



شبكة المعلومات الجامعية

بِسْمِ اللَّهِ الرَّحْمَنِ الرَّحِيمِ



شبكة المعلومات الجامعية

جامعة عين شمس

التوثيق الالكتروني والميكرو فيلم

قسم

نقسم بالله العظيم أن المادة التي تم توثيقها وتسجيلها
علي هذه الأفلام قد أعدت دون أية تغيرات



يجب أن

تحفظ هذه الأفلام بعيدا عن الغبار

في درجة حرارة من ١٥-٢٥ مئوية ورطوبة نسبية من ٢٠-٤٠%

To be Kept away from Dust in Dry Cool place of
15-25- c and relative humidity 20-40 %



شبكة المعلومات الجامعية التوثيق الالكتروني والميكروفيلم

بعض الوثائق الأصلية تالفة

بالرسالة صفحات لم ترد بالاصل

TOWARDS MINING WEB CONTENT OUTLIERS

By

Ayman Hassan Tanira

A Thesis Submitted to the
Faculty of Computers and Information
Cairo University
in Partial Fulfillment of the
Requirements for the Degree of

Master of Science

in

COMPUTER SCIENCE

Under the Supervision of

Prof. Dr. Ahmed Abd El-Wahed Rafea

Computer Science Department, School of Science and Engineering
The American University in Cairo

Dr. Hesham Ahmed Hassan

Computer Science Department, Faculty of Computers and Information
Cairo University

CP
CXCE

**FACULTY OF COMPUTERS AND INFORMATION
CAIRO UNIVERSITY, EGYPT
NOVEMBER, 2007**

TOWARDS MINING WEB CONTENT OUTLIERS

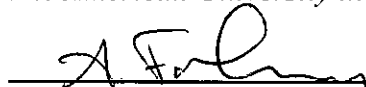
By

Ayman Hassan Tanira

A Thesis Submitted to the
Faculty of Computers and Information
Cairo University
in Partial Fulfillment of the
Requirements for the Degree of
Master of Science
in
COMPUTER SCIENCE

Approved by the Examining Committee:


Prof. Dr. Ahmed Abd El-Wahed Rafea
*School of Science and Engineering,
The American University in Cairo*


Prof. Dr. Aly Aly Fahmy
*Dean of Faculty of Computers and Information,
Cairo University*


Prof. Dr. Mohsen Abd El-Razik Rashwan
*Faculty of Engineering,
Cairo University*


Dr. Hesham Ahmed Hassan
*Faculty of Computers and Information,
Cairo University*

FACULTY OF COMPUTERS AND INFORMATION
CAIRO UNIVERSITY, EYGPT
NOVEMBER, 2007

CERTIFICATION

I certify that this work has not been accepted in substance for any academic degree and is not being concurrently submitted in candidature for any other degree.

Any portions of this thesis for which I indebted to other sources are mentioned and explicit references are given.

Student Name: *Ayman Hassan Tanira*

Signature: *Ayman Tanira*

ACKNOWLEDGEMENTS

First of all, my greatest thanks go to Almighty ALLAH who has made all things possible for me.

I would like to express my gratitude and thanks to my supervisor, *Prof. Dr. Ahmed Rafea* from Computer Science Department, School of Science and Engineering, the American University in Cairo, for his continuous interest and his technical directions through all stages of this research. I am grateful to him for giving me much of his time discussing many aspects of this study.

I would also like to thank my second supervisor *Dr. Hesham Hassan* from Computer Science Department, Faculty of Computers and Information, Cairo University, for his indispensable helping and encouragement through all stages of this research.

Special thanks go to *Dr. Malik Agyemang* from Computer Science Department, University of Calgary, Canada, for his scientific advices in the early stages of this research.

Finally, I would like to thank my supportive wife, *Iman*, and my wonderful sons; *Mohaned* and *Mohammed*, for their patience and shouldering the alienation for me.

ABSTRACT

The task of outlier detection is to find small fraction of data that are exceptional when compared with rest large amount of data. Finding outliers from huge data repositories is like finding needles in a haystack. The existing outlier detection algorithms were designed for mining numeric data which cannot be applied directly to mine outliers from Web datasets because the Web contains data of different types such as: text, hypertext, images, video, audio, etc.

Web content outliers are Web documents with varying contents compared to other documents taken from the same category. Mining Web document outliers may lead to the identification of competitors, emerging business trends in electronic commerce, improving the quality of results obtained from a Web search engine, and cleaning corpus used in Web documents classification.

This thesis concentrates on enhancing current approaches for detecting Web document outliers. It introduces a Web document outlier mining system aiming to detect outlying Web documents from a particular category. A major advantage of the proposed system is that it does not depend on a domain dictionary. Thus, it can be used for any domain of the interest of a miner. The system achieves its objectives by passing through three phases: preparation, processing, and detection. The preparation phase converts each Web document in the collected documents into what is called a feature frequency profile. This work provides two types of features: single word (i.e., keyword) and character N-grams (i.e., 4-grams, 5-grams). The processing phase uses a vector space model to represent the document collection as a matrix of feature weights. Then it computes pair-wise document dissimilarities resulting document dissimilarity matrix. The detection phase exploits document dissimilarity matrix for detecting Web document outliers. In this research, we adapted a novel algorithm called FindWDO algorithm for the detection phase of our proposed system. The FindWDO algorithm uses a k -nearest neighbors' method for identifying the top outlying Web documents. It also uses a simple pruning rule to reduce the running time required for identifying the closest neighbors for every document in the collection.

