Searching for Recent Celebrity Images in Microblog Platform

Na Zhao†, Richang Hong†, Meng Wang†, Xuegang Hu†, Tat-Seng Chua‡
† School of Computer and Information, Hefei University of Technology
‡ School of Computing, National University of Singapore
zhaona311@gmail.com, hongrc@hfut.edu.cn, eric.mengwang@gmail.com,
jsjxhuxg@hfut.edu.cn, chuats@comp.nus.edu.sg

ABSTRACT

With the explosive growth and widespread accessibility of image content in social media, many users are eagerly searching for most recent and relevant images on topics of their interests. However, most current microblog platforms merely make use of textual information, specifically, keywords, for image search which cannot achieve satisfactory results since in most cases image content is inconsistent with textual content. In this paper we tackle this problem under the application of searching for celebrity image. The proposed method is based on the idea of refining the initial text-based search results by utilizing multimedia plus social information. Given a text search query, we first obtain an initial text-based result. Next, we extract a seed tweet set whose images contain faces recognized as celebrities and texts contain the expanded keywords. Third, we extend the seed set based on visual and user information. Lastly, we employ a multi-modal graph based learning method to properly rank the obtained tweets by integrating social and visual information. Extensive experiments on data collected from Tencent Weibo demonstrate that our proposed method could approximately achieve 3-fold improvement in results as compared to the text baseline, typically used in microblog search service.

Categories and Subject Descriptors

H.3.3 [**Information Search and Retrieval**]: Search process

Keywords

Microblog Image Search; Social Content; Visual Content; Graph Learning

1. INTRODUCTION

In recent years, social media has undergone a rapid development. Among the social media, microblog platforms have attracted the most attention, and have been used as daily tools to acquire prevailing information. As a result, information search especially image search in microblog plays a more and more important role due to the availability of large amount of images uploaded by users every day.

Generally, users pay more attention to fresh information. Hence current microblog platforms, like Twitter and Tecent Weibo, provide a "most recent" image search service by ordering the resultant images in chronological order. However, there are two key problems with the current services. One is that they just perform retrieval based on text content which is frequently inconsistent with its associated image. According to statistic in [4], only 30% of tweets have the consistency between textual and image information. As a result, a large number of resultant images are irrelevant. Take Tecent Weibo as an example, if we input "Kai-Fu Lee"[1] as a query for image search, among the top 5 returned results sorted in time sequence, only one image tweet is really relevant, as shown by the red box in Figure 1(a). The other problem is that when ranking the returned images, they just take time factor into account while ignoring the relevance of images. But for users, they care both recency and relevance. From the above, we can see that currently available text based image search service in microblog is far from satisfactory. In addition, by performing analysis of large-scale query logs and supplemental qualitative data, Teevan et al. [6] demonstrated that people are more interested with queries of celebrity names in microblog search. Therefore, in this paper we tackle these two problems in most recent microblog image search with celebrity queries.

There have been many works devoted to conducting image search. Image search engines like Google Image Search offer the service based on the textual information in web page instead of the real visual content. In recent years, content based image retrieval has become popular and it supports the use of image as the query like in [5] but the obtained results are only similar to the exact input image and the appearance variation is not accounted for. More recently, multimedia information has been utilized for image search by combining ink, textual, and visual information for image search result clustering and found to be helpful in [2]. Different from web image search, text in microblog is always noisy and the textual content is frequently inconsistent with visual content [4]. On the other hand, we can take advantage of some specific features like *Mention* and *Hashtag* for better performance. To search for images related to celebrities, face recognition is intuitively the most effective method.

MM'14, August 24–27, 2014, New York, NY, USA.
Copyright 2014 ACM 978-1-4503-3063-3/14/11 ...$15.00.
http://dx.doi.org/10.1145/2647868.2655011.

[1] It corresponds to "李开复" in simplified Chinese.

