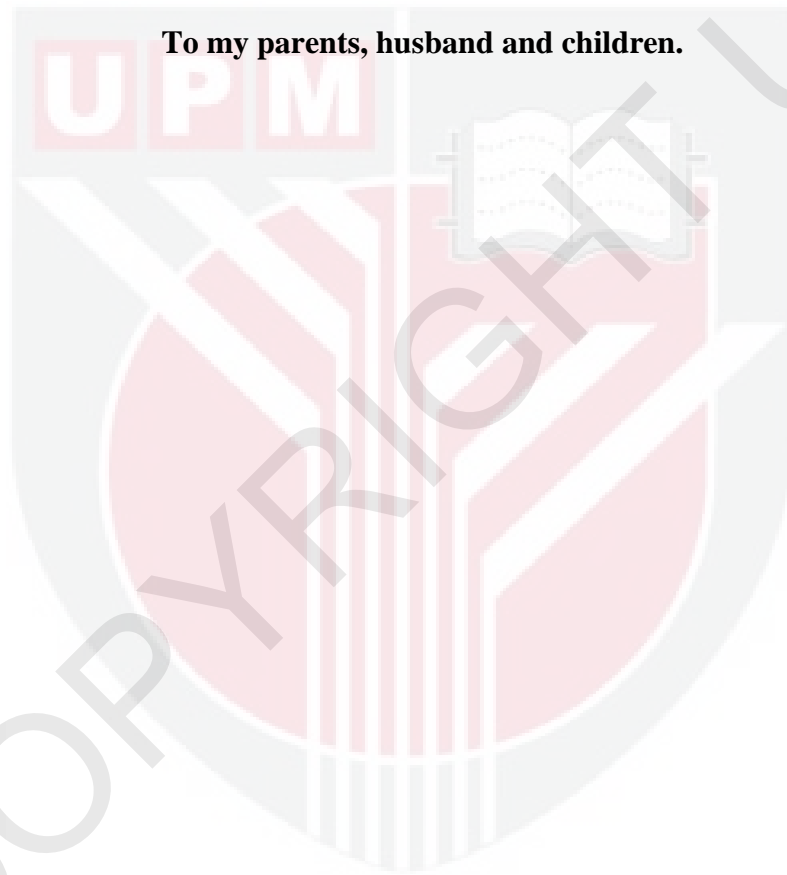
**By**

**FATEMEH AZIMZADEH**

**Thesis Submitted to the School of Graduate Studies, University Putra Malaysia,  
in Fulfilment of the Requirement for the Degree of Doctor of Philosophy**

**January 2013**

**To my parents, husband and children.**



© COPYRIGHT UPM

Abstract of thesis presented to the Senate of Universiti Putra Malaysia in fulfilment of the requirements for the degree of Doctor of Philosophy

**DOCUMENT RANKING USING INFORMATION QUALITY CRITERIA IN  
WEBLOG SEARCH ENGINE**

By

**FATEMEH AZIMZADEH**

**January 2013**

**Chair: Associate Professor Abd Rahman Ramli, PhD**

**Faculty: Engineering**

Social media has revolutionized the Web industry. Weblog medium, fundamentally, is an innovation in personal publishing. It has also come to engender a new form of social interaction on the web. Because much firsthand information is recorded in blog posts, more and more people tend to search their wanted information on blog sites. A major problem is that a weblog includes nontraditional features of the Web pages such as Weblog post, links, tags, and comments. Thus, the use of traditional rank algorithms like PageRank and HITS in general search engines are not appropriate to evaluate the Weblog posts because such algorithms do not consider the blog specific features.

On the other hand, information quality criteria are important factors for the users. From Weblogs, which have unfiltered information without expert peer review, users expect that search engines deliver quality information for their queries. There has

been little framework which consider information quality criteria in the Weblog search engine. This thesis establishes an integrated framework which incorporates information quality criteria into the ranking function of search engine on Persian weblogs. The presented framework rank Weblogs and posts based on the selected information quality criteria. Then, the ranking scores are merged with relevancy in the search engine. A ranking method is developed for the Weblog search engine where the post is considered as the document retrieved. This thesis proposes two ranking functions in the search engine which are combined with the information quality criteria, and then compared with a PageRank based ranking function. The results reveal that combination of quality criteria with relevancy, without suitable weight for each one, does not lead to user's satisfaction. Instead, applying proper weights to both information quality factors and relevancy intelligibly improve the results of the search engine and consequently lead to user satisfaction.

