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

Number *Page*

LIST OF TABLES

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 the 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.