MASTER'S THESIS

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

Leung, Kwan Wai

Date of Award:
2011

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 evolutional 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