Blogs Search Engine Adopting RSS Syndication Using Fuzzy Logic

ATHRAA JASIM MOHAMMED

UNIVERSITI UTARA MALAYSIA

2012

Blogs Search Engine Adopting RSS Syndication Using Fuzzy Logic

A project submitted to the School of Computing
In partial fulfillment of the requirements for the degree
Master of Science (Intelligent Systems)
Universiti Utara Malaysia

By

Athraa Jasim Mohammed

PERMISSION TO USE

In presenting this project in partial fulfillment of the requirements for a postgraduate degree from University Utara Malaysia, I agree that the University Library may make it freely available for inspection. I further agree that permission for coping of this project in any manner, in whole or in part, for scholarly purpose may be granted by my supervisor(s) or, in their absence by the Dean of the Graduate School.

It is understood that any copying or publication or use of this project or parts thereof for financial gain shall not be allowed without my written permission. It is also understood that due recognition shall be given to me and to University Utara Malaysia for any scholarly use which may be made of any material from my project. Requests for permission to copy or to make other use materials in this project, in whole or in part should be addressed to:

Dean of Awang Had Salleh Graduate School
of Arts and Sciences
University Utara Malaysia
06010 UUM Sintok
Kedah Darul Aman

ABSTRAK

Perkembangan pesat Internet meningkatkan jumlah laman blog. Kadang-kadang laman blog ini memberi tumpuan kepada menyelesaikan beberapa masalah penting. Untuk mencari blog tertentu adalah masalah keras untuk pengguna kerana banyak blog-blog ini mengandungi maklumat kurang tepat seperti iklan dalam talian, notis dan bunyi bising yang meminimumkan ketepatan laman blog. Malah untuk mendapatkan blog yang lebih relevan adalah satu lagi masalah yang merendahkan prestasi carian. Kajian ini mencadangkan blog enjin carian menerima pakai sindikasi RSS menggunakan logik fuzi. Enjin blog carian terdiri daripada tiga fasa utama iaitu 'crawling' menggunakan suapan algoritma RSS, algoritma weblog pengindeksan dan teknik mencari dengan logik fuzi. Dalam ;crawling' RSS, merangkak proses suapan RSS perlu dikumpulkan untuk mengekstrak maklumat yang berguna seperti tajuk, pautan, menyiarkan masa dan keterangan. Weblog pengindeksan meng gunakan pautan untuk mendapatkan semula laman blog untuk pemprosesan teks dan membina pangkalan data pengindeksan. Dalam usaha untuk mendapatkan maklumat yang diperlukan oleh mana-mana pengguna, terdapat antara muka pengguna untuk mencari kata kunci dengan darjah kepentingan dan mengira ketumpatan kata kunci daripada pangkalan data pengindeksan. Ketepatan halaman yang dikira berdasarkan nilai 'fuzzy weighted average'. Prototaip dibina menggunakan Visual Basic 2008 untuk mengesahkan enjin carian blog yang dicadangkan. Ia adalah sebuah aplikasi window dengan protokol sambungan http. Dalam penilaian sistem, dua pengukuran digunalcan iatitu 'precision' dan 'mean average precision'. Ketepatan parameter adalah bergantung kepada responden yang menentukan

jumlah yang diambil pautan dan jumlah pautan yang berkenaan untuk keputusan carian kata kunci. Bilangan kata kunci yang digunakan dalam sistem ujian adalah lima pasang kata kunci. Keputusan eksperimen menunjukkan ketepatan min purata adalah 81.7% daripada keseluruhan prestasi sistem. Peratus daripada responden adalah 80% yang tahu dan merupakan pengguna blog dan 20% tidak mempunyai pengetahuan tentang blog. Masa pelaksanaan sistem yang berdasarkan responden adalah 70% antara 3-5 minit dan 30% kurang daripada 3 minit. Peratusan ini adalah baik. Dengan mengambil kira Kadar kepuasan bagi sistem adalah 80% berpuas hati dan 20% amat berpuas hati.

