J. Vis. Commun. Image R. 40 (2016) 838–846

Contents lists available at ScienceDirect

J. Vis. Commun. Image R.

journal homepage: www.elsevier.com/locate/jvci

Quality biased multimedia data retrieval in microblogs ☆

Shuhan Qi [a], Peiguang Jing [b], Xuan Wang [a,*], Liqiang Nie [c]

[a] Computer Application Research Center, ShenZhen Graduate School, Harbin Institute of Technology, ShenZhen, China
[b] School of Electronic Information Engineering, Tianjin University, Tianjin, China
[c] School of Computing, National University of Singapore, Singapore

ARTICLE INFO

Article history:
Received 15 April 2016
Revised 14 July 2016
Accepted 20 August 2016
Available online 26 August 2016

2010 MSC:
00-01
99-00

Keywords:
Microblog retrieval
Quality model
Multiview embedding

ABSTRACT

With the rapid development of social media platforms, huge amount of user generated contents (UGC) are generated ceaselessly. In recent years, content based microblog retrieval has attracted extensive research attention. Effective microblog retrieval services complex analysis of short text and multimedia contents. In this paper, we present a quality biased multimedia microblog retrieval framework. First, we develop an anchor graph based multiview embedding framework which maps the multimedia content features into a unified latent space. Then, the content matching scores of testing microblogs related to the query are obtained by a Markov random field. Further, we employ an quality model to incorporate both microblog quality and content matching. As compared with the state-of-art methods, experimental results demonstrate the effectiveness of the proposed approach.

1. Introduction

Social media platforms [1,2], such as Facebook, Twitter and Sina Weibo, have provided services for consumers to generate a large amount of user-generated-contents (UGC) and social media stream data. An example is the Super Bowl 2013, which attracts up to 24 million tweets in total and the number of tweets about just the blackout is over 231K per minute. Various user-shared hot contents may spread quickly and widely across the entire social network. The large amount of UGC data not only offers important data resources for a wide range application, but also requires the technology for targeting the data that users really needed. For example, organizations such as enterprises and governments have the requirement of obtaining the social media data that related to some entities (like brand, event and person, etc.) [3,4]; in this way these organizations can evaluate the effects of marketing strategy or the influence of these entities. Even for individual users, searching for useful entity-related information on social media platform helps them to make decisions. Therefore, identifying microblogs that relate to specific entities in social media data is very significant.

Content based microblog identification in social media has attracted extensive research attention in recent years [5–7]. Different from those methods that mainly focus on microblog classification in the social media streams, in this paper, we concentrate on the entity-related microblog retrieval task. Specifically, in the social media data streams, given some predefined entity-related texts, tags or images as seeds, the aim is taking these seeds as queries to retrieve the post data which have relevance with the predefined entity. Most of existing microblog retrieval models only take time and the textual content as the evidence of relevance, while the multimedia contents of social media such as videos and images which are indispensable for social media analysis. It is essential to further explore the role of visual contents for microblog retrieval. As compared with traditional search tasks, the microblog retrieval problem has many challenges. First, due to the short and conversation content of social posts, textual contents sometimes cannot provide enough information for social analysis. Besides, the topics and keywords may change rapidly as time shifts. Under this situation, the traditional methods which keep a fixed keyword dictionary are not adequate for searching the representative data. Second, the contents in social media have rapidly shifted from text to multimedia such as image and vedio. Users often use multi-media contents to express their opinions or feelings. Moreover, the meanings of different modalities content sometimes irrelevant or even incompatible with each other. For example, according to our recent investigation, more than 40% Sina

☆ This paper has been recommended for acceptance by Zicheng Liu.

* Corresponding author.
E-mail address: wangxuan@cs.hitsz.edu.cn (X. Wang).

http://dx.doi.org/10.1016/j.jvcir.2016.08.015
1047-3203/

Weibo microblogs[1] are posted with images, but only a small proportion (about 30%) of these microblogs have compatible textual contents.

The social media platforms encourage various forms of contents such as news reports, personal updates, babbles, conversations, etc. In this case, to improve the microblog retrieval, we need to measure the informative microblogs and take into account the quality of the microblog contents. Quality model begins to attract researchers' attention in recent years. There is a variety of new development of quality model, especially in pagerank [8], web searching [9], and biologically inspired media quality modeling [10,11]. Some recent works show that the deterministic quality-biased ranking model improves the performance of information retrieval [12,13]. Most of these content based quality models only consider the textural content of microblog, while ignore the visual content.

