

**Analyzing and Summarizing User Contributions
in Wikis**

by

Fong Kin Fong



Master of Science in Software Engineering

2013



**Faculty of Science and Technology
University of Macau**



Analyzing and Summarizing User Contributions in Wikis

by

Fong Kin Fong

A thesis submitted in partial fulfillment of the
requirements for the degree of

Master of Science in Software Engineering

Faculty of Science and Technology
University of Macau

2013

仁義禮知信

Approved by _____

Supervisor

Date _____



In presenting this thesis in partial fulfillment of the requirements for a Master's degree at the University of Macau, I agree that the Library and the Faculty of Science and Technology shall make its copies freely available for inspection. However, reproduction of this thesis for any purposes or by any means shall not be allowed without my written permission. Authorization is sought by contacting the author at

E-mail: peter.kf.fong@gmail.com

Signature _____

Date _____





University of Macau

Abstract

ANALYZING AND SUMMARIZING USER
CONTRIBUTIONS IN WIKIS

by Fong Kin Fong

Thesis Supervisor: Assistant Professor, Dr. Robert P. Biuk-Aghai
Master of Science in Software Engineering

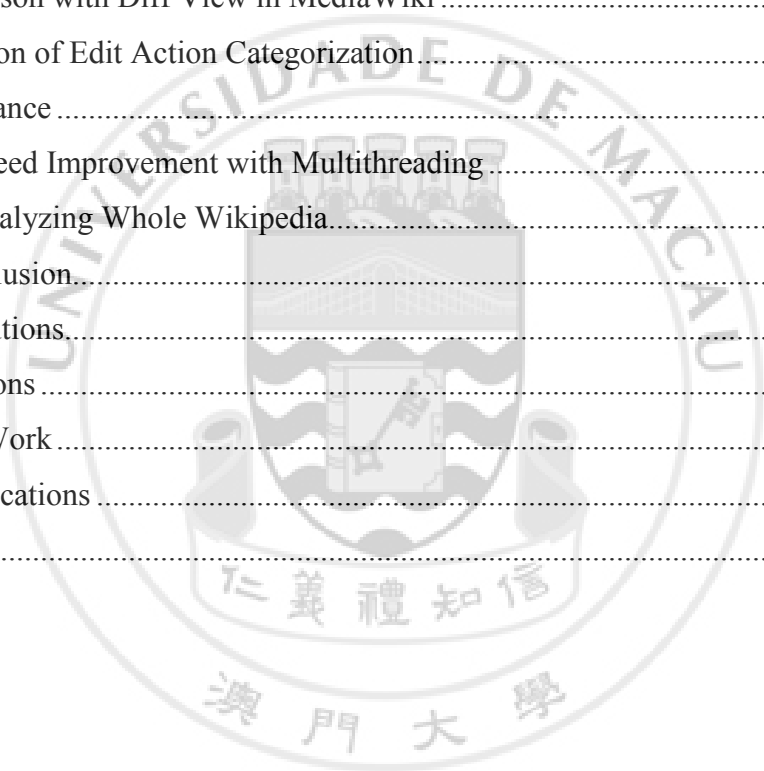
Wiki system, which allows users to collaboratively write articles online in a convenient way, has become popular in recent years. By design, wiki encourage a large number of users to make many edits rapidly. This characteristic makes it impractical to identify editor's contributions of an article by examining each edit manually. However, finding major contributors are, in many cases, useful for wiki community members. Therefore, a method to automatically identify and summarize user contribution can be helpful. In this thesis, we proposed a framework to programmatically analyze wiki article edits, present it in a way that is close to human understanding, and measure the significance of the edit using a metric. With the significance value of each edit on hand, we can then perform different kinds of aggregation to find out who is the major contributor of an article, what topic writers are interested, what type of contribution are made by them, and many others.



