User-Oriented Social Analysis across Social Media Sites

Ming Yan, Zhengyu Deng, Jitao Sang, and Changsheng Xu

National Lab of Pattern Recognition, Institute of Automation
CAS, Beijing 100190, China
China-Singapore Institute of Digital Media, Singapore, 139951, Singapore
{ming.yan,zydeng,jtsang,csxu}@nlpr.ia.ac.cn

Abstract. The vast amount of user-generated data in various and disparate social media sites contains rich and diverse information about what is happening around the world. Digging into such user-generated data distributed in different social media sites helps us better understand what people are interested in and how they feel about certain topics. In this paper, we investigate into users' behavior data in Twitter and YouTube to figure out whether people's attention on certain topics has some sort of temporal order between Twitter and YouTube on user level. We collected a real world dataset of 8,518 users with account associations between Twitter and YouTube as well as all their behavior data with timestamp since Jan. 2012. The results demonstrate that more users tend to get access to certain events earlier in Twitter than in YouTube and the ratio is somewhat topic-sensitive.

Keywords: temporal, cross-network, user-oriented, social behavior analysis.

1 Introduction

With the emergence and popularity of various and disparate social media sites, users are now frequently participating in multiple social media sites simultaneously. According to statistics, on average every user owns as many as 5.6 social media accounts while they visit about 3 different social media sites per day. Large amounts of information is generated, shared and exchanged by the users across various social media sites, which capture how people are spending their time and what they are interested in. This also makes user's information distributed in multiple social media sites, thus the aggregation and analysis of user-related information in multiple sites will inevitably give us a better understanding of users themselves.

Nowadays many social media aggregation tools such as about.me and FriendFeed make it convenient for users to provide and aggregate their separate accounts, which provides possibilities for the cross-network analysis on user level. We also find that more and more users are willing to provide their accounts in other platforms when registering into social network sites. For example, we

A. Petrosino, L. Maddalena, P. Pala (Eds.): ICIAP 2013 Workshops, LNCS 8158, pp. 482–490, 2013.

have observed from our Google+ dataset that a considerable proportion of users provide the external links to their other social sites such as YouTube, Flickr and Twitter at their Google+ homepages.

In recent years, cross-network social media analysis and application has attracted more and more academic attentions. The works mainly consist of multi-platform social network analysis [1][2], cross-network user identification [3][4] and cross-network collaboration [5,6,7,8], etc. [1][2] compared the social network topology structure and some corresponding SNA metrics of different social network sites and analyzed how information propagates through different network structures. [4] developed a new re-identification algorithm utilizing only the network topology to re-identify the anonymous Twitter graph from Flickr graph with relatively small error rate. [5][6] analyzed the characteristics of different social tagging sites and proposed some cross-network user modeling strategies. [8] explored ways in which event content identified on one social media site can be used to retrieve additional relevant event content on other social media sites so as to better understand the related events.

In this paper, our cross-network social analysis work focuses on investigating into the temporal order of user's access or reaction to certain topics shared across different social media sites: Twitter and YouTube. Take the event "US presidential election 2012" as an example, our motivation is to figure out if most users first post or reshare the tweets about "Obama vs Romney" or other topics related to this event in Twitter and then go to watch related videos in YouTube for more information or vice versa. Some researchers have already begun to explore and compare the speed of the emergence and spread of certain events among different text-based social sites [9][10]. [9] made some analysis on the identification of events using Twitter and Wikipedia and found that Wikipedia lags behind Twitter by about two hours. [10] found six main temporal shapes of attention of online content between Twitter and Webblogs. However, these existed cross-network temporal analysis works only focus on the global popularity of certain events. Our work made this temporal analysis on a user level, i.e., we try to find whether the temporal order exists for the majority of users. To the best of our knowledge, this is among the first work to make a temporal analysis of certain events in multiple platforms on a user level. Finally, we also find the temporal order is somewhat related to the category of the events. For example, a larger proportion of users first share their opinions about the newly-released electronic products in Twitter and then view the videos in YouTube which describe the new features of the products.

2 Data Collection

In this section, we first describe how we collect our cross-network user dataset. Then we introduce the way we extract some certain topics which are frequently talked about both in Twitter and YouTube and how we represent or track these topics on user level in detail.

