UNIVERSITI PUTRA MALAYSIA

TERM FREQUENCY AND INVERSE DOCUMENT FREQUENCY WITH POSITION SCORE AND MEAN VALUE FOR MINING WEB CONTENT OUTLIERS

WAN RUSILA BINTI WAN ZULKIFELI

FSKTM 2013 8

TERM FREQUENCY AND INVERSE DOCUMENT FREQUENCY WITH POSITION SCORE AND MEAN VALUE FOR MINING WEB CONTENT OUTLIERS

By

WAN RUSILA BINTI WAN ZULKIFELI

Thesis Submitted to the School of Graduate Studies,
Universiti Putra Malaysia, in Fulfillment of the
Requirements for the Degree of
Master of Science

December 2013

Alhamdulillah..
I lovingly dedicated this thesis to my..

Husband..
Who supported me each step of the way,
Thank you for your love, encouragement and sacrifices.

Son..
Who gave me strength and motivation to finish my study.

Parents ~ Mama, Papa, Mak, Abah..
Who believe in me,
Thank you for all the moral support, guidance and patience.

Sister..
Who offered time, thought and lough,
Thank you for the hand you always lend me.

This is a thesis that I..
Truly wasn't able to finish without anyone of them.

Thank you.
Barokallahufikum.

Abstract of thesis presented to the Senate of Universiti Putra Malaysia in fulfilment of the requirement for the degree of Master of Science

TERM FREQUENCY AND INVERSE DOCUMENT FREQUENCY WITH POSITION SCORE AND MEAN VALUE FOR MINING WEB CONTENT OUTLIERS

By

WAN RUSILA BT WAN ZULKIFELI

December 2013

Chairman: **Norwati Mustapha, PhD**

Faculty: **Computer Science and Information Technology**

In the past few years, there was a rapid expansion of activities in the Web Content Mining area. However, the focus was only on the technical, visual design and frequent web content pattern while less frequent web content pattern called outliers was undervalued. Mining Web Content Outliers is used to detect irrelevant web content within a web portal. It is important to detect outliers especially when a web portal is hacked. Recently, there are only a few approaches suggested to Mining Web Content Outliers such as Signed-with-Weight technique and mining through mathematical approach. The mathematical approach developed is based on two way rectangular representations and correlation method. However the approaches do not take the advantage of position score and stemmed domain dictionary. Position score and stemmed domain dictionary are very useful in mining web content outliers because it may effects on reduction the relevance of documents.

Therefore, this study was made to resolve the problems in Mining Web Content Outliers by combining the strength of word-based techniques, position score weighting technique and stemmed domain dictionary. The existing weighting technique was transformed to the Term Frequency and Inverse Document Frequency with Position Score and Mean Value (TF.IDF.PSM) weighting technique by implementing a standard weighting technique from Information Retrieval called Term Frequency and Inverse Document Frequency (TF.IDF) and a weighting technique from Text Categorization called the Term Frequency and Relevance Frequency (TF.RF) into Web Content Mining. This technique is started with extracting the web pages, preprocess it and then generate the full word profile. Depending on the length of the character, the respective index on the stemmed domain dictionary is searched. Positive count is incremented by one, if the word is present in the dictionary and document. Then word frequency in a web page and in every web pages and position score are counted. Finally the dissimilarity measure is computed to determine outliers. In the dissimilarity measure part, the TF.IDF.PSM is used not only to calculate and analyze the relevant words but also to consider the importance of the irrelevant words by assigning weight based on the word position in a page. A statistical approach '*mean*' is added to balance the weight of position score.

The technique has been tested on 431 web pages from the Course folder of University Wisconsin, provided by World Wide Knowledge Base. While the 43 benchmark dataset is from Science Medical folder provided by The 20 Newsgroups Dataset. Term Frequency and Inverse Document Frequency (TF.IDF) weighting technique from Information Retrieval (IR) and the Term Frequency and Relevance Frequency (TF.RF) weighting technique by Text Categorization are used during experimental phase and the results are qualified by two parameters which is the percentage of the accuracy and the F1-measure. The experimental results show that the TF.IDF.PSM weighting technique achieves up to 98.95% of accuracy, which is about 3.21% higher than the Signed-with-Weight technique. Besides, it also achieves up to 94.19% of F1-measure, which is a 18.12% improvement from the Signed-with-Weight technique.