TABLE OF CONTENTS

List of Figures	iii
List of Tables	v
Glossary	vi
Chapter 1: Introduction	1
1.1. Wiki.....	2
1.1.1. Wikipedia.....	3
1.1.2. MediaWiki	4
1.2. Motivation.....	5
1.3. Research Problems.....	5
1.4. Solution Approach	6
1.5. Outline of this Thesis.....	7
Chapter 2: Literature Review.....	8
2.1. Wiki Edit Analysis Methods.....	8
2.1.1. Differencing Algorithms.....	8
2.1.2. Edit Categorization	9
2.1.3. Edit Significance Calculation	10
2.2. Wikipedia Analysis and Visualization.....	11
2.2.1. Revision Analysis and Visualization	11
2.2.2. Category Analysis and Visualization.....	15
2.2.3. Recent Changes Analysis and Visualization.....	18
Chapter 3: Wiki Edit History Analyzer	20
3.1. Analysis Method	20
3.1.1. Lexical Analyzer.....	21
3.1.2. Differencing Engine.....	25
3.1.3. Edit Action Categorizer	34
3.1.4. History Summarizer	41
3.2. Edit Significance	41
Chapter 4: Prototype Implementation.....	44

4.1. Prototype Analyzer	44
4.2. Mediawiki Extensions.....	45
4.2.1. History View	45
4.2.2. Diff View	46
4.2.3. Article Statistics	48
4.2.4. Editor Statistics	49
4.3. Batch Processing.....	50
Chapter 5: Case Study on Wikipedia.....	52
5.1. Comparison with Diff View in MediaWiki	52
5.2. Evaluation of Edit Action Categorization.....	54
5.3. Performance	56
5.3.1. Speed Improvement with Multithreading.....	57
5.3.2. Analyzing Whole Wikipedia.....	58
Chapter 6: Conclusion.....	61
6.1. Contributions.....	62
6.2. Limitations	63
6.3. Future Work.....	64
Associated Publications	66
Bibliography	67



LIST OF FIGURES

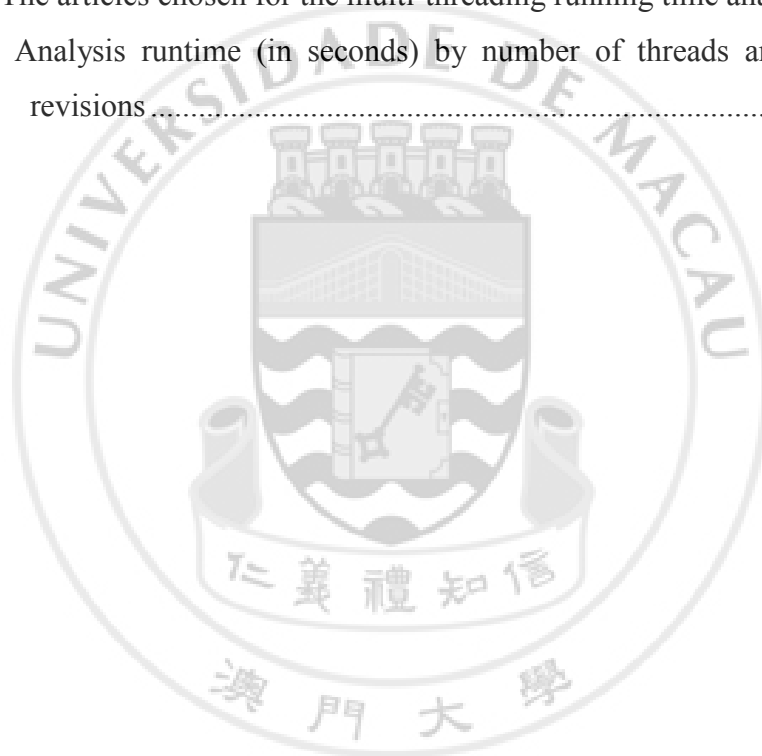
<i>Number</i>	<i>Page</i>
Figure 1: WikiDashboard on English Wikipedia article “United States presidential election, 2008” (reproduced from [28])	12
Figure 2: The Chromograms application, block view (reproduced from [29])	13
Figure 3: History flow for article “Abortion” in English Wikipedia, spaced by date (reproduced from [30])	14
Figure 4: Degree of co-authorship output calculated by CoAuthor MediaWiki extension (reproduced from [4])	14
Figure 5: Wikipedia category visualization (reproduced from [31])	16
Figure 6: Overview map of categories of Simple English Wikipedia (reproduced from [32])	17
Figure 7: Radial visualization of categories in English Wikipedia (reproduced from [33])	17
Figure 8: Recent change visualization for English Wikipedia, edit pattern chart (reproduced from [34])	19
Figure 9: Conceptual structure of Wiki Edit History Analyzer	20
Figure 10: Three level of lexical analysis and differencing	22
Figure 11: Edit significance bar in MediaWiki “history” view	46
Figure 12: Diff view of a revision of the article “City of Manchester Stadium”	47
Figure 13: Edit analysis statistics page for article “Jupiter”	48
Figure 14: Contribution from user “Eptalon”	50
Figure 15: Difference page of article “Film noir” produced by MediaWiki	53
Figure 16: Difference page of article “Film noir” produced by our edit history analyzer	54
Figure 17: Relationship between number of analyzer threads and total analysis time	59
Figure 18: Relationship between running time and number of revisions	59