2.1 Cross-network User Data Collection

In order to obtain a collection of users who have both accounts in Twitter and YouTube, we started from Google+ website where people provide many external links to their other social network homepages and collected about 100,000 users in total. Then we kept only the users who have both Twitter and YouTube accounts and removed those who have less than 10 videos in their upload list, which results in the final cross-network dataset with 8,518 users. The users' rich social behavior data from Jan. 2012 to Apr. 2013 were also downloaded from Twitter and YouTube respectively. In Twitter, we collected all the users' posted tweets with timestamp (including retweets). In YouTube, all the available information for a user such as uploading and favoriting a video were downloaded. As a result, we got more than 8 million tweets and 0.6 million video-related behaviors for our 8,518 user dataset. The following experiments are all based on this dataset.

2.2 Topic Extraction and Representation

To find some topics which are widely spread between Twitter and YouTube in our dataset, we combine the official statistics of the trending topics in 2012 with a simple sorting method by word frequency. In Twitter we use tweet words and hashtags to identify or represent a topic while in YouTube the video tags are utilized to identify a topic. We started from the top trending searches of 2012 revealed by Google [1] and collected all the trending topics with different locations and different categories it mentioned. Then we aggregated all the tags of the YouTube videos and all the tweet words as well as the hashtags involved in our dataset respectively. We further counted the word and tag frequency upon all the behavior data in Twitter and YouTube respectively and sorted the words and tags according to their frequency. Finally, we selected the trending topics with high frequency in our behavior dataset in both networks from all those collected through Google official data as our ultimate topics. As a result, we obtain 22 trending topics shown in Table 1. The following cross-network data analysis works in section 3 are all based on these 22 trending topics and we will only show the topic number in the subsequent experiment results for brevity.

Next, we will describe how we identify and represent the selected trending topics in Twitter and YouTube respectively since different social networks tend to use different terms to indicate the trending topics. For instance, people may use "obama election 2012" or "mitt romney election" to indicate the topic "US presidential election 2012" in YouTube while they may adopt the hashtags such as #USelection, #voteobama or #obama2012 to indicate the same topic in Twitter. To capture all the terms which can represent the trending topics in Twitter and YouTube respectively, we use the YouTube Search API to search for all the 22 trending topics in YouTube engine and aggregate the video tags of

[1] `http://www.marketingcharts.com/wp/topics/entertainment/google-reveals-2012s-top-trending-searches-25381/`

Table 1. The final selected trending topic list

Topic	Topic	Topic
1. US presidential election 2012	9. Samsung Galaxy S III	17. google glasses
2. gangnam style	10. Michael Jackson	18. call me maybe
3. super bowl 2013	11. Christmas 2012	19. Spider Man
4. Olympic 2012	12. Google Nexus 4 release	20. Skyfall
5. Justin Bieber	13. Iphone 5 release	21. End of the World 2012
6. star wars film	14. Call of Duty: Black Ops II	22. Whitney Houston
7. The Dark Knight Rises	15. Doctor Who TV Series	
8. Minecraft Game	16. Prometheus	

the returned videos while in Twitter we search for the related tweets from a wider range of tweet dataset we downloaded since the Twitter Search API can only search the tweets currently posted. Then we adopt the same word frequency sorting method as in the previous Topic Extraction procedure and manually select the high frequency terms which can represent the topic as our indicator for the topic. In this way, we count how many of the 8,518 users have referred to the different selected topics via their user tag cloud in Twitter alone, YouTube alone and both in Twitter and YouTube respectively. The result is shown in Table 2. We can see that users are more active in Twitter and for all the topics the number of involved users are larger in Twitter than YouTube. Moreover, the topic overlap on user level between Twitter and YouTube is relatively small that only a small proportion of users pay attention to certain topics both in Twitter and YouTube.

Table 2. The user number who have referred to each of the selected trending topics (we only show 20 topics due to the space and T stands for Topic) in Twitter alone, YouTube alone and both in the two respectively.

