

Map-like Wikipedia Visualization

by

Pang Cheong Iao



Master of Science in Software Engineering

2011



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

Map-like Wikipedia Visualization

by

Pang Cheong Iao

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

2011



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: inbox@patrickpang.net



Signature _____

Date _____

University of Macau

Abstract

MAP-LIKE WIKIPEDIA VISUALIZATION

by Pang Cheong Iao

Thesis Supervisor: Assistant Professor, Dr. Robert P. Biuk-Aghai

Master of Science in Software Engineering

Wikipedia, the largest online collaboratively authored encyclopedia, allows anyone to easily contribute to its content. As it operates in a collaboratively authoring model, authors are not only responsible for writing up articles, but also assigning articles into categories in a way to classify by their topics. This makes the category system, which contains topic classifications under authors preferences is worth studying. However, given its large data volume and idiosyncrasies of its data structure, the analysis of Wikipedia's category data is challenging. To facilitate ease and efficiency of understanding, we have designed a new visualization – a map-like visualization representing Wikipedia's category data in a form similar to a geographic map. It allows readers to intuitively perceive large-scale patterns of categories and articles in the Wikipedia. In this thesis, we describe the algorithm of creating this novel type of visualization, and introduce our approaches for tackling the difficulties of handling the Wikipedia data.

TABLE OF CONTENTS

List of Figures	3
List of Tables	5
Glossary	6
Acknowledgments.....	7
Dedication	8
Chapter 1: Introduction	9
1.1 Collaborative Authoring and Wikis	9
1.2 Wikipedia and MediaWiki	11
1.3 Research Problems	12
1.4 Solution Approach	13
1.5 Outline of this Thesis	14
Chapter 2: Literature Review	15
2.1 Information Visualization	15
2.2 Wikipedia Data Processing	19
2.3 Wikipedia Visualization.....	20
2.4 Map-like Visualization.....	26
2.5 Force-directed Layout Algorithms.....	27
2.6 Overlap Removal Algorithms	28
Chapter 3: Map-like Visualization.....	29
3.1 Overview of the Visualization	29
3.2 Data Source Requirements.....	30
3.3 Sorting Algorithm by Similarities.....	31
3.4 Preliminary Layout	32
3.5 Hexagonal Visualization	35
3.5.1 The Hexagon Canvas	35
3.5.2 Selection of Hexagons	37
3.5.3 Clutter Reduction.....	38
3.5.4 Coloring Scheme.....	40

3.5.5 Text Labeling	41
Chapter 4: Wikipedia data analysis.....	43
4.1 Database Schema of MediaWiki.....	43
4.2 Obtaining Wikipedia Data	45
4.3 Transforming Category Graph.....	46
4.3.1 Choosing Semantic Root.....	46
4.3.2 Removing Non-Content Categories.....	47
4.3.3 Building a Category Tree.....	49
4.4 Category Relationships	50
Chapter 5: Case Study on Wikipedia.....	55
5.1 Performance	55
5.2 Discussion on Applications.....	56
5.2.1 Understanding Category Composition.....	57
5.2.2 Displaying Category Relationship	59
5.2.3 Overview of Multiple Wikipedias	60
5.2.4 Comparison of Topic Areas in Multiple Wikipedias.....	61
5.3 Compare with Other Wikipedia Visualizations	62
5.3.1 Comparison with Semantic Coverage Visualization	63
5.3.2 Comparison with Matrix Visualization.....	64
5.3.3 Comparison with Radial Visualization.....	65
5.3.4 User Evaluation.....	67
Chapter 6: Conclusion.....	70
6.1 Contribution	71
6.2 Future Work	71
Associated Publications	73
References.....	74
Appendix A: Results of Wikipedia Map-like Visualization.....	78
Appendix B: Importing Wikipedia Database Dumps	83
VITA.....	84

LIST OF FIGURES

<i>Number</i>	<i>Page</i>
Figure 1: Treemap visualization of a folder in a file system (reproduced from [10])	15
Figure 2: A treemap with one million of items (reproduced from [11]).....	16
Figure 3: Two representations of a tree: (a) the radial layout (b) the balloon layout (reproduced from [13])	17
Figure 4: Connections added to: (a) a radial tree (b) a balloon tree; both visualizing function calls in a software (caller in green and callee in red) (excerpt from [14]).....	17
Figure 5: The screenshot of SpaceTree (reproduced from [17]).....	18
Figure 6: Haber and McNabb’s visualization model	19
Figure 7: Five levels deep, centered on Politics	21
Figure 8: Close up of Chris Harrison’s WikiViz output	21
Figure 9: Entity view with labeled nodes in WikiVis (reproduced from [23]).....	22
Figure 10: WikiVis category view (reproduced from [23]).....	22
Figure 11: Wikipedia category visualization (reproduced from [7])	23
Figure 12: The Chromograms application: showing users’ activity in blocks (reproduced from [24])	24
Figure 13: History flow visualization (reproduced from IBM Research).....	25
Figure 14: Author text analysis of a section of text with high depth of collaboration (reproduced from [26])	25
Figure 15: Edit significance bars of revisions of an article (reproduced from [27])	25
Figure 16: Summary of different type of edit behaviors (reproduced from [27])	26
Figure 17: Visualization of a 10-year period by cartographic means (reproduced from [28]).....	27
Figure 18: Comparison of overlap removal algorithms (reproduced from [32]).....	28
Figure 19: The bottom-up layout approach	32
Figure 20: Approaches for radial placement of sorted entities	34
Figure 21: Hexagons and their coordinates in the memory	35

