Mining Cross-Network Association for Youtube Video Promotion of Britneyspears to Advertise Gangnam Style

Gayam Yamini Reddy [1], Y.Lakshmi Prasanna[2], Dr.M.V.Siva Prasad[3]

[1] *M.Tech student, Department of CSE, Anurag Engineering College, Nalgonda District , Telangana.*
[2] *Assistant Professor, Department of CSE, Anurag Engineering College, Nalgonda District , Telangana.*
[3] *Principal &Professor, Department of CSE, Anurag Engineering College, Nalgonda District , Telangana.*

Abstract— We introduce a novel cross-network collaborative problem in this work: given YouTube videos, to find optimal Twitter followees that can maximize the video promotion on Twitter. Since YouTube videos and Twitter followees distribute on heterogeneous spaces, we present a cross-network association-based solution framework. Three stages are ad-dressed: (1) heterogeneous topic modeling, where YouTube videos and Twitter followees are modeled in topic level; (2) cross-network topic association, where the overlapped user-s are exploited to conduct cross-network topic distribution transfer; and (3) referrer identification, where the query YouTube video and candidate Twitter followees are matched in the same topic space. Different methods in each stage are designed and compared by qualitative as well as quantitative experiments. Based on the proposed framework, we also discuss the potential applications, extensions, and suggest some principles for future heterogeneous social media utilization and cross-network collaborative applications.

Keywords: video promotion, cross-network analysis, social media.

1. INTRODUCTION

Since the launch in 2005, YouTube has established itself as the world's largest video sharing platform. Latest statistics show that within every minute, 100 hours of video are up-loaded to YouTube 1, resulting in an estimate of more than 2 billion videos totally. People act on purpose. It has been recognized that YouTube users share videos with an obvious extrinsic motivation of receiving attentions (e.g., video view), especially for the profit-seeking video content providers 2. In spite of the fact that billions of videos are consumed in YouTube each day, the massive volume makes the exploration of individual videos very difficult. According to research, YouTube video view count distribution exhibit-s a power-law pattern with truncated tails [3]. Most videos have a short active life span, receiving half of the total views in the first 6 days after being published, and with fewer and fewer access thereafter [4]. Therefore, the mismatch between high attention expectation and rare access opportunity calls for YouTube video promotion to broaden the viewership.

Generally speaking, within YouTube, video can be accessed from internal search, related video recommendation, channel subscription or front page highlight. Some work has been devoted to utilizing these sources to promote internal video views. Zhou et al. studied the impact of related video recommendation on video views, with goal to design a strategy to drive YouTube video popularity. In, YouTube search bias phenomenon is investigated to optimize video discovery in YouTube's internal search results. However, essentially as a content repository, YouTube exhibits limited promotion efficiency with the internal mechanisms. Very recent research shows that external referrers, such as external search engines and other social media websites, arise to be important sources to lead users to YouTube videos. Among the social media websites, Twitter has been quickly growing as the top referrer source for web video discovery 3.
Twitter allows users to embed videos in their tweets by posting video links. Followers to these users then receive the tweet feed and become the potential viewers of these videos. The followee follower architecture has established Twitter as a great platform to promote and engage with the audiences and distinguished itself with the significant information propagation efficiency. Twitter followees, especially those with a lot of followers (which we refer to as popular followee), play important roles under social media circum-stances by: (1) acting as "we media", via the control of information dissemination channels to millions of audiences, and (2) acting as influential leaders, via their potential impact on the followers' decisions and activities. YouTube video "Gangnam Style" went viral to become the first web video that reaches one billion views in 5 months,

resulting mainly from its successful strategy of roping in some popularly followed musicians on Twitter, such as Britney Spears, Justin Bieber and Katy Perry. In this context, if we can identify "proper" followees to help disseminate videos, their significant audience accessibility and behavioral impact will guarantee the promotion efficiency. Therefore, the problem of this work is: For specific YouTube video, to identify proper Twitter followees with goal to maximize video dissemination to the followers.

