Accepted Manuscript

Multimedia Venue Semantic Modeling based on Multimodal Data

Wei-Zhi Nie, Wen-Juan Peng, Xiang-yu Wang, Yi-liang Zhao, Yu-Ting Su

PII: S1047-3203(16)30244-9
DOI: http://dx.doi.org/10.1016/j.jvcir.2016.11.015
Reference: YJVCI 1900

To appear in: *J. Vis. Commun. Image R.*

Accepted Date: 22 November 2016

Please cite this article as: W-Z. Nie, W-J. Peng, X-y. Wang, Y-l. Zhao, Y-T. Su, Multimedia Venue Semantic Modeling based on Multimodal Data, *J. Vis. Commun. Image R.* (2016), doi: http://dx.doi.org/10.1016/j.jvcir.2016.11.015

Multimedia Venue Semantic Modeling based on Multimodal Data

Wei-Zhi Nie, Wen-Juan Peng, Xiang-yu Wang, Yi-liang Zhao, Yu-Ting Su,

***Abstract*—A huge amount of text and multimedia (images and videos) data concerning venues is constantly being generated. To model the semantics of these venues, it is essential to analyze both text and multimedia user-generated content (UGC) in an integral manner. This task, however, is difficult for location-based social networks (LBSNs) because their text and multimedia UGCs tend to be uncorrelated. In this paper, we propose a novel multimedia location topic modeling approach to address this problem. We first utilize Recurrent Convolutional Networks to build the correlation between multimedia UGCs and text. Then, a graph model is structured according to these correlations. Next, we employ a graph clustering method to detect the latent multimedia topics for each venue. Based on the obtained venue semantics, we propose techniques to model multimedia location topics and perform semantic-based location summarization, venue prediction and image description. Extensive experiments are conducted on a cross-platform dataset, and the promising results demonstrate the superiority of the proposed method.**

***Index Terms*—Location-based, multimedia modeling, Graph Clustering, Image Description.**

I. Introduction

In the era of Web 2.0, the rapid development of location-based social networks (LBSNs), such as Foursquare[1], Gowalla[2], and Jiepang[3], has resulted in a huge amount of location-oriented user-generated contents (UGCs). When users physically visit certain venues, they typically check in at these places to let the world know of their current presence. They may also share location-related contents such as textual descriptions or images. In addition to traditional user textual descriptions and comments, images are becoming increasingly popular in LBSNs because it is more convenient to simply shoot and share an image than a short text when using location-enabled smart phones [1]. The huge and fast-growing amounts of location-oriented UGC have given rise to new opportunities and challenges for the effective searching, browsing, and organization of these multimedia contents.

There are numerous UGCs for different venues in LBSNs. These UGCs describe different aspects of the venues. It will be useful to extract venue semantics from these UGCs to better understand the functions and activities of these venues. Given the semantics of different venues, we can then measure the semantic similarity between different venues and use this similarity as a basis to recommend venues with similar semantics to users. Thus, the extraction of location semantics from UGCs will be beneficial to various location-based services.

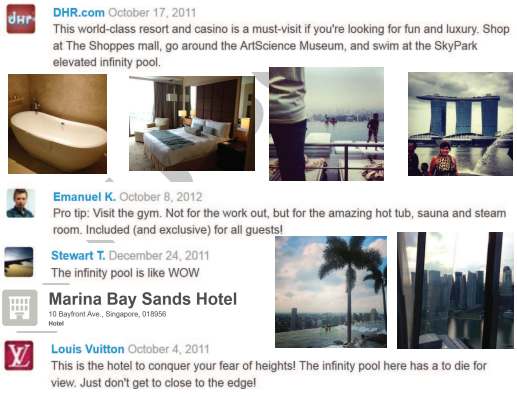

Fig. 1. Examples of UGCs at Marina Bay Sands Hotel in Singapore

A. Motivation

In recent years, many algorithms are proposed to utilize UGCs to handle some classic problem such as image annotation[2], multimodal feature learning[3], venue recommendation[4], event detection[5], etc. However, UGCs are unstructured and usually include different modalities and various topics. For example, Figure 1 presents a sampled subset of UGCs from the Marina Bay Sands Hotel in Singapore, contributed by users on Foursquare. The UGCs contain both texts and images and cover various topics. However, these systems generally do not provide annotation facilities for users to annotate their shared images, which makes it difficult to analyze the multimedia content of these images. To address this problem, we propose a *multimedia location topic model* to discover and annotate venue related UGCs using diverse and representative thematic information. In contrast to traditional text-based topic models, such as LDA [6], [7], [8], the *multimedia location topic model* addresses not only texts but also visual information.

Many classic methods [9], [10], [11], [12] have been proposed to address this problem. An intuitive solution to modeling multimedia topics from heterogeneous location-oriented UGCs is to directly fuse them together ([9], [12]).

Wei-Zhi Nie, Wen-Juan Peng and Yu-Ting Su are with School of Electronic Information Engineering, Tianjin University, Tianjin, 300072, China. (Email: weizhinie@tju.edu.cn)

Xiang-yu Wang and Yi-liang Zhao are with School of Computing, National University of Singapore.

[1]http://www.foursquare.com/

[2]http://www.gowalla.com/

[3]http://www.jiepang.com/