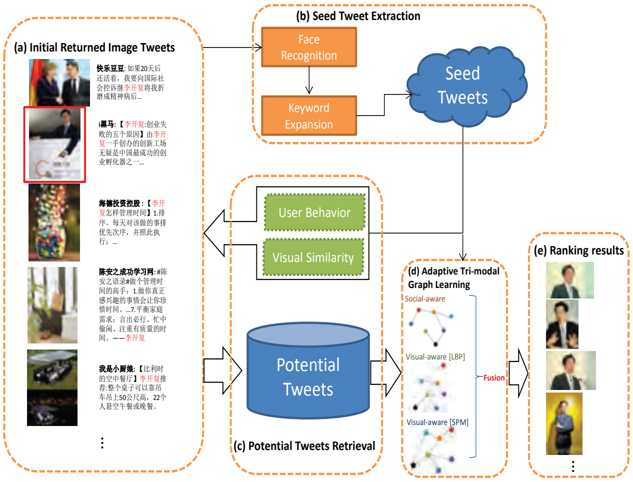

(a) Initial Returned Image Tweets
快乐豆豆: 如果20天后还活着，我要向国际社会控诉继李开复将我折磨成精神病后...
i黑马:【李开复:创业失败的五个原因】由李开复一手创办的创新工场无疑是中国最成功的创业孵化器之一...
海德投资控股:【李开复怎样管理时间】1.排序。每天对该做的事排优先次序，并照此执行；...
陈安之成功学习网:#陈安之语录#做个管理时间的高手：1.做你真正感兴趣的事情会让你珍惜时间。...7.平衡家庭需求：言出必行、忙中偷闲、注重有质量的时间。——李开复
我是小厨娘:【比利时的空中餐厅】李开复推荐:整个桌子可以靠吊车吊上50公尺高，22个人悬空午餐或晚餐。
⋮
(b) Seed Tweet Extraction
Face Recognition
Keyword Expansion
Seed Tweets
User Behavior
Visual Similarity
Potential Tweets
(c) Potential Tweets Retrieval
(d) Adaptive Tri-modal Graph Learning
Social-aware
Visual-aware [LBP]
Visual-aware [SPM]
Fusion
(e) Ranking results
⋮

Figure 1: The framework of our proposed method.

However, although the most recent face recognition work can achieve a high accuracy [3], it will fail frequently when used on images in microblog as those images are usually fuzzy, in low resolution and with small faces or side-view faces. Besides, there are many images where no face shows up but the back of celebrity can be clearly identified. For such images, face recognition cannot work.

In this work we propose a novel approach to search for recent celebrity images in microblog. To our knowledge, there is no existing work that deal with this problem. Taking recency and relevance into account, we incorporate textual, social and visual information into the proposed approach. Based on the initial text based search results, face recognition is first conducted. To complement the limited accuracy of face recognition, we extract extended keywords to generate more highly relevant tweets called *seed tweets*. Then we extend seed tweet set to gather more potential tweets by utilizing user behavior and visual similarity. To rank all the obtained seed and potential tweets, we adopt a tri-modal graph-based learning algorithm to adaptively integrate social and visual information which are expressed in three modalities: social-aware and two visual-aware modalities with two different visual features. Finally, we evaluate the proposed method on a real-world dataset containing 0.1 million microblogs of 100 different person names crawled from Tencent Weibo. To ensure efficiency, we limit the computation in the initial text-based results. Experimental results demonstrate that we could approximately achieve three-fold improvement as compared to the text-based results.

2. THE PROPOSED METHOD

2.1 Seed Tweet Extraction

In order to search for the most recent and relevant images of celebrities in microblog, our first step is to extract a set of seed tweets whose images are regarded as highly related to this celebrity. Given a text query with the name of the celebrity using the "most recent" ranking option, microblog platform returns search results in chronological order. We denote the results as $T_o = \{T_1, ..., T_n\}$ where each T_i is a tweet. As we can see in Figure 1(a), a high proportion of returned images are irrelevant as the search criteria is text based. To obtain a set of seed tweets where the results are mostly relevant, we adopt a two step scheme consisting of face recognition and keyword expansion.