It is not trivial to measure the "properness" of Twitter followees for specific YouTube videos. The challenge lies in two-fold: (1) The level of "properness" is not necessarily proportional to the number of followers (#follower). While a popular followee with a large #follower will guarantee a huge audiences, what video promotion cares is the number of "effective" audiences, who are likely to show interest to the video and with higher probability to take subsequent consuming actions like watch, reshare, etc. A close analogy to advertising can be made, where the followee is viewed as advertising media, whose bid price is decided by #follower. Twitter followee identification is analogous to advertising media selection 4, with goal to achieve the maximum coverage and exposures in a target audience with the minimum cost. (2) Based on the above discussion, whether a Twitter followee is proper for the promotion task is actually decided by the interest his/her followers show to the YouTube videos. However, we only know the followers' activities on Twitter, based on what only the demographics or interest-s on the general level can be inferred. While, the YouTube videos are known to distribute more on specific semantic level. The discrepancy in topic granularity and affiliated platform makes it impractical to directly evaluate Twitter followers' interest to YouTube videos, let alone the costly computation in evaluating each follower and the subsequent aggregation.

Our solution to address the above challenges is inspired by the fact that the same individual usually involves with different social media networks, including media sharing YouTube and Flickr, microblogging Twitter and Tumblr, private/professional social networks LinkedIn and Facebook.

Anderson Analytics shows that the different social media networks share remarkable percentage of overlapped user-s 5. In this context, if we know the corresponding Twitter accounts of YouTube users who show interest to a given video (e.g., upload, favorite, add to playlist), it is confident to identify the Twitter followee that these Twitter accounts jointly followed as the optimal promotion referrer. In practice, it is impossible to obtain all the overlapped accounts between different networks 6. Moreover, a practical solution should be not limited to the specific video and followee, but generalizable on the alike sets. Therefore, in this work, we propose to investigate the problem in YouTube video and Twitter followee topic level, and exploit the observed overlapped users to mine the cross-network topic association for solution. Specifically, based on users' interactions with YouTube videos and Twitter followees, we first build heterogeneous video topic and followee topic, respectively. After that, the topic association is mined from the over-lapped users' distributions on the two topics. Finally, the optimal Twitter promotion referrer is identified by matching with the transferred video distribution on the Twitter followee topic space.

Our contributions in this work can be summarized in the following three-fold:

1. We introduce a new problem of YouTube video promotion on Twitter platform by identifying proper Twitter followees. There exist both trends and demands in exploring external referrers towards promoting social media content.
2. A cross-network association-based solution framework is presented, under which alternative methods have been examined. The solution is validated to discover heterogeneous topic association and facilitate effective video-followee matching in the same topic space.

3. The discussion in Section 5 on the idea of exploiting overlapped users' activities in different networks towards cross-network knowledge mining opens up possibilities to the utilization of heterogeneous social media sources. This will be the key takeaway for future cross-network analysis and applications.

2. RELATED WORK

This section reviews the related topics. Instead of a complete coverage, we only review some representative work in each topic, with goal to position this work in the coordinate of existing work for better understanding the addressed problem as well as the proposed solution.

2.1 Cross-network Collaboration

With various social media networks growing in prominence, netizens are using a multitude of social media ser-vices for social connection and information sharing. Cross-network collaborative applications have recently attracted attentions. One line is on cross-network user modeling, which focuses on integrating various social media activities. In the authors introduced a cold-start recommendation problem by aggregating user profiles in Flickr, Twitter and Delicious. Deng et al. has proposed a personalized YouTube video recommendation solution by incorporating us-er information from Twitter. Another line is devoted to taking advantage of different social networks' characteristics towards collaborative applications. Suman et al. exploited the realtime and socialized characteristics of the Twitter tweets to facilitate video applications in YouTube. In Twitter event detection is conducted by employing Wikipedia pages as the authoritative references. Our work belongs to the second line, where a collaborative application is designed to exploit the propagation efficiency of Twitter to meet the YouTube video promotion demands.

