شبكة المعلومات الجامعية

بسم الله الرحمن الرحيم

شبكة المعلومات الجامعية

جامعة عين شمس

التوثيق الالكتروني والميكروفيلم

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 الاصلية 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

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

TOWARDS
MINING WEB CONTENT OUTLIERS

By

Ayman Hassan Tanira

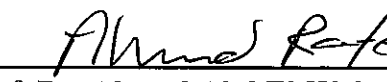
A Thesis Submitted to the
Faculty of Computers and Information
Cairo University
in Partial Fulfillment of the
Requirements for the Degree of

Master of Science

in

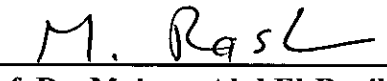
COMPUTER SCIENCE

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 mush 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

LIST OF TABLES

LIST OF FIGURES