The experimental results on two different datasets with embedded motifs showed that FindWDO with N-grams outperforms similar algorithms in the same domain with respect to the accuracy of results.

TABLE OF CONTENTS

APPROVAL PAGE.....	ii
CERTIFICATION.....	iii
ACKNOWLEDGEMENTS.....	iv
ABSTRACT.....	v
TABLE OF CONTENTS.....	vi
LIST OF TABLES.....	viii
LIST OF FIGURES.....	x
LIST OF ABBREVIATIONS.....	xii
 CHAPTER 1: INTRODUCTION.....	 1
1.1 Background.....	2
1.1.1 Web Mining.....	2
1.1.2 Web outliers.....	3
1.2 Problem Definition.....	4
1.3 Motivation for Mining Web Content Outliers.....	5
1.4 Thesis Objective and Research Approach.....	6
1.5 Thesis Organization.....	6
 CHAPTER 2: OUTLIER DETECTION: A SURVEY.....	 8
2.1 Overview.....	8
2.2 Outlier Definition.....	9
2.3 Taxonomy of Outlier Detection Techniques.....	9
2.4 Statistical Techniques.....	10
2.4.1 Distribution-Based Techniques.....	10
2.4.2 Depth-Based Techniques.....	15
2.5 Machine Learning Techniques.....	18
2.5.1 Distance-Based Techniques.....	18
2.5.2 Density-Based Techniques.....	22
2.5.3 Clustering-Based Techniques.....	26
2.5.4 Rule-Based Techniques.....	30
2.6 Textual Techniques.....	33
2.6.1 Ontology-Based Techniques.....	33
2.6.2 Vector-Based Techniques.....	35
2.7 Concluding Remarks.....	40
 CHAPTER 3: A PROPOSED WEB DOCUMENT OUTLIER MINING SYSTEM.....	 42
3.1 Overall Structure of the System.....	42
3.2 Document Extraction.....	46
3.3 Content Preprocessing.....	47
3.3.1 Data Selection.....	48
3.3.2 Tokenization.....	48
3.3.3 Stop-Words Removing.....	49
3.4 Feature Profile Generation.....	50
3.4.1 Feature Type Selection.....	50

3.4.1.1 Keyword Feature.....	51
3.4.1.2 N-Gram Feature.....	51
3.4.2 Feature Frequency Counting.....	53
3.5 Vector Space Model (VSM).....	56
3.5.1 General Vector Creation.....	56
3.5.2 Dimensionality Reduction.....	57
3.5.3 Document Representation.....	57
3.6 Term-Weighting Method.....	58
3.7 Dissimilarity Computation.....	60
3.8 Outlier Detection.....	61
3.8.1 WCOND-Mine Algorithm.....	62
3.8.2 LOF-Mine Algorithm.....	64
3.8.3 The Proposed Algorithm (FindWDO).....	67
3.9 Outlier Analysis.....	70
CHAPTER 4: EVALUATION OF THE PROPOSED SYSTEM.....	71
4.1 Evaluation Datasets.....	71
4.1.1 HTML Pages Dataset.....	71
4.1.2 Email Messages Dataset.....	72
4.2 Evaluation Methodology.....	73
4.3 Evaluation Criteria.....	74
4.4 Experimental Results.....	74
4.4.1 Testing Effect of the Number of Neighbors on FindWDO Algorithm..	75
4.4.2 Testing Effect of the Pruning Rule on FindWDO Algorithm.....	76
4.4.3 Testing Accuracy on HTML Pages Dataset.....	78
4.4.3.1 WCOND-Mine Algorithm Results.....	78
4.4.3.2 LOF-Mine Algorithm Results.....	80
4.4.3.3 FindWDO Algorithm Results.....	82
4.4.3.4 Comparisons.....	83
4.4.4 Testing Accuracy on Email Messages Dataset.....	85
4.4.4.1 WCOND-Mine Algorithm Results.....	85
4.4.4.2 LOF-Mine Algorithm Results.....	87
4.4.4.3 FindWDO Algorithm Results.....	88
4.4.4.4 Comparisons.....	90
4.4.5 Testing for Running Time.....	91
4.4.5.1 WCOND-Mine Algorithm Results.....	93
4.4.5.2 LOF-Mine Algorithm Results.....	94
4.4.5.3 FindWDO Algorithm Results.....	95
4.4.5.4 Comparisons.....	96
4.5 Discussion.....	98
CHAPTER 5: CONCLUSIONS AND FUTURE WORK.....	101
5.1 Conclusions.....	101
5.2 Future Work.....	103
REFERENCES.....	105