Abstrak tesis yang dikemukakan kepada Senat Universiti Putra Malaysia sebagai memenuhi keperluan untuk ijazah Doktor Falsafah

**KEDUDUKAN DOKUMENT DENGAN MENGGUNAKAN KRITERIA  
KUALITI MAKLUMAT DALAM ENJIN WEBLOG CARIAN**

Oleh

**FATEMEH AZIMZADEH**

**Januari 2013**

**Pengerusi: Profesor Madya Abd Rahman Ramli, PhD**

**Fakulti: Kejuruteraan**

Media sosial telah merevolusikan industri Web. Medium Weblog pada dasarnya adalah satu inovasi dalam penerbitan peribadi juga telah menghasilkan satu bentuk interaksi sosial yang baru dalam Web. Disebabkan banyak maklumat yang berterusan direkodkan dalam pos blog, semakin banyak orang berkecenderungan untuk mencari maklumat yang diperlukan oleh mereka dalam laman blog. Satu masalah utama ialah sesebuah Weblog memasukkan sifat-sifat yang bukan tradisi. Contohnya adalah pos Weblog, capaian, para tanda dan komen. Sejurus itu, algoritma tradisi dalam enjin carian yang biasa tidak sesuai untuk menilai pos-pos Weblog kerana algoritma tersebut tidak mempertimbangkan sifat blog yang spesifik.

Selain daripada itu kriteria kualiti maklumat adalah satu faktor yang penting untuk pengguna. Tidak seperti Weblog yang maklumatnya tidak disaring tanpa ulasan teliti daripada pakar, pengguna-pengguna menjangkakan enjin carian itu mengeluarkan maklumat yang ber kualiti mengikut kehendak mereka. Setakat ini, walau bagaimanapun, telahpun ada beberapa rangka kerja yang mempertimbangkan

kriteria kualiti maklumat dalam enjin carian Weblog. Tesis ini memperkenalkan satu integrasi rangka kerja yang menggabungkan kriteria kualiti maklumat ke dalam fungsi kedudukan enjin carian dalam Weblog Parsi. Sistem yang dibangunkan ini adalah enjin carian pertama yang didedikasikan kepada Weblog Persian. Kedudukan rangka kerja Weblog dan pos yang dibentangkan terdiri daripada kriteria kualiti maklumat yang terpilih. Kemudian, skor kedudukan itu digabungkan dengan perkaitan dalam enjin carian itu. Tesis ini mencadangkan dua fungsi kedudukan dalam enjin pencarian yang digabungkan dengan kriteria kualiti maklumat, dan kemudian membandingkan mereka dengan fungsi PageRank berdasarkan kedudukan. Keputusan menunjukkan balawa penggabungan kriteria kualiti dengan perkaitan tanpa pemberat yang sesuai selalunya tidak memenuhi kehendak pengguna. Sebaliknya, penggunaan pemberat yang sesuai ke atas kedua-dua faktor kualiti maklumat dan perkaitan secara logiknya meningkatkan hasil carian enjin dan akibatnya menjurus kepada kepuasan pengguna.

## ACKNOWLEDGEMENTS

Acknowledgement is not a play of words, but an attitude of mind. If words are considered as the symbol of approval and tokens of appreciation, then let the words play the heralding role to expressing my gratitude.

Thanks to Allah, who with His willing give me the opportunity to complete this thesis.

I am heartily thankful to my supervisor, Associate Professor Abdul Rahman Ramli, whose encouragement, guidance and support me; I would like to express my deepest thanks co- supervisors Professor Borhanuddin Mohd Ali and Professor Hamidah Ibrahim for their advice and constant support.

Deepest thanks and appreciation to my husband, parents, family, my children, and others for their cooperation, encouragement, constructive suggestion and full of support for the report completion, from the beginning till the end. Also thanks to all of my friends and everyone, those have been contributed by supporting my work and help myself during this four years thesis progress till it is fully completed.



