

MASTER'S THESIS

Commentary-based social media clustering with concept and social network discovery

Leung, Kwan Wai

Date of Award:
2011

[Link to publication](#)

General rights

Copyright and intellectual property rights for the publications made accessible in HKBU Scholars are retained by the authors and/or other copyright owners. In addition to the restrictions prescribed by the Copyright Ordinance of Hong Kong, all users and readers must also observe the following terms of use:

- Users may download and print one copy of any publication from HKBU Scholars for the purpose of private study or research
- Users cannot further distribute the material or use it for any profit-making activity or commercial gain
- To share publications in HKBU Scholars with others, users are welcome to freely distribute the permanent URL assigned to the publication

**Commentary-based Social Media Clustering with
Concept and Social Network Discovery**

LEUNG Kwan Wai

**A thesis submitted in partial fulfillment of the requirements
for the degree of
Master of Philosophy**

Principal Supervisor: Dr. C. H. Li

Hong Kong Baptist University

August 2011

Abstract

Due to the huge amount of videos in social media sites, categorizing videos with similar contents can help users to search videos more efficiently. We propose to facilitate searching of video in social media site with clustering from commentary-based matrix factorization and to improve indexing via the generation of new concept words. Videos are clustered by modified multi-assignment NMF clustering. Factorized component entropies are introduced for handling the difficult problem of vocabulary construction for concept discovery in social media. Since the categorization is learnt from user feedback, it can accurately represent the user sentiment on the videos. Experiments conducted using empirical data collected from YouTube shows the effectiveness of our proposed methodologies. Comparing with popular tag-based method, our proposed methodologies are shown to be more robust.

Besides concept discovery of online videos, user comments are also a good source for discovering relationships of celebrities for social network studies. However, the evolutionary characteristic and the daunting complexity of the interrelationship among singers made the problem technically intriguing. In this thesis, we present a novel commentary-based social network analysis (CBSNA) methodology to analyze the relationships between singers which is based on user comments on YouTube videos. Commentary-based singer social network is compared to tag-based network which is in a common network construction approach. Weighting schemes are developed for handling noisy social network.

To study and analyze a social network, visualizing the network is an important step. However, due to the tremendous growth of social network, the resulting social network graphs (SNGs) are always massive in size. Moreover, huge amount of uncleaned data col-

lected from the public contains lots of noise that would affect the effectiveness of further analysis. In the visualization of large social networks, it is popular to construct binary social network based on predetermine values of distance between nodes. In this thesis, we study different kinds of network building methods, including edge-cut with GeoDeg, MaxDir and MinDir normalization, and propose to construct a simplified network for presentation which the network is functionally equivalent to the original network, retains the main properties of nodes and democratically that every node can has its important relationships presented. Weighted-cut and authority scores are used as a measurement for evaluating the performance of different network presentation methods. Apart from SNGs, our proposed approach is also applicable on any types of graph with small-world property.

Table of Contents

Declaration	i
Abstract	ii
Acknowledgements	iv
Table of Contents	v
List of Figures	viii
List of Tables	ix
Chapter 1 Introduction	1
1.1 Social Media	1
1.2 Social Network	1
1.3 Social Network Visualization	2
1.4 Organization Roadmap	2
1.5 Chapter Outline	3
Chapter 2 Concept Discovery in Youtube.com Using Factorization Method	4
2.1 Overview	4
2.2 Related Works	6
2.3 Public Attention Based Video Concept Discovery and Categorization for Video Searching	8
2.4 Dataset collection	10

2.5	Data Pre-processing	11
2.5.1	Data Cleaning	11
2.5.2	Text Matrix Generation	12
2.6	Video Processing via Clustering	12
2.6.1	Video Clustering and Concept Discovery	14
2.6.2	Factorized Component Entropy Measures for Vocabulary Construction	17
2.7	Experimental Evaluation	20
2.7.1	Empirical Setting	20
2.7.2	Video Categories and Concepts	22
2.7.3	User Comments vs User Tags	27
Chapter 3 Commentary-based Hong Kong Singer Social Network		32
3.1	Overview	32
3.2	Related Works	33
3.3	Commentary-based Singer Social Network Discovery	34
3.4	Network Construction	35
Chapter 4 Visualization of Social and Other Small World Networks		38
4.1	Overview	38
4.2	Related Works	39
4.3	Dense Network Compression	40
4.3.1	k-Nearest Neighbor approach	40
4.3.2	Edge-cut approach	41
4.3.3	Weight Normalization	42
4.3.4	Comparisons	43
4.3.5	Hubs and Authorities	47
Chapter 5 Singer Network for Analysis		52
5.1	Singer Network Visualization	52
5.2	Network Weighting Schemes	60

5.2.1	Binary Weighting	60
5.2.2	Logarithm Weighting	60
5.3	Network Evaluation	61
5.4	Commentary-based Vs Tag-based	63
Chapter 6 Conclusion and Future Work		79
Bibliography		82
Publications		88
Curriculum Vitae		89