LIST OF TABLES

Table 3.1: A proposed simple coding method shows token's location and its code...	49
Table 3.2: An output table of tokenization step shows each token and its location...	49
Table 3.3: An example shows the three feature generation methods keyword, 4-grams, and 5-grams.....	53
Table 3.4: An example shows document feature distribution with features T_1 , T_2 , T_3 , T_4 , and T_5 occurring in Title, Meta, and Body tags.....	55
Table 3.5: A feature frequency distribution of three documents D_1 , D_2 , and D_3 computed using Equations 3.1 and 3.2 with data from Table 3.4.....	55
Table 3.6: A feature frequency profile of document D_1 with data from Table 3.5....	55
Table 3.7: An example shows document dissimilarity matrix consisting of 5 documents D_1 , D_2 , D_3 , D_4 , and D_5 and the average of document dissimilarities computed by WCOND-Mine algorithm.....	63
Table 3.8: LOF-Mine computation with data from Table 3.7.....	65
Table 3.9: FindWDO computation with data from Table 3.7.....	68
Table 4.1: The number of outliers detected and accuracy computed from FindWDO algorithm over HTML dataset by using different number of neighbors at top-120 outliers' level.....	76
Table 4.2: The running time of FindWDO algorithm by using different number of neighbors and feature word over HTML dataset at top-120 outliers' level.....	77
Table 4.3: The number of outliers detected and accuracy computed from WCOND-Mine algorithm over HTML dataset at different top-n outliers.....	79
Table 4.4: The number of outliers detected and accuracy computed from LOF-Mine algorithm over HTML dataset at different top-n outliers.....	81
Table 4.5: The number of outliers detected and accuracy computed from FindWDO algorithm over HTML dataset at different top-n outliers.....	82
Table 4.6: The accuracy of WCOND-Mine, LOF-Mine, and FindWDO algorithms over HTML dataset at Top-120 outliers' level.....	84
Table 4.7: The number of outliers detected and accuracy computed from WCOND-Mine algorithm over Email dataset at different top-n outliers.....	86
Table 4.8: The number of outliers detected and accuracy computed from LOF-Mine algorithm over Email dataset at different top-n outliers.....	87
Table 4.9: The number of outliers detected and accuracy computed from FindWDO algorithm over Email dataset at different top-n outliers.....	89
Table 4.10: The accuracy of WCOND-Mine, LOF-Mine, and FindWDO algorithms over Email dataset at Top-100 outliers' level.....	91

Table 4.11: Showing six groups of HTML pages dataset, their sizes (MB), and number of embedded motifs..... 92

Table 4.12: The size of the general vector generated from six different groups of HTML page..... 92

Table 4.13: The running time of WCOND-Mine algorithm over six different groups of HTML pages..... 93

Table 4.14: The running time of LOF-Mine algorithm over six different groups of HTML pages..... 94

Table 4.15: The running time of FindWDO algorithm over six different groups of HTML pages..... 95

LIST OF FIGURES

Figure 2.1:	Taxonomy of outlier mining techniques.....	10
Figure 2.2:	A data distribution with 5 outliers (V, W, X, Y and Z). The data is adapted from the Win dataset (Blake and Merz, 1998).....	12
Figure 2.3:	A box plot of the y-axis values for the data in Figure 2.2 (Hodge and Austin, 2004).....	12
Figure 2.4:	A 2-dimensional space with one outlying observation (lower left corner).....	13
Figure 2.5:	First 20 depth contour for a dataset containing 5000 points (Johnson et al., 1998).....	16
Figure 2.6:	Depth contours computed by FDC for concentric data points (Johnson et al., 1998).....	17
Figure 2.7:	A 2-Dimensional dataset shows the idea of the density-based techniques to detect outliers (Breunig et al., 2000).....	23
Figure 2.8:	Three clusters with five outlier points (V, W, X, Y, and Z). The data is adapted from the Win dataset (Blake and Merz, 1998).....	26
Figure 3.1:	Block diagram of the proposed Web document outlier mining system...	43
Figure 3.2:	Overall structure of the proposed Web document outlier mining system	44
Figure 3.3:	General feature vector of document collection in Table 3.5.....	57
Figure 3.4:	General frequency vector of document collection in Table 3.5.....	57
Figure 3.5:	Feature Frequency Matrix (FFM) shows the frequency of feature per document.....	58
Figure 3.6:	Feature Weight Matrix (FWM) shows the weight of feature per document.....	60
Figure 3.7:	Document Dissimilarity Matrix (DDM) shows the pair-wise document dissimilarities.....	61
Figure 3.8:	The WCOND-Mine algorithm.....	62
Figure 3.9:	The LOF-Mine algorithm.....	66
Figure 3.10:	The proposed algorithm (FindWDO).....	69
Figure 4.1:	The accuracy of FindWDO algorithm over HTML dataset by using different number of neighbors at top-120 outliers' level.....	76
Figure 4.2:	The running time of FindWDO algorithm by using different number of neighbors with feature word over HTML dataset at top-120 outliers' level.....	78
Figure 4.3:	The accuracy of WCOND-Mine algorithm over HTML dataset at different top-n outliers.....	79
Figure 4.4:	The accuracy of WCOND-Mine algorithm over HTML dataset at different top-n outliers.....	80