During offline training process of face recognition, we collected images containing celebrity faces by Google Image Search. To crop the face from images as positive samples, we employ face detection as proposed in [1]. Subsequently we implement the Eigenfaces algorithm [8] for training the face recognition model. Thus, given a tweet image, face recognition is run to verify whether it is the celebrity. Then the tweet images likely to contain the celebrity faces are kept as part of seed set, denoted as T_f. How T_f is related to the celebrity depends on the face recognition performance. However, in our work, the average precision of face recognition has just reached 75.8% and the average recall is only 46.52% which are not good enough to identify all the highly relevant tweets in T_o.

Therefore we need to do filtering on T_f in order to remove those false positives. Here we propose a text based method based on keyword expansion technique. To generate expanding keywords, we compute the frequency of each word in T_f. The top N frequent words which are not in stop word set are chosen as the expanded keywords. The tweets among T_f containing the expanded keywords are set as the seed tweets which we denote as T_s. According to experiments, the average accuracy of T_s is 85.6%. T_s contains highly relevant images but it covers only a small fraction of images related to the celebrity because many relevant images are not searchable by current set of keywords; and face recognition often fails to detect faces or recognize the detected face correctly.

2.2 Potential Tweets Retrieval

In this section, we extend the seed tweets to include more possibly relevant tweets. To efficiently obtain more relevant tweets, we analyze the seed tweets from two aspects. First, we identify the behavior pattern of users. The users who posted more than k number of seed tweets are selected as active users. Here we set k to 2 as the set T_s is small. We set all the tweets posted by active users, denoted as T_u, as one part of potential tweet set. Meanwhile, the visual information is also taken into account. We respectively calculate visual distance between every seed image and images in T_o and add the top K similar images, denoted as T_v, to the potential tweet set. Here we employ Euclidean distance on image features which will be detailed at Section 2.4. Finally, we obtain the final potential tweet set, denoted as $T_p (T_p = T_u \bigcup T_v)$.

2.3 Adaptive Tri-modal Graph Learning Algorithm

Having obtained the seed set T_s and the expanded potential set T_p, the next step is to rank these tweets according to the relevance to the celebrity. Here we propose a tri-modal graph learning algorithm to adaptively integrate visual and social information.

First, we assume that there are m tweets in $\mathcal{T}(T_s \bigcup T_p)$, $\mathcal{T} = \{t_1, t_2, \dots, t_m\}$, and the relevance score vector of $\mathcal{T}$ is $\mathbf{r} = [r_1, r_2, \dots, r_m]^{\mathrm{T}}$. Next a graph-based ranking framework is used to predict whether the tweets in $\mathcal{T}$ are relevant to the input celebrity. This problem is formulated as minimizing the following cost function [7]

$$\mathcal{Q}(\mathbf{r}) = \mathcal{R}_{NLap}(\mathbf{r}, \mathcal{T}) + \sigma \times \mathcal{D}(\mathbf{r}, \bar{\mathbf{r}}). \tag{1}$$

The two terms on the right side of equation (1) respectively represent the regularization term that maintains visual consistency and the distance term that measures the difference between ranking score list and original ranking list. Here σ is a scaling parameter for modulating the effect of distance. $\bar{\mathbf{r}} = [\bar{r}_1, \bar{r}_2, \ldots, \bar{r}_m]^{\mathrm{T}}$ is the initial ranking score list.

There are usually two forms for regularization term: graph Laplacian regularizer and normalized graph Laplacian regularizer. Here we utilize the latter as Tian et al. [7] verified its effectiveness over graph Laplacian:

$$\mathcal{R}_{NLap}(\mathbf{r}, \mathcal{T}) = \sum_{i,j=1}^{n} W_{ij} \parallel \frac{r_i}{d_{ii}} - \frac{r_j}{d_{jj}} \parallel^2 = \mathbf{r}^T L_n \mathbf{r}, \qquad (2)$$

where $L_n = \mathbf{D}^{-1/2}(\mathbf{D}-\mathbf{W})\mathbf{D}^{-1/2}$ is the normalized Laplacian matrix. Here $\mathbf{W}$ is a similarity matrix in which W_{ij} indicates the similarity of t_i and t_j, and $\mathbf{D}$ is a diagonal matrix with $d_{ii} = \sum_j W_{ij}$.