ABSTRACT

The rapid development of Internet increases the writers of blog sites. Sometimes these blog sites focused on solving some important problems. To find specific blogs are hard problem for the users because a lot of these blogs contain unuseful information such as online advertisements, notice and noise which minimize the rank of blog site. Furthermore to retrieve more relevant blogs is another problem which lowering the search performance. This study proposes blogs search engine adopting RSS syndication using Fuzzy logic. The blogs search engine consists of three main phases which are crawling using RSS feeds algorithm, indexing weblogs algorithm and searching technique with Fuzzy logic. In RSS crawling process RSS feeds need to be gathered to extract useful information such as title, links, publish time and description. Indexing weblogs use the links to retrieve the blogs sites for text processing and construct indexing database. In order to retrieve such information needed by any user, there is user interface to search for keyword with importance degree and compute the density of keyword from the indexing database. The rank of the pages is computed based on fuzzy weighted average value. A prototype is built using visual basic 2008 to validate the proposed blogs search engine. It is a windows application with http connection protocol. In system evaluation used two measurement performances which are precision and mean average precision. The parameters of precision determine based on respondents whom determine the total retrieved links and the total relevant links for the keyword search result. The number of keywords that used in testing system is five pairs keywords. The experimental results show that the mean average precision is

81.7% of the whole system performance. The percent of respondents is 80% who knows and uses the blogs and 20% don't have knowledge. The execution time of the system based on respondents is 70% between 3-5 minute and 30% less than 3 minute. This percentage is good considering the rate of satisfaction for system is 80% satisfied and 20% strongly satisfied.

DEDICATION

I would like to dedicate this thesis to my country "IRAQ"

ACKNOWLEDGMENTS

My gratitude sincere and my appreciation is given to my research supervisor, Dr. Husniza binti Husni, for her invaluable insights, professional advice, scholarly guidance.

I would like to extend my thanks to all my lecturers of the University Utara Malaysia, for their supported during my study.

Special thanks to my family, my husband and my children, for their encouragement, their help, and their continued support during the period of my studies.

Thank you everyone.

TABLE OF CONTENT

CHAPTER FOUR: PROTOTYPE DEVELOPMENT

CHAPTER FIVE: RESULTS AND FINDINGS

CHAPTER SIX: CONCLUSION AND RECOMMENDATIONS

LIST OF TABLE

LIST OF FIGURE

ABBREVIATION

RSS ……..……..….Really Simple Syndication

XML …….……...……Extensible Markup Language

URL …….....…….Uniform Resource Locator

HTML …….....…Hypertext Markup Language

Vlog …….…..……..……..Video blogging

SE …….…….……...Sentiment Evaluation

PMI-IR …….….....Pointwise Mutual Information and Information Retrieval

WDM ……..……...…….Web-tree determination module

WMM ……..……..…Weight Measure Module

TEM ……..…….....Top-k Enumeration Module

WISDOM ……..…Web Information Spread Data Operations Model

FTCA ……...…….…..Fuzzy Transduction-Based Clustering Algorithm

TRM ……...……Transduction Based Relevance Model

STC …….……...Suffix Tree Clustering

PSA ……...……Porter Stemming Algorithm

FWA ……..…….Fuzzy Weighted Average

HTTP ……...……Hypertext Transfer Protocol

CPU ……..……Central Processing Unit

RAM …….……..Random Access Memory

MAP …….……...Mean Average Precision

LIST OF APPENDICES

CHAPTER ONE

INTRODUCTION

In the past few years, the great popularity of online communities has caused huge amounts of web data, which led to the development of people, consumption of information and **i**ncreasing content delivered in streams. A stream is a series of text documents that arrive over time led to subscribe to blogs. A weblog or blog is a "frequently updated Web page with dated entries in reverse chronological order, usually containing links with commentary" (Gao, Tian, Huang & Yang, 2010). Mainly people use streams as daily activities to talk about their interests and opinion and also share information with others. Usually users seek for streams on a specific topic that is published regularly for the subscription purpose so that they can get new updates. Instead of encountering streams by accident while on the move Internet, a search engine has become necessary to find interested streams (Park, Shin, Kim & Chung, 2010).