I certify that a Thesis Examination Committee has met on 31 January 2013 to conduct the final examination of Fatemeh Azim Zadeh on her thesis entitled “Document Ranking Using Information Quality Criteria in Weblog Search Engine” in accordance with the Universities and University College Act 1971 and the Constitution of the Universiti Putra Malaysia [P.U.(A) 106] 15 March 1998. The Committee recommends that the student be awarded the Doctor of Philosophy.

Members of the Examination Committee were as follows:

**M. Iqbal bin Saripan, PhD**

Associate Professor  
Faculty of Engineering  
Universiti Putra Malaysia  
(Chairman)

**Wan Azizun binti Wan Adnan, PhD**

Senior Lecturer  
Faculty of Engineering  
Universiti Putra Malaysia  
(Internal Examiner)

**Syamsiah binti Mashohor, PhD**

Senior Lecturer  
Faculty of Engineering  
Universiti Putra Malaysia  
(Internal Examiner)

**Sachio Hirokawa, PhD**

Professor  
Kyushu University  
Japan  
(External Examiner)

---

**SEOW HENG FONG, PhD**

Professor and Deputy Dean  
School of Graduate Studies  
Universiti Putra Malaysia

Date: 21 March 2013

This thesis submitted to the Senate of University Putra Malaysia and has been accepted as fulfilment of the requirement for the degree of Doctor of Philosophy.

The members of the Supervisory Committee were as follows:

**Abd Rahman Ramli, PhD**

Associate Professor  
Faculty of Engineering  
Universiti Putra Malaysia  
(Chairman)

**Borhanuddin Mohd Ali, PhD**

Professor  
Faculty of Engineering  
Universiti Putra Malaysia  
(Member)

**Hamidah Ibrahim, PhD**

Professor  
Faculty of Computer Science and Information Technology  
Universiti Putra Malaysia  
(Member)

---

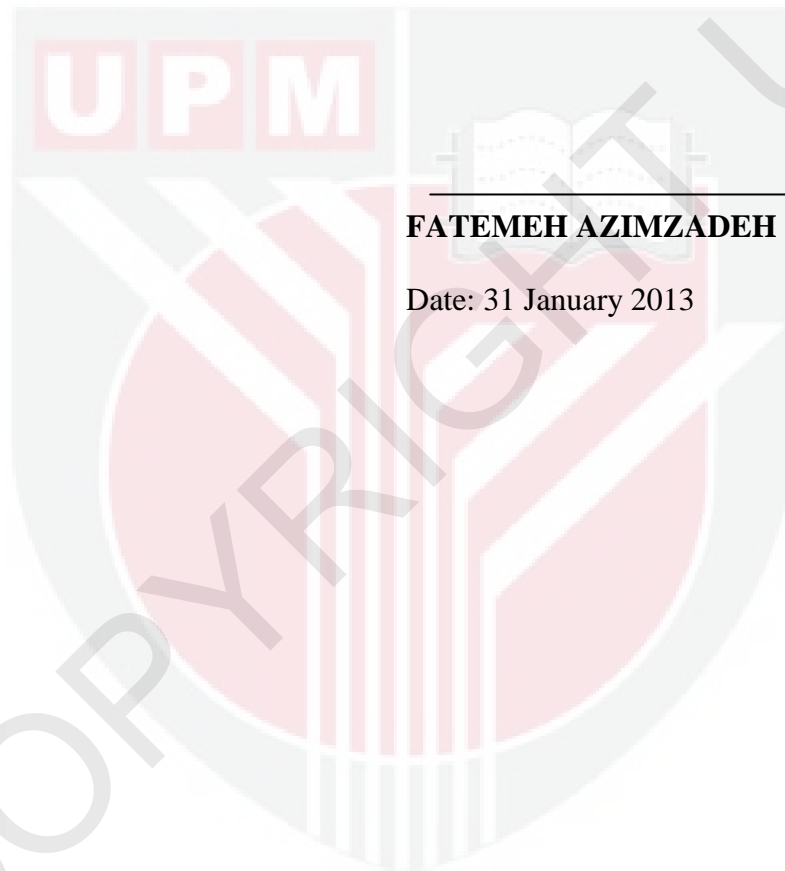
**BUJANG BIN KIM HUAT, PhD**

Professor and Dean  
School of Graduate Studies  
Universiti Pura Malaysia

Date:

## DECLARATION

I declare that the thesis is my original work except for quotations and citations which have been duly acknowledged. I also declare that it has not been previously or concurrently, submitted for any other degree at Universiti Putra Malaysia or at any other institution.