We estimate the similarity between the i-th and j-th tweets with a Mahalanobis distance metric which takes into account the correlations of the data set and can be learned by an optimization framework:

$$W_{ij} = \exp(-(t_i - t_j)^{\mathrm{T}}\mathbf{S}(t_i - t_j)) = \exp(-||\mathbf{M}(t_i - t_j)||^2), \quad (3)$$

where $\mathbf{S}$ is a symmtric positive semi-definite real matrix, which can be decomposed as $\mathbf{S} = \mathbf{M}^{\mathrm{T}}\mathbf{M}$. Here we denote $\mathbf{M}$ as a d-by-d diagonal matrix. Then equation (3) is equivalent to transforming each tweet t_i to $\mathbf{M}t_i$.

In order to integrate visual and social information into this graph learning framework, we regard each feature description of the tweet, such as visual or social feature, as a modality. Hence, according to extracted feature described in Section 2.4, there are three different modalities: two kinds of visual aware modalities and one social aware modality. Then we consider integrating different modalities to model adaptive multiple graph learning. In this work, we linearly combine the normalized graph Laplacian regularizers of the three modalities:

$$\mathcal{R}_{NLap}(\mathbf{r}, \mathcal{T}, \lambda) = \sum_{i,j} \lambda_1 W_{ij}^s \parallel \frac{r_i}{d_{ii}^s} - \frac{r_j}{d_{jj}^s} \parallel^2 + \sum_{i,j} \lambda_2 W_{ij}^l \parallel \frac{r_i}{d_{ii}^l} - \frac{r_j}{d_{jj}^l} \parallel^2 + \sum_{i,j} \lambda_3 W_{ij}^v \parallel \frac{r_i}{d_{ii}^v} - \frac{r_j}{d_{jj}^v} \parallel^2 + \varphi \parallel \lambda \parallel^2, \qquad (4)$$

where $W_{ij}^k = \exp(-||\mathbf{M}^k(t_i^k - t_j^k)||^2)$. Here the first term of right side of equation (4) is the social aware modality while the second (LBP) and third (SPM) are the visual aware modalities. The last term is used to adaptively modulate the impacts of these three modalities. λ_k is the weight for each modality, it satisfies $0 \leq \lambda_k \leq 1$ and $\sum_{k=1}^{3} \lambda_k = 1$. φ is a coefficient to learn the weights.

As for distance term in the cost function, we choose squared loss term as it can easily solve the optimization framework. Accordingly, the cost function can be formulated as the following form:

$$\mathcal{Q}(\mathbf{r}, \lambda) = \sum_{i,j} \lambda_1 W_{ij}^s \parallel \frac{r_i}{d_{ii}^s} - \frac{r_j}{d_{jj}^s} \parallel^2 + \sum_{i,j} \lambda_2 W_{ij}^l \parallel \frac{r_i}{d_{ii}^l} - \frac{r_j}{d_{jj}^l} \parallel^2 + \sum_{i,j} \lambda_3 W_{ij}^v \parallel \frac{r_i}{d_{ii}^v} - \frac{r_j}{d_{jj}^v} \parallel^2 + \sigma \parallel \mathbf{r} - \bar{\mathbf{r}} \parallel^2 + \varphi \parallel \lambda \parallel^2 . \qquad (5)$$

For equation (5), we adopt an alternating optimization method to solve it, which is analogous to the solution in [9].

2.4 Tweet Features

For each image tweet, we extract the following two sets of features:

1) Social features.

- 6-dimensional social features. The features include 3-dimensional relationship features (whether the tweet poster is a followee, follower or friend of the target person), the related mention feature (indicated by a "@" symbol to explicitly address the target person in the tweets), the related hashtag feature (tags surrounded with # symbol to indicate discussed topics about the target person), and the URLs presence feature.

2) Visual features.