A search engine is simply defined as "a web site used to easily locate internet resources" (Meghabghab & Kandel, 2008). Search engines able to do the information retrieval process by adopting Artificial Intelligence techniques (Meghabghab & Kandel, 2008). A search engine consists of three parts; the first part is the crawlers or spiders. The crawlers are software programs that work on the Internet to gather information. The second part is the indexer, whereby it created a database from website after many preprocessing techniques. This database operates behind the search webpage. The search webpage is the last part of search engine. It is interface of search engine where user writes his/her keyword to find relevant information (Bouras, Poulopoulos & Silintziris, 2009). There are one technique and one format

The contents of the thesis is for internal user only

References

Bouras, C., Poulopoulos, V., & Silintziris, P. (2009). Personalized news search in www: adapting on user's behaviour**,** *fourth international conference on internet and web applications and* services, 125-130, doi: 10.1109/ICIW.2009.25.

Bracewell, D., Gustafson, S., Moitra, A., & Steuben, G. (2010). WISDOM from light-weight information retrieval, *IEEE international conference on social computing / IEEE international conference on privacy, security, risk and trust*, 347-354, doi: 10.1109/SocialCom.2010.57.

Chong, T. (2010). A kind of algorithm for page ranking based on classified tree in search engine, *International conference on computer application and system modeling (ICCASM 2010)*, 04 November 2010, v13-538 - v13-541, doi: 10.1109/ICCASM.2010.5622891.

Ding, L., Finin, T., Joshi, A., Pan, P., Cost, R., Peng, Y., Reddivari, P., Doshi, V., & Sachs, J. (2004). Swoogle: a search and metadata engine for the semantic web, CIKM'04, November 8–13, 2004, Washington, DC, USA, ACM,652-659, doi: 10.1145/1031171.1031289.

Gao, W., Tian ,Y., Huang ,T., & Yang ,Q. (2010). Vlogging: a survey of videoblogging technology on the web, ACM Comput. Surv. 42(4), Article 15 (June 2010), 57 pages, doi:10.1145/1749603.1749606.

Gulli, A. (2005). The anatomy of a news search engine, *ACM*, May 10–14,2005, Chiba, Japan,880-881, doi: 10.1145/1062745.1062778.

Hirokawa, S., Yin, C. , & Nakatoh, T. (2011). Component-based search engine for blogs, *2011 IEEE international conference on fuzzy systems*, June 27-30, 2011, Taipei, Taiwan, 1074-1078, doi: 10.1109/FUZZY.2011.6007650.

Jiang, Z., & Deng, X. (2010). A personalized search engine model based on RSS user's interest, *2010 2nd international conference on future computer and communication*, V2-196 - V2-199, doi: 10.1109/ICFCC.2010.5497371.

Keong, B., & Anthony, P. (2011). Meta search engine powered by DBpedia, *2011 international conference on semantic technology and information retrieval*, 28-29 June 2011, Putrajaya, Malaysia, 89-93,doi: 10.1109/STAIR.2011.5995770.

Kim, K., & Cho, S. (2001). A personalized web search engine using fuzzy concept network with link structure, *IFSA world congress and 20th NAFIPS international conference, 2001*. Joint 9th, 81 - 86 vol.1, doi: 10.1109/NAFIPS.2001.944231.

Lai, L., Wu, C., Lin, P., & Huang, L. (2011). Developing a fuzzy search engine based on fuzzy ontology and semantic search, *2011 IEEE international conference on fuzzy systems*, June 27-30, 2011, Taipei, Taiwan, 2684-2689, doi: 10.1109/FUZZY.2011.6007378.

Laughlin, A., Olson, J., Simpson, D., & Inoue, A. (2011). Page ranking refinement using fuzzy sets and logic, *Proceedings of The 22nd Midwest artificial intelligence and cognitive science conference 2011*,Cincinnati, USA, April 16-17, 2011,40-46.