Abstrak tesis yang dikemukakan kepada Senat Universiti Putra Malaysia sebagai memenuhi keperluan untuk ijazah Master Sains

PEMBERAT KEKERAPAN PERKATAAN DAN KEKERAPAN DOKUMEN SONGSANG BERSAMA PERINCIAN KEDUDUKAN DAN NILAI PURATA UNTUK MELOMBONG KANDUNGAN *OUTLIERS* DI LAMAN SESAWANG

Oleh

WAN RUSILA BT WAN ZULKIFELI

December 2013

Pengerusi: **Norwati Mustapha, PhD**

Fakulti: **Sains Komputer dan Teknologi Maklumat**

Dalam beberapa tahun yang lepas, bidang perlombongan kandungan laman sesawang telah mengalami perkembangan yang amat pesat. Bagaimanapun, fokus hanya diberikan kepada aspek teknikal, rekabentuk visual dan corak kandungan laman sesawang yang kerap, dimana corak kandungan laman sesawang yang kurang kerap yang dipanggil 'outliers' sering diabaikan. Perlombongan kandungan laman sesawang 'outliers' digunakan untuk mengesan kandungan laman sesawang yang tidak relevan didalam sesuatu portal laman sesawang. Adalah penting untuk mengesan 'outliers' terutamanya apabila portal laman sesawang itu telah digodam. Sehingga kini, hanya terdapat beberapa pendekatan yang mencadangkan penggunaan perlombongan kandungan laman sesawang 'outliers', antaranya teknik 'Signed-with-Weight' dan perlombongan menggunakan pendekatan matematik. Pendekatan matematik dibangunkan berdasarkan kepada teknik Perwakilan dan Korelasi Segi Empat Tepat Dua Hala. Bagaimanapun pendekatan ini tidak menggunakan kelebihan yang ada pada perincian kedudukan dan perpustakaan domain yang telah disunting. Perincian kedudukan dan perpustakaan domain yang telah disunting adalah sangat berguna

dalam perlombongan kandungan sesawang 'outliers, kerana ianya memberi kesan pada pengurangan dokumen-dokumen yang relevan.

Disebabkan itu, kajian ini telah dibuat untuk menyelesaikan masalah di dalam perlombongan kandungan laman sesawang 'outliers' dengan menggabungkan kekuatan yang ada pada teknik berdasarkan teks, teknik pemberat perincian kedudukan dan juga perpustakaan domain 'stem'. Teknik pemberat yang asal telah diubahsuai menjadi teknik pemberat Kekerapan Perkataan dan Kekerapan Dokumen Songsang bersama Perincian Kedudukan dan Nilai Purata (TF.IDF.PSM) dengan menggabungkan teknik pemberat asas daripada 'information retrievel' (TF.IDF) dan 'text categorization' (TF.RF) ke dalam perlombongan kandungan laman sesawang. Teknik ini dimulakan dengan menguraikan halaman laman sesawang, kemudian ia diproses dan profil perkataan penuh akan dijana. Indeks yang berkaitan di dalam perpustakaan domain 'stem' akan di cari bergantung kepada panjang perkataan tersebut. Jika perkataan itu dijumpai di dalam perpustakaan dan juga di dalam dokumen, kiraan positif akan ditambah satu. Frekuensi perkataan di dalam halaman laman sesawang dan juga di setiap halaman laman sesawang dan markah kedudukan akan dibilang. Akhir sekali, timbangan ketidaksamaan akan di kira untuk menentukan 'outliers'. Pada bahagian timbangan ketidaksamaan, teknik Kekerapan Perkataan dan Kekerapan Dokumen Songsang bersama Perincian Kedudukan dan Nilai Purata (TF.IDF.PSM) digunakan bukan sahaja untuk mengira dan menganalisa perkataan yang berkaitan, tetapi juga untuk mempertimbangkan kepentingan perkataan yang tidak berkaitan dengan meletakkan pemberat berdasarkan kepada kedudukan perkataan di dalam sesebuah halaman. Pendekatan statistik 'min' ditambah untuk mengimbangkan pemberat pada perincian kedudukan.