Figure 19: Relationship between article length and average running time per
revision.....60



LIST OF TABLES

<i>Number</i>	<i>Page</i>
Table 1: Comparison of wiki markups and HTML	3
Table 2: Some basic wiki markup in MediaWiki	24
Table 3: Common content types in MediaWiki	37
Table 4: Article chosen for evaluation	55
Table 5: The articles chosen for the multi-threading running time analysis.....	57
Table 6: Analysis runtime (in seconds) by number of threads and number of revisions	58



GLOSSARY

Article: A content page in a wiki.

Contributor: A person who edits an article in a wiki.

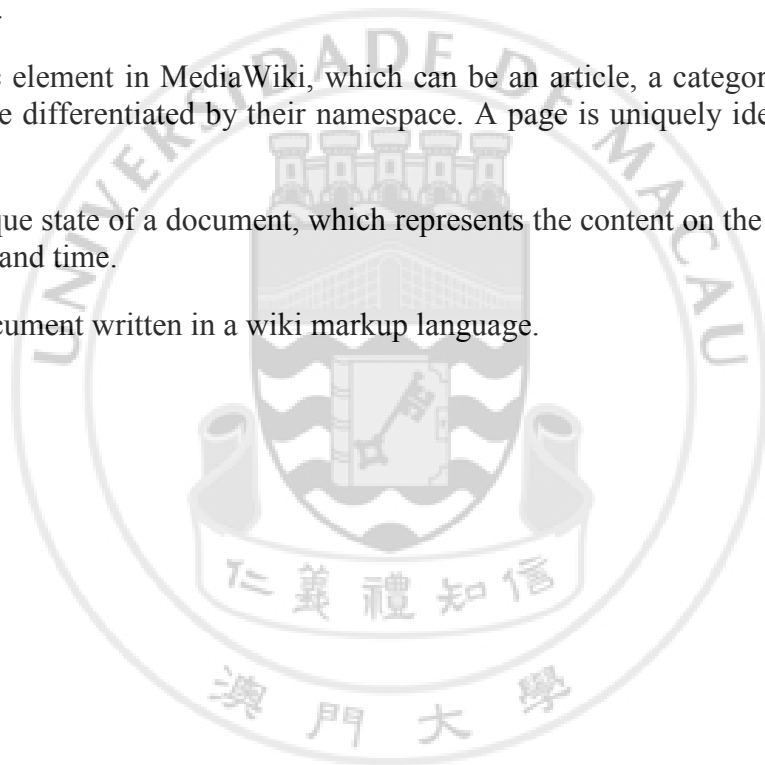
Edit: Modification of the content in a page, which creates a new version of the page.

Namespace: Namespaces are used in wikis to separate content by purpose. For example, articles are stored in the main namespace, while discussions are stored in the Talk namespace.

Page: The basic element in MediaWiki, which can be an article, a category or other object, which are differentiated by their namespace. A page is uniquely identified by its page ID.

Version: A unique state of a document, which represents the content on the document at a certain date and time.

Wikitext: A document written in a wiki markup language.



ACKNOWLEDGMENTS

I would like to express my most sincere gratitude to my supervisor, Dr. Robert P. Biuk-Aghai, who kept guiding me and encouraging me along the way of this research. With his great vision and experience, he helped me to find my research direction, inspired me with different new ideas, pointed the way to solve the problems and suggested numerous possible improvements. It is impossible for me to complete this thesis without his leadership and experience. Moreover, he has put a lot of effort in helping me to improve my academic writing skills. I truly appreciate his effort, patience and help during these years of supervision.

I would also like to thank teachers and students, both former and current, in the Faculty of Science and Technology, especially Mr. Miguel Gomes da Costa Junior and Mr. Pang Cheong-lao, Patrick, for generously sharing their knowledge and experience related to this research work.

Last but not least, I would like to thank my parents and all my friends for their support and encouragement on both my research and my life.