2.2 Social Media Influencer Mining
Previous analysis on Twitter has found that popular users with high in-degree are not necessarily influencers for propagation, which calls for research onto the problem of influencer mining. One line is to identify the domain or topic experts. Representative solutions include the extensions to Page Rank by considering topical similarity, e.g., Twitter Rank, and incorporating auxiliary sources like Twitter lists. Another line is concerned with maximizing influence spread by initializing some seed users. David et al. first defined this problem, which is then applied to product adoption and viral marketing .

Our introduced problem of Twitter followee identification can be viewed as a special case of influencer mining. The existing influencer mining methods mainly focus on single net-work and need an explicit relevance metric, e.g. the topical relevance between follower and followee, and the accept rate between the propagation item and follower. In our problem, the relevance of influencer is designed by items distributed on another network. It is difficult to explicitly define the relevance metric between cross-network knowledge. More-over, to focus on addressing cross-network association, we pay no attention to the complicated social network structure as in the standard maximizing influence problems. What we care is actually the propagation efficiency in the first level of followee-follower network.

2.3 Heterogeneous Topic Association

The core of our solution lies in the heterogeneous topic association between Twitter followee and YouTube video. Typical applications of existing heterogeneous topic association work include cross-media retrieval and heterogeneous face recognition, where invariant feature extraction and sub-space learning based solutions are extensively investigated. Invariant feature extraction methods are devoted to reducing the heterogeneous gap by exploring the most insensitive feature patterns. Klare et al. proposed to extract the SIFT and Multiscale LBP for forensic sketch and mug shot photo matching. In the intradifference and inter-difference are jointly considered into a discriminant local feature learning framework. The basic idea of subspace learning is to learn a new space where the observed heterogeneous data can be well represented provides good surveys of CCA and its extensions to learn a semantic representation from multimodal data. Multimodal topic modeling can also be viewed as one type of subspace learning, where multimodal representations are projected to a shared topic space.

Subspace learning methods focus on maintaining the smoothness for retrieval, i.e., the projected coefficients of two items should be similar if they constitute a training pair. This is different from our goal for heterogeneous topic association and transfer. Invariant feature extraction aims to extract and learn low-level discriminative features, which will largely fail in case of complicated association like heterogeneous social media topics. In this work, we propose a solution framework based on users collaborative involvement in heterogeneous topics. This avoids low-level analysis and can be viewed as a high-level crowd sourcing strategy.
CROSS-NETWORK YOUTUBE VIDEO

This section introduces the cross-network YouTube video promotion problem and the proposed solution. We first formally define the problem:

DEFINITION 1 (CROSS-NETWORK YOUTUBE VIDEO PRO-MOTION).
Imagine we have a collection of YouTube videos V where each v ∈ V is represented by its contained textual words and visual key frames and a collection of Twitter users UT whose followees construct the Twitter followee user collection U followee ⊂ UT . The goal of You tube video promotion is: for a given YouTube video v ∈ V, to identify Twitter followee u ∈ U whose followers are most likely to be interested in v.
3.1 Framework

Our solution consists of three stages: Heterogeneous Topic Modeling, Cross-network Topic Association and Referrer Identification(as illustrated in Fig. 2). The goal of Stage 1 is to discover the latent structure within YouTube video and Twitter user spaces, and facilitate the subsequent anal-ysis and applications in topic level. We conduct this by employing generative topic models, with video as document, textual word and visual feature of key frames as the multi-modal word in YouTube, and user as document, followee as word in Twitter. Through this stage, each YouTube video and Twitter user can be represented as distributions in the derived corresponding topic spaces.

As discussed in the introduction, the discrepancy between the cross-network topic spaces prevents from direct analysis. Stage 2 is designed to address this issue by mining the cross-network topic association. Note that traditional semantic-based criteria tend to fail in capturing the association between heterogeneous entities of video and user. We propose a solution that first aggregates YouTube video distribution to user level, and then exploit the overlapped users among different networks as bridge for association mining. The basic premise is that: if the same group of users heavily involve with topic A in network X and topic B in network Y, it is very likely that topic A and B are closely associated. With the derived topic association, topical distribution transfer between different networks is enabled, i.e., given users' topical interest in YouTube videos, we can infer their most probably followed Twitter followee topics.