Figure 22: Arithmetic of a hexagon	36
Figure 23: Example of hexagon selection.....	37
Figure 24: Before and after clutter reduction.....	39
Figure 25: A topographic map uses colors for showing elevation.....	40
Figure 26: Legend of a map-like visualization	40
Figure 27: Rotation of text labels in a region	41
Figure 28: Result of the text label placement	42
Figure 29: Eliminating edges in the graph (edge indicates “parent-to-child” relationship)	50
Figure 30: Comparison of different experimental methods for aggregating similarity	52
Figure 31: The current way to obtain category distribution information.....	58
Figure 32: Highlight of category “Mathematics” in the Wikipedia.....	58
Figure 33: Cluster of related categories “Environment”, “Life” and “Geography” placed in close proximity of one another	59
Figure 34: Category “Science” in Wikipedia: (a) Danish (b) Swedish (c) Chinese	62
Figure 35: Highlight of Holloway’s Wikipedia visualization.....	63
Figure 36: A multi-level matrix visualization of category hierarchies	64
Figure 37: A radial visualization of Wikipedia (reproduced from [37])	66
Figure 38: Extract of the radial visualization near the edge of the circle (reproduced from [37])	67
Figure 39: Overview map of the Simple English Wikipedia.....	78
Figure 40: Overview map of the Danish Wikipedia	79
Figure 41: Overview map of the Chinese Wikipedia.....	79
Figure 42: Overview map of the Swedish Wikipedia.....	80
Figure 43: Overview map of the German Wikipedia.....	81
Figure 44: Overview map of the English Wikipedia	82

LIST OF TABLES

<i>Number</i>	<i>Page</i>
Table 1: Comparison of wikitext and HTML	11
Table 2: Examples of hierarchical data (number in the parentheses represents the size)	30
Table 3: Sorting algorithm with similarity pairs	32
Table 4: Algorithm for randomly selecting hexagons	38
Table 5: Namespaces in MediaWiki	44
Table 6: Definition of the “categorylinks” table	44
Table 7: Sizes of Wikipedias	45
Table 8: Top content categories in different Wikipedia language editions	46
Table 9: Comparison of name similarities under category “Aircraft 1950-1959” and “Computer science”	49
Table 10: Cosine similarity for categories in the English Wikipedia with data from different levels of sub-categories	51
Table 11: Standard deviations of similarity aggregation experimental expressions.....	53
Table 12: Aggregated similarity of category “Mathematics” in the Wikipedia	54
Table 13: Statistics of visualization output.....	56
Table 14: Statistics of the German and English Wikipedia	60
Table 15: Statistics of category “Science” in the Danish, Swedish and Chinese Wikipedia	62
Table 16: Responses of the user evaluation	68
Table 17: SQL commands for importing database dumps.....	83

GLOSSARY

Article: An encyclopedia article in the Wikipedia.

Category: A classification defined in the Wikipedia that groups articles with relevant information.

Category Assignment: An article is assigned to a category, and becomes part of the content of that category.

Category Relationship: A measurement of relatedness of categories. If contents within one category are more similar to the contents of another category, then we define they have a closer relationship or vice versa.

Co-assignment of Categories: If an article is assigned with multiple categories at the same time, the article is co-assigned with these categories.

Cosine Similarity: Cosine similarity is a measure of similarity between two vectors by measuring the cosine of the angle between them. It is widely used in measuring similarities of documents and strings.

Directed Acyclic Graph: A directed acyclic graph is a directed graph with no directed cycles. For example: for any starting point of vertex v , there is no way to follow a sequence of edges that eventually loops back to v again.

Namespace: Namespaces are used in the Wikipedia to separate contents for different purposes, such as articles and categories are stored under Main and Category namespace respectively for public access, while Talk namespace consists of discussions that intended for the authoring community. In this way, different types of users are able to concentrate on the data which intended for their use.

Page: The basic element in the Wikipedia, which can be an encyclopedia article, a category or other object that differentiate by its namespace. A page is uniquely identified by its page ID.

Sub-category: A category that is assigned under an upper level (parent) category. Usually sub-categories are defined as more detailed topics of their parent's topic.

Top Categories: The highest level of categories for classifying articles by their topics, such as Science, People, History, etc.

ACKNOWLEDGMENTS

I would like to thank my thesis supervisor, Dr. Robert P. Biuk-Aghai, who kept guiding and encouraging me along the way of this research. He inspired me with different new ideas, points to improve and possible solutions to problems. Without his leadership and experience, it is impossible for me to complete this thesis. Moreover, he paid so much effort in helping me improve the academic writing skills as well as publish publications. I truly appreciate his effort and help during these three years of supervision.

I am also grateful to my colleagues in the University of Macau, especially my unit head Steve Lai, team leaders Jack Wu and Christina Lou and my project manager Evi Ian, who helped me in my work and coordinated the workload so that I could successfully finish this degree.

I would also thank the former and current students of the Faculty of Science and Technology, such as Libby Tang, Henry Leong, Alex Ieong, Kin Lei and Peter Fong, for sharing their experiences and knowledge related to this research.

I gratefully acknowledge the support for the publications of this research from the Macau Special Administrative Region – Science and Technology Development Fund (grant number 021/2011/A).

Last but not least, I will never forget the support and encouragement from my parents and all my friends on both my research and my life.

DEDICATION

I would like to dedicate this thesis to my parents, who continuously support my studies and my life all the time.