**FATEMEH AZIMZADEH**

Date: 31 January 2013



## TABLE OF CONTENTS

	<b>Page</b>
<b>ABSTRACT</b>	iii
<b>ABSTRAK</b>	v
<b>ACKNOWLEDGEMENTS</b>	vii
<b>DECLARATION</b>	x
<b>APPROVAL</b>	viii
<b>LIST OF TABLES</b>	xiii
<b>LIST OF FIGURES</b>	xiv
<b>LIST OF ABBREVIATIONS</b>	xv
<b>CHAPTER</b>	
<b>1 INTRODUCTION</b>	<b>1</b>
1.1 Background	1
1.2 Motivation and Problem Statements	5
1.3 Research Aim and Objectives	8
1.4 Scope of the Research	8
1.5 Research Contributions	9
1.6 Brief Methodology	10
1.7 Organization of the Thesis	10
<b>2 LITERATURE REVIEW</b>	<b>11</b>
2.1 Introduction	11
2.2 Weblog and Social Networks	11
2.2.1 Blog Posts and Blog Links	13
2.3 Information Quality Criteria	15
2.3.1 Weblog Information Quality Criteria	15
2.4 Information Retrieval Systems	17
2.4.1 Web Search Engine	18
2.5 Ranking in Web Search	23
2.5.1 Relevancy	25
2.6 Quality Search Engine	28
2.6.1 Criteria in General Search Engine	32
2.6.2 Weblog Quality Search Engine	34
2.6.3 Evaluation of the Quality-Based Search Engine	41
2.7 Background of Materials	43
2.7.1 Content Management System (CMS)	44
2.7.2 Information Retrieval Policy	46
2.7.3 Analytic Hierarchy Process (AHP)	49
2.7.4 Bubble Sort Distance	51
2.7.5 Score Normalization	52
2.8 Summary	53
<b>3 RESEARCH METHODOLOGY AND DESIGN</b>	<b>54</b>
3.1 Introduction	54
3.2 Methodology Overview	55
3.3 Architecture	57
3.4 Flowchart of the Research Methodology	60
3.5 IQ Measurement Subsystem	64

	3.5.1 Weblog Quality Score	64
	3.5.2 Weblog Quality Criteria	66
	3.5.3 Weblog IQ Criteria in Search Engines	67
	3.5.4 Post IQ Criteria in Search Engines	69
	3.5.5 IQ Criteria and Assessment Method	69
	3.5.6 Total Weblog Quality Score	76
	3.5.7 Post Quality Score	76
	3.5.8 Incorporating IQ Framework into Weblog Management System	78
	3.6 Weblog Construction	79
	3.7 Exporting the Database Output to SPSS	81
	3.7.1 Data Cleaning	81
	3.7.2 Pearson Correlation	83
	3.8 Practical Indexing	84
	3.9 Searching and Ranking by Lucene	86
	3.9.1 Lucene Ranking Functions	88
	3.10 Ranking Function	92
	3.10.1 Merging the Relevancy Score and the Quality Score	96
	3.10.2 Applying Weights to Criteria in Search Engine	99
	3.10.3 B2Rank	106
	3.10.4 Comparison	112
	3.11 Summary	113
4	<b>RESULTS AND DISCUSSIONS</b>	115
	4.1 Introduction	115
	4.2 Data Collection	116
	4.3 Data Cleansing	116
	4.4 Correlation Analysis on Data Sets	119
	4.4.1 Weblog Data Set	120
	4.4.2 Post Data Set	126
	4.4.3 Search Engine Data Set	127
	4.5 Information Quality Dimensions for Weblogs	130
	4.6 PageRank and Information Quality Criteria Approach	136
	4.7 Evaluation of the Ranking Functions in Search Engine	138
	4.8 Summary	142
5	<b>CONCLUSION AND RECOMMENDATION FOR FUTURE RESEARCH</b>	144
	5.1 Conclusion	144
	5.2 Suggestion for Future Works	146
	<b>REFERENCES</b>	148
	<b>APPENDICES</b>	158
	<b>BIODATA OF STUDENT</b>	164
	<b>LIST OF PUBLICATIONS</b>	165