Lee, W., Jung-Hoon, J., Kim, Y., & Kai-Sang, C. (2009). AnchorWoman: top-k structured mobile web search engine, CIKM'09, November 2–6, 2009, Hong Kong, China, ACM, 2089-2090, doi: 10.1145/1645953.1646317.

Li, G., Ji, S., Li, C., Wang, J.,& Feng, J. (2010). Efficient fuzzy type-ahead search in TASTIER, *ICDE conference 2010*, IEEE, 1105-1108, doi: 10.1109/ICDE.2010.5447804.

Lin, Y., Lai, L., Wu, C.,& Huang, L. (2010). A self-adaptation approach to fuzzy-go search engine, *Computer symposium (ICS), 2010 International IEEE*, 1020-1025, doi: 10.1109/COMPSYM.2010.5685543.

Matsumoto, T., & Hung, E. (2010). Fuzzy clustering and relevance ranking of web search results with differentiating cluster label generation, *Fuzzy Systems (FUZZ), IEEE international conference on2010*, 23 September 2010, 1 – 8, doi: 10.1109/fuzzy.2010.5584771.

Meghabghab, G., & Kandel, A. (2008). *Search engines ,link analysis ,and user's web behaviour*, Berlin Heidelberg: Springer-Verlag.

Negnevitsky, M. (2011). *Artificial intelligence: a guide to intelligent systems*, third edition, Wesley.

Park, J., Shin, Y., Kim, K., & Chung, B. (2010). Searching the long tail of social media streams on the web, *IEEE intelligent systems*, 09 November 2010, doi: 10.1109/MIS.2010.115.

Pavlacka ,O.,& Talasova, J. (2006). Application of the fuzzy weighted average of fuzzy numbers in decision making models.

Phoey Lee, T., Abdul Ghani, A., Ibrahim, H., & Atan, R. (2009). Coalescence of XML-based Really Simple Syndication(RSS) aggregator for blogosphere, *Proceedings of the 7th International Conference on Advances in Mobile Computing and Multimedia*, December 14–16, 2009, Kuala Lumpur, Malaysia, ACM, 530-534, doi: 10.1145/1821748.1821850.

Shang, W., Wang, T., & Lv, R. (2011). The key technology research of intelligent information syndication, *2011 fourth international joint conference on computational sciences and optimization*, 865-867, doi: 10.1109/cso.2011.275.

Snasel, V., Kromer, P., Nyongesa, H., Musilek, P., & Husek, D. (2007). Fuzzy modeling of user needs for improvement of web search queries, *Fuzzy information processing society, 2007. NAFIPS '07. annual meeting of the north American*, 24-27 June 2007, 446 – 451,doi: 10.1109/nafips.2007.383881.

Topac, V. (2010). Efficient fuzzy search enabled hash map, *4th international workshop on soft computing applications*, 15-17 July, 2010 -Arad, Romania, IEEE, 39-44, doi: 10.1109/SOFA.2010.5565628.

Xu, G., Zhang, Y., & Li, L. (2011). Web mining and social networking, techniques and application, New York, Springer, doi 10.1007/978-4419-7735-9.

Yang, S., Zi-tao, L., Cheng, L., & Ye, L. (2009). Research on social network based on meta-search engine, *2009 sixth web information systems and applications conference*, 179-183, doi 10.1109/wisa.2009.21.

Zhang, X., Xu, C., Cheng, J., Lu, H., & Ma, S. (2009). Effective annotation and search for video blogs with integration of context and content analysis. *IEE*E transactions on multimedia, 11(2), 272-285, doi: 10.1109/TMM.2008.2009689.

Zhou, Y., Chen, X., & Wang, C. (2006). A self-organizing search engine for RSS syndicated web contents, *Proceedings of the 22nd international conference on data engineering workshops (ICDEW'06)*, 24 April 2006, Atlanta, GA, USA, IEEE Computer Society, doi: 10.1109/ICDEW.2006.19.

Zhu, J., & Wang, H. (2010). Application of E-commerce personality searching based on RSS, *2010 2nd IEEE international conference on information management and engineering (ICIME)*, 197– 199, doi: 10.1109/ICIME.2010.5478085.