Since the ultimate goal is to match video to followee. After the offline Stage 1 and Stage 2, in the online Stage 3, we view each test video as a virtual YouTube user who holds identical topical distribution. It is easy to understand that the virtual user actually represents the typical users in YouTube showing significant interest to the test video, who are exactly potential fans and thus the targeted users. Therefore, after topical distribution transfer, it is promising to identify the Twitter followee that best matches the followee topical distribution of the targeted users as the optimal promotion referrer for the video. In Table 1 we summarize the inputs and outputs for each stage.

3.2 Heterogeneous Topic Modeling

3.2.1 YouTube Video Topic Modeling

In YouTube, the video topics are expected to span over both textual and visual spaces. We introduce a modification to the multi-modal topic model, Corr-LDA. Corr-LDA proposed for the problem of image annotation, by model-ing the correspondence between image segments and caption words. It assumes a generative process that first generates the segment descriptions and subsequently the caption word-s. In our problem, each YouTube video is represented as a pair (f ; w), where f = {f1, • • • , fN } is a collection of N visual feature vectors associated with the extracted key frames, w = {w1, • • • , wM } is the collection of M caption and tag words. Different from image where each word corresponds to one segment, video caption and tag word usually distribute in several key frames.

Therefore, we modified the standard Corr LDA and introduce inverse Corr-LDA (I Corr-LDA) to discover the YouTube video multimodal topics. In particular, we first gen-erate M textual words from the standard LDA model. Then, for each of the N key frames, one of the words is selected and a corresponding key frame is drawn, conditioned on the same topic generating the word. The graphical model of I Corr-LDA is depicted in Fig. 3. After topic modeling, each video v ∈ V can be represented as v = {v1, • • • , v KY }, where KY is the number of topics in the derived YouTube video space, vk = p(zkY |v) is video v's topic distribution on the kth topic.

3.2.2 Twitter Followee Topic Modeling

Since the properness of Twitter followee is decided by the followers, we are interested in investigating into the followee-follower architecture in Twitter. Therefore, we represent each Twitter user (document) with all his/her followees (words) and apply the standard LDA for topic modelling.

Since topic modeling exploits co-occurrence relationships, like the YouTube video topics capturing the frequently co-occurred visual features and textual words in videos, the derived Twitter topics actually capture the shared followees by a subset of Twitter users. Particularly, high topic-word distribution indicates the popularity of followees in a group of Twitter followers, and high document-topic distribution indicates users' significant interest in a class of Twitter followees.

After topic modeling, we can obtain Twitter user topic distribution matrix U T = {uT1 , • • • , uT|UT |}. Each user u ∈UT is represented as uT = {uT1 , • • • , uTKT }, where KT is the number of topics in the derived Twitter followee space, uTk = p(zkT |u) is user u's topic distribution on the kth topic.

3.3 Cross-network Topic Association

3.3.1 YouTube User-Topic Distribution Aggregation

YouTube user's topic distribution can be obtained by aggregating his/her interested videos' distributions. Specifically, for YouTube user u, we construct the interested video set. Vu ⊂ V from his/her uploaded videos, favorite videos and videos in the playlists.

Transition Probability-based Association (TP)

With the derived YouTube and Twitter user topic distributions, we present the solutions for topic association mining. Recall that the basic idea is: if many overlapped users who take interests in the ith YouTube topic also follow the jth Twitter topic, the association between the two topics aij tends to be strong. One direct way is to examine the joint involvement of cross-network topics in the overlapped users.

Regression-based Association:

probability-based method directly calculates over all overlapped users, where noisy user topic distributions will deteriorate the derived association matrix. Alter-native way to obtain the association matrix is to solve an optimization problem. Rewriting the user topic distribution matrices.

3.3.4 Latent Attribute-based Association (LA)

The aforementioned two association methods are devoted to finding the cross-network association matrix A. Actually, to conduct the topical distribution transfer, the association matrix is not necessarily needed. Moreover, such a matrix exists under the assumption of linear association, which does not hold in complicated cases.