Teknik ini telah diuji pada 431 halaman laman sesawang daripada fail kursus Universiti Wisconsin yang disediakan oleh 'World Wide Knowledge Base'. Manakala 43 set data penanda aras adalah daripada fail Sains Perubatan yang disediakan oleh 'The 20 Newsgroup Dataset'. Teknik pemberat Kekerapan Perkataan dan Kekerapan Dokumen Songsang (TF.IDF) daripada perolehan maklumat dan teknik pemberat Kekerapan Perkataan dan Kekerapan Relevan (TF.RF) dengan pengkategorian perkataan telah digunakan semasa fasa eksperimen dan keputusan yang diperolehi telah diukur berdasarkan dua parameter iaitu peratusan ketepatan dan juga 'F1-measure'. Keputusan eksperimen menunjukkan yang teknik pemberat Kekerapan Perkataan dan Kekerapan Dokumen Songsang bersama Perincian Kedudukan dan Nilai Purata (TF.IDF.PSM) telah mencapai ketepatan sehingga 98.95% yang mana merupakan 3.21% lebih tinggi daripada teknik 'Signed-with-Weight'. Selain itu, ia juga memperolehi sehingga 94.19% markah 'F1-measure', yang merupakan 18.12% lebih tinggi daripada markah yang diperolehi dengan teknik 'Signed-with-Weight'.

ACKNOWLEDGEMENTS

I would like to extend my gratitude to people who helped to bring this research successful. First, I would like to thank Associate Professor Norwati Mustapha for her help, professionalism, valuable guidance, and support throughout my entire program of study.

Secondly, I would like to express my sincere thanks and appreciation to the supervisory committee member, Dr. Aida Mustapha for her guidance and valuable suggestions in making this work a success. My sincere appreciation also goes to all my friends in UPM for their continuous help and sharing of knowledge.

I would also like to thank the expert who was involved in the validation for this research project, Assistant Professor Pookuzhali Sugumaran. Without her passionate participation and input, the validation could not have been successfully conducted.

Finally, I must express my very profound gratitude to my parents and to my husband for providing me with unfailing support and continuous encouragement throughout my years of study and through the process of researching and writing this thesis. This accomplishment would not have been possible without them. Thank you.

Wan Rusila Bt Wan Zulkifeli
December 2013

APPROVAL

This thesis was submitted to the Senate of Universiti Putra Malaysia and has been accepted as fulfilment of the requirement for the degree of Master of Science. The members of the Supervisory Committee were as follows:

Norwati Binti Mustapha, PhD
Associate Professor
Faculty of Computer Science and Information Technology
Universiti Putra Malaysia
(Chairman)

Aida Binti Mustapha, PhD
Faculty of Computer Science and Information Technology
Universiti Putra Malaysia
(Member)

____

BUJANG BIN KIM HUAT,
PhD

Professor and Dean
School of Graduate Studies
Universiti Putra Malaysia

Date:

Declaration by graduate student

I hereby confirm that:

- this thesis is my original work;
- quotations, illustrations and citations have been duly referenced;
- this thesis has not been submitted previously or concurrently for any other degree at any other institutions;
- intellectual property from the thesis and copyright of the thesis are fully-owned by Universiti Putra Malaysia, as according to the Universiti Putra Malaysia (Research) Rules 2012;
- written permission must be obtained from supervisor and the office of Deputy Vice-Chancellor (Research and Innovation) before thesis is published in book form;
- there is no plagiarism or data falsification/fabrication in the thesis, and scholarly integrity is upheld as according to the Universiti Putra Malaysia (Graduate Studies) Rules 2003 (Revision 2012-2013) and the University Putra Malaysia (Research) Rules 2012. The thesis has undergone plagiarism detection software.

Signature: ______________ Date: ______________

Name and Matric No.: Wan Rusila Binti Wan Zulkifeli, GS22215

____
___

Declaration by Members of Supervisory Committee

This is to confirm that:

- the research conducted and the writing of this thesis was under our supervision;
  - supervision responsibilities as stated in the Universiti Putra Malaysia (Graduate Studies) Rules 2003 (revision 2012-2013) are adhered to.

Signature: ____________________
Name of
Chairman of
Supervisory
Committee: ____________________

Signature: ____________________
Name of
Member of
Supervisory
Committee: ____________________

TABLE OF CONTENTS