- 11,564-dimensional LBP features. We first resize each image into 64 × 64 pixels. Then we extract local binary pattern (LBP) histograms from it and concatenated them into a single, spatially enhanced feature histogram efficiently representing the whole image.
- 21,504-dimensional SPM features. The SIFT features are extracted from densely located patches centered at every 6 pixels on each image and the size of the patches is fixed as 16 × 16 pixels. We construct a visual word dictionary containing 1024 words from the training samples via K-means clustering. Each SIFT feature vector is encoded into a 1024-dimensional code vector via vector quantization. Then the code vector from each image are pooled into a 21,504-dimensional feature vector in a spatial pyramid manner.

3. EXPERIMENTS

3.1 Experimental Settings

In this work, we demonstrate the performance of the proposed method by using Tencent Weibo, China's leading microblogging service. As we focus on celebrity image search, we selected Top 100 names based on Celebrity Popularity List which ranked by number of followers. These celebrities include actors, musicians, politicians, athletes and so on. We then used image search function of Tencent using the celebrity names as queries with "most recent" ranking option to collect the image tweets as well as their associated user relationship information for the period from 1th March to 30th June 2012. Due to the insufficiency of returned results for some queries, we discarded these queries. Finally, we obtained a dataset consisting of 80 names, 125,900 image tweets and 23,481,896 user relationship pairs. As there is no ground truth for our dataset, each image is annotated by three volunteers with a three level of relevance, using the scores of 0,1,2 to respectively denote very relevant, relevant, and irrelevant images.

In our method, there are two parameters, i.e., φ and σ (equation(5)). To study the effect of the two parameters, we set up a tuning dataset by randomly selecting a subset from the dataset. The tuning set contains 8 names. As there should be no overlap between tuning set and testing set, we define the tweets from the remaining names as testing set.

To construct face model of face recognition, for each query we respectively crawled 800 images as the positive samples and 1000 images as negative samples from Google. These images were all preprocessed by face detection to produce face images. Since the average precision of face detection method is about 71.04%, we artificially removed those false detected results so as to build superior face models.

In order to measure the performance of our method, we adopted Mean Average Precision (MAP) which is a standard evaluation criterion in information retrieval.

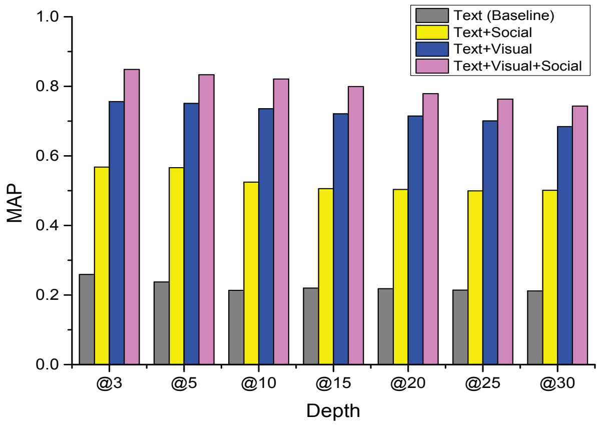


Figure 2: Comparison of the MAP@N measurement for results obtained by different methods.

3.2 Experimental results

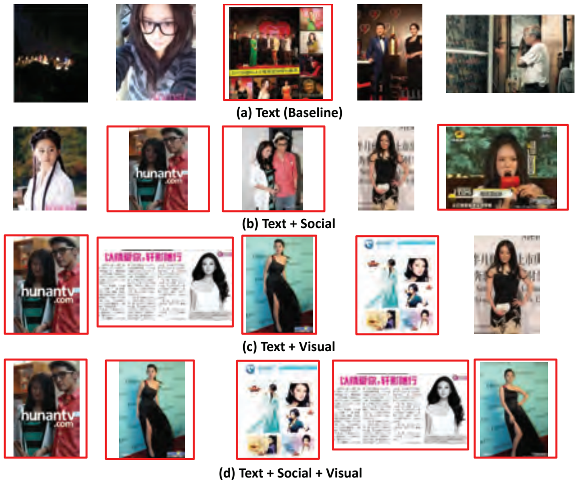
In this section, we compare our proposed approach that integrates visual and social information into initial text-based results with the methods that only leverage one kind of information. We consider the text-based search result as the baseline, denoted as "Text"; and we use "Text+Visual" to denote the method that leverages visual information containing two kinds of visual features described in Section 2.4. Similarly, "Text+Social" is used to denote the method that leverages social information. Finally, "Text+Visual+Social" represents our proposed method.