3.4 Referrer Identification

With the cross-network distribution transfer function F, we can estimate arbitrary user's Twitter followee topic distribution by inputting his/her YouTube video topic distribution. In our video promotion problem, given a test YouTube video vt, we simulate a virtual user with identical topic distribution vtY = p(zY |vt) to represent the typical YouTube users liking the video 8. After distribution transfer, the virtual user's Twitter followee topic distribution vtT = p(z T |vt) actually reflects the most probable Twitter following pat-terns for the video fans.

4. EXPERIMENTS

Since no ready cross-network dataset is available, we construct a new dataset with user account linkage between YouTube and Twitter. Google+ encourages users to provide the external links to their other social media network ac-counts. We first collected 143,259 Google+ users, among which 38,540 provide YouTube account, 39,400 provide Twitter account, 11,850 provide both accounts. For each You Tube user, we further downloaded his/her uploaded videos, favorite videos, playlists and the involved video information via YouTube API. For each Twitter user, we downloaded his/her followee set and user profiles via Twitter API.

4.2 Heterogeneous Topic Modeling

4.2.1 Topic Number Selection

In topic modeling, the selection of topic number is very important. We resort to the perplexity in this paper, which is a standard measure for estimating how well one generative model fits the data. The lower the perplexity score is, the better the performance. We test the perplexity with different topic number KY and KT on 490,000 held-out You Tube videos and 9,400 held-out Twitter users, respectively. The perplexity scores on different topic numbers can be seen that on both YouTube and Twitter, the perplexities decrease dramatically first before reaching a relatively stable level and then have a tendency to increase when the models are over fit. Since larger topic number requires more computational cost and has over fitting risk, we prefer the smallest topic number that leads to perplexity on the stable level. Therefore, we choose the topic number KY = 40 for YouTube and KT = 80 for Twitter.

4.2.2 Visualization of Discovered Topics

In order to interpret the derived topic spaces, we visual-ize some of the discovered topics in YouTube and Twitter, respectively. Table 3 shows two sampled YouTube video topics. For each topic, we provide the top-5 probable word-s and 3 most representative videos. Representative videos are ranked based on the video-topic distribution $p(z_k^Y |v)$ and represented by the key frames and video titles in Table 3. By visualizing both the semantic and visual information, it is very easy to interpret the domain knowledge associated with each topic. Moreover, the discovered video topics show high consistency between textual semantics and visual patterns.

sampled Twitter followee topics, with each visualized by the top-3 probable followees and the followees' profile information. It is conceived that the discovered Twitter topics have a quite wide coverage: the general topic #43 addressing the game-related popular followees, the specific topic #10 consisting of Forbes influencers, and even the geographic topic #38 with the top followees all
Twitter Referrer Identification:

Experimental Setting

2,061 videos that more than 15 overlapped users have shown interest to are selected to construct the YouTube test video set V_t. Meanwhile, 79,169 Twitter followees who
are followed by more than 50 users construct the candidate Twitter followee set Ut f ollowee.

We use Normalized Discounted Cumulative Gain (NDCG) as the evaluation metric, which is widely used in retrieval problems. NDCG is defined as:

We consider the following settings for comparison:

- Random: randomly select k followees from U_t f ollowee;

- Popularity: select k popular Twitter followees with the most #followers;

- Regression+Direct: distribution transfer by

  Regression l1, matching by Direct product;

- Regression+Weighted: distribution transfer by

  Regression l1, matching by Weighted product;

- LA all+Direct: distribution transfer by LA all, matching by Direct product;

- LA all+Weighted: distribution transfer by LA all, matching by Weighted product.

4.4.2 Experimental Results and Analysis

We show NDCG@5 for different settings in Fig. 7. It is observed that: (1) Popularity fails to identify the optimal

Twitter referrer. This is easy to understand. While high #follower guarantees the coverage of potential viewers (precision), the retrieved follower set is expected to al-so include many uninterested users (recall), which deviates our goal towards target promotion. (2) Conducting distribution transfer by LA all+Direct and LA all+Weighted obtain better performance than Regression+Direct and Regression+Weighted. This coincides with our motivation that more accurate distribution transfer contributes to improved referrer identification. (3) The settings with weighted product-based matching consistently outperform those with direct product. This demonstrates the advantage of topic weight optimization. One possible interpretation is that different topics contribute differently in view of referrer identification.