In this paper, we propose a quality biased multimedia microblog retrieval approach. The main framework of the proposed method is shown in Fig. 1. There are large scale multimedia contents in social media. To make an unified feature representation of the data and compute the matching score between queries and testing data, we propose an anchor graph based multiview feature embedding model. By looking for the complementary relationships in different feature spaces, the multiview features are mapped into a unified latent subspace. The new feature space not only maintains the information in the original feature space but also be more discriminative. Meanwhile, we employ anchor graph to handle the high complexity problem of solving the large scale multiple graph. After extracting the features in the new latent space, content matching scores between queries and testing data are inferred by a Markov random field. Further, we also propose a quality biased model as a complement of the content matching. By using some quality features that are extracted from the textual contents and visual contents, the quality biased model improves the performance of the retrieval results. The effectiveness of proposed ranking method has been evaluated on the Brand-Social-Net dataset [7]. The contribution of this paper is in three fold: first, a Quality-based Microblog retrieval framework is proposed, in this framework, multiply content quality factors of microblog are utilized to retrieve related microblogs from dataset. Second, a multi-graph based multi-view embedding model is obtained by unsupervised learning. The multi-view embedding model is able to map the multiple microblog features into a optimal unified latent space; third, to solve the problems of large storage requirement and extensive computation in multi-view embedding model, an anchor graph based efficient graph construction method is proposed.

2. Related work

In this section, we briefly introduce the related work on social media analysis and quality model.

2.1. Social media analysis

In this section, we briefly review the related work on data analysis in social media platforms. Allan et al. [14] proposed a method that uses a tf-idf variant and a time-based threshold to measure the relevance between events and documents. In [15], the microblog identification problem is treated as a clustering problem, and a similarity learning metric is introduced to identify the distance between events and social media streams. Weerlamp et al. proposed a two-level textual credibility target describing approach, in which the two-level are the single post level and the blog level [16]. Wang et al. [17] developd a social event detection method that detects social, physical-world events from photos in social media sites. In addition to the metadata such as time, locations, tags and descriptions, they incorporated online social interaction features in the detection of physical-events.

For entity-related microblog analyzing in social media. In [6], a microblog identification framework was proposed, which includes an offline relevance detection step and an online rectification step for topic-related data gathering and noisy data filtering. Gao et al. [7] proposed a multi-faceted brand tracking method for data gathering and content analysis in social media streams. The proposed method gathers relevant data base on keywords, visual contents and social factors. Then they extended their work to brand-related social events detection in [5]. To handle the noisy and short microblogs, they proposed an intermediate semantic entity named microblog clique which explores the correlated information among microblogs. Base on the microblog cliques, the social events are detected by a bipartite graph.

2.2. Quality model

Content based quality model for information retrieval has been studied in a variety of settings. Bendersky et al. [9] proposed a quality-biased ranking method which is based on the features of the web content, readability, and other web quality evidences, for web documents search. Huang et al. [12], built a linear regression model that includes a quality regularization factor for ranking the short microblogging documents. Jaeho et al. suggested a low-cost quality model using surrogate judgments based on user behaviors [13].

Most of the content based quality models for information retrieval are based on textual contents. To evaluate the quality of multimedia content of social media, Nie et al. [18] proposed a quality model for venue recommendation by utilizing multimedia data to predict venue interestingness in location based service (LBS) network. Hu and Nie [19] proposed discriminative image quality measurement based framework for image recognition.

Recently, many multi-task dictionary learning models are applied in multimedia and computer vision. They can improve the performance of learning algorithms by learning a problem together with other related problems at the same time, using a shared representation. Besides, these methods can be used to learn high-order potential descriptors for multimedia data [20,21]. Further, multi-task dictionary learning can be used as a generate framework to improve multimedia classification [22,23], recognition [24], retrieval [25,26] and clustering [27].

3. Scalable microblog multiview embedding

Suppose that there is a data set $\chi = \{\chi_1, \chi_2, \ldots, \chi_N\}$ with N microblog. For each microblog there are K types of feature with different modalities. Different modality features are obtained from different perspectives (such as the textual content feature and visual content feature of microblog), which have different physical meanings and statistical properties. Motivated by the manifold regularization methods [28,29], in this section, we proposed a manifold regularization based multiview embedding method, which generates a unified microblog feature representation by encoding the correlation among different modality features.

3.1. Manifold regularization based multiview embedding

Given the K modalities of feature for each sample in the dataset, i.e., $x_i = \{x_i^{(k)}\}$, $x_i^{(k)} \in R^{m \times k}$, where $x_i^{(k)}$ is the feature of k-th modality.

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