We first compare the MAP measures with different depth of these methods. Figure 2 illustrates the MAP@3, MAP@5, MAP@10, MAP@15, MAP@20, MAP@25, MAP@30 measurements obtained by these methods in testing set. From the figure we can see that all the methods achieve encouraging improvements as compared to the baseline. Moreover, the "Text+Visual+Social" approach consistently and substantially achieves better performance than "Text+Social" and "Text+Visual" methods. From these results, we can conclude that visual and social information are both helpful for searching the most recent and relevant images. Furthermore, we observe that our proposed method is able to improve the MAP metric twice or thrice as compared to the baseline.

Next we compare the most recent results obtained by different methods. The top 5 results for an example query "An Yixuan" are demonstrated in Figure 3. The first row shows the most recent images in the initial text-based ranking list, where most of them are irrelevant. From the second and third rows, we can observe that the obtained results by "Text+Social", "Text+Visual" methods have been improved to different degrees. However, they still contain at least one irrelevant image in the top 5 results. The last row ranked by our proposed method obtains the best performance in which the most recent top 5 images are all relevant.

4. CONCLUSIONS

In this work, we proposed an effective framework to tackle the problem of recent celebrity image search in microblog. Face recognition and expanded keywords matching were employed to generate a seed tweet set based on which more potential tweets were gathered by using user behavior and visual similarity. To rank these tweets, a tri-modal graph learning method was used. In this way, our method successfully integrated textual, visual and social information. Experimental results on a real life dataset from Tencent Weibo demonstrated the effectiveness of our proposed approach.


(a) Text (Baseline)

(b) Text + Social

(c) Text + Visual

(d) Text + Social + Visual

Figure 3: Top 5 results obtained by Baseline and different methods for the example query "An Yixuan". The relevant images are labeled by red rectangle box.

Acknowledge

This work was partially supported by the State Key Development Program of the Basic Research Program of China (grant 2013CB336500) and NUS-Tsinghua Extreme Search (NExT) project (grant R-252-300-001-490).

5. REFERENCES

[1] Y. Abramson, B. Steux, and H. Ghorayeb. Yet even faster (yef) real-time object detection. *IJISTA*, 2(2):102–112, 2007.

[2] D. Cai, X. He, Z. Li, W.-Y. Ma, and J.-R. Wen. Hierarchical clustering of www image search results using visual, textual and link information. In *Proceedings of MM*, pages 952–959. ACM, 2004.

[3] H. Fan, Z. Cao, Y. Jiang, Q. Yin, and C. Doudou. Learning deep face representation. *arXiv*, 2014.

[4] Y. Gao, F. Wang, H. Luan, and T.-S. Chua. Brand data gathering from social media streams. In *Proceedings of ICMR*. ACM, 2014.

[5] J. Sivic and A. Zisserman. Video google: A text retrieval approach to object matching in videos. In *Proceedings of ICCV*, volume 2, pages 1470–1477, 2003.

[6] J. Teevan, D. Ramage, and M. R. Morris. # twittersearch: a comparison of microblog search and web search. In *WSDM*, pages 35–44. ACM, 2011.

[7] X. Tian, L. Yang, J. Wang, Y. Yang, X. Wu, and X.-S. Hua. Bayesian video search reranking. In *Proceedings of MM*, pages 131–140. ACM, 2008.

[8] M. A. Turk and A. P. Pentland. Face recognition using eigenfaces. In *CVPR*, pages 586–591. IEEE, 1991.

[9] M. Wang, H. Li, D. Tao, K. Lu, and X. Wu. Multimodal graph-based reranking for web image search. *TIP*, 21(11):4649–4661, 2012.