5. DISCUSSION

5.1 Application and Extension

The proposed framework also enables solutions to other applications. From Stage 2, we actually obtain the association between YouTube video interests and Twitter following patterns. Based on this association, cross-network personalized recommendation problems on two directions can be enabled: recommending Twitter followee topic or Twitter lists given YouTube video interest [18], and recommending YouTube videos given Twitter followee list.

Moreover, for LA-based solutions, a careful investigation into the derived latent attributes, e.g., checking the coupled factors' distribution in the two topic distribution, will gain understanding into the examined user collection and facilitate cross-network collaborative applications like user clustering. User classification can also be conducted if we annotate the derived user attributes.

Another promising application is on examining the value of Twitter followees. Current methods value Twitter followees by directly analysing their followers' demographic-s information, e.g., the followee has a lot of young female followers. The proposed framework in this work facilitates application-oriented Twitter followee value analysis, by associating Twitter followee topic space with the needed topic spaces. For example, our work can be viewed as valuing Twitter followee promotion efficiency to YouTube videos. This significantly expands understanding into the value of Twitter followees.

Extension. Our current solution only employs the con-tent feature of YouTube test videos, i.e., title, tags and key frames. One extension is to combine with social features, e.g., who uploads or favourites the video. The consideration of user social network is also expected to contribute to improved cross-network association.

Moreover, the current referrer identification is on the individual level, i.e., no interaction between followees is considered. In practice, when choosing a group of followees as the promotion referrers, follower intersection of the candidate followees need to be modelled. Analogous to advertising, as discussed in the introduction, this work actually addresses the problem of advertising media selection. Other problem-s in advertising include advertising anchor text generation (i.e., optimizing video description for promotion), and advertising slot bid (i.e., followee reshare time selection).

CONCLUSION

We have proposed an overlapped user-based association solution framework, to address the novel cross-network You Tube video promotion problem. Alternative methods have been developed and evaluated, to demonstrate the effectiveness of exploiting user collaboration towards heterogeneous knowledge association. The proposed framework is quite flexible, and can be generalized to other cross-network collaborative problems. We hope that this paper could serve as a good chance to emphasize the collective utilization of social media sources and further the agenda of cross-network analysis and application in social multimedia research.

7. REFERENCES

[1] Bernardo A Huberman, Daniel M Romero, and Fang Wu. Crowdsourcing, attention and productivity. Journal of Information Science, 35(6):758–765, 2009.

[2] Jean Burgess and Joshua Green. YouTube: Online video and participatory culture. John Wiley & Sons, 2013.

[3] Meeyoung Cha, Haewoon Kwak, Pablo Rodriguez, Yong-Yeol Ahn, and Sue Moon. I tube, you tube, everybody tubes: analyzing the world's largest user generated content video system. In IMC 2007, pages 1–14.

[4] Xu Cheng, Cameron Dale, and Jiangchuan Liu. Statistics and social network of youtube videos. In IWQoS 2008, pages 229–238.

[5] Renjie Zhou, Samamon Khemmarat, and Lixin Gao. The impact of youtube recommendation system on video views. In IMC 2010, pages 404–410.

[6] Youmna Borghol, Sebastien Ardon, Niklas Carlsson, Derek Eager, and Anirban Mahanti. The untold story of the clones: Content-agnostic factors that impact youtube video popularity. In KDD 2012, pages 1186–1194.

[7] Flavio Figueiredo, Jussara M Almeida, Marcos Andr´e Gon¸calves, and Fabr´ıcio Benevenuto. On the dynamics of

social media popularity: A youtube case study. arXiv preprint arXiv:1402.1777, 2014.

[8] Delip Rao, David Yarowsky, Abhishek Shreevats, and Manaswi Gupta. Classifying latent user attributes in twitter. In SMUC 2010, pages 37–44.