	T1	T2	T3	T4	T5	T6	T7	T8	T9	T10
Twitter	2908	3850	1107	1376	1071	2385	2251	857	1164	519
YouTube	949	1181	239	310	405	1171	638	572	458	321
Both Two	521	602	82	115	78	350	219	221	192	62
	T11	T12	T13	T14	T15	T16	T17	T18	T19	T20
Twitter	4155	1434	2708	890	1114	791	1704	897	951	1254
YouTube	1270	361	497	174	586	231	658	508	264	249
Both Two	729	189	246	63	177	75	269	117	82	85

3 Cross-Network Data Analysis

In this section, we first present a global temporal dynamic analysis of the popularity of certain topics regarding all the users' behavior data in our dataset. Then we further investigate whether there are some temporal behavior patterns across different social media sites on user level and topic level respectively.

3.1 Global Attention of Certain Topics Starts Earlier in Twitter than in YouTube

To capture a full dynamic of some certain topics and verify the effectiveness of our method to represent and track certain topics, we select two popular topics in our trending topic list in Table 1, i.e. "US presidential election 2012" and "super bowl 2013", we then count how many users in our dataset begin to pay attention to these topics each day during an observation of two months. For the topic "US presidential election 2012", we observe from Oct.1 to Nov.30 in 2012 of which the final election day happened on Nov.6. Here we mainly present and illustrate our results on this topic for the sake of space limit. The temporal dynamic result of the users' attention on this topic is shown in Figure 1. We further track this topic with the real-world timeline in its wikipedia page [2] and the real events are labeled in the figure. We can find that the statistics from our dataset well capture the real events happened in "US presidential election 2012", since it can be seen from the Figure 1 that a peak of user attention occurs near each of the important events in the topic which in turn indirectly verify the effectiveness of our mechanism to represent and track the topics.

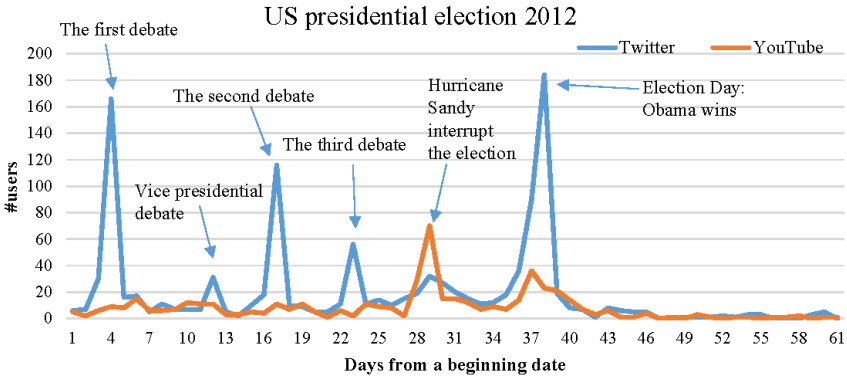

Fig. 1. The temporal dynamics of global user attention for the topic "US presidential election 2012"

Moreover, for a qualitative analysis of the temporal order of users' attention on this topic between Twitter and YouTube in a global view we can see from the figure that many users in Twitter begin to follow this topic since the first, second and third debate while users in YouTube mainly begin to follow this topic since the event "Hurricane Sandy interrupts the election" occurs. Finally, the attention in both networks achieves the peak when the final election day comes and Obama wins the election. Since Twitter network can identify more sub-topics and users

[2] http://en.wikipedia.org/wiki/United_States_presidential_election,_2012_timeline

are more active and productive, users on average tend to get access to certain topics faster in Twitter than in YouTube. This is also reasonable by realizing the fact that users can speak almost whatever they like in Twitter just by typing some words while users in YouTube have to spend more efforts recording and uploading the videos before people can get to know this. The difference of the efforts it takes to create the content accounts for this gap to some extent. The similar characteristic also goes for the topic "super bowl 2013" and we omit it here in case of space limit.

Nevertheless, we've also observed an interesting phenomenon that most the emergence of the peak time of user attention lags behind the occurrence time of the real events by one day. For instance, the first, second and the third debate happened in Oct.3, Oct.16 and Oct.22 respectively according to Wikipedia statistics while our peak time for each of the three events all comes one day later than the actual occurrence time. It may be due to the fact that many users are not that active in social networks and they just begin to pay attention to this topic when they notice their friends or some real-time event reporters post some tweets about this topic. Besides, it takes some time that information spread widely to the general public in social media network .

3.2 Attention to Certain Topics is Earlier in Twitter on User Level

After the global temporal analysis of user behavior between Twitter and YouTube, we further investigate into the temporal patterns across different social media sites based on each single user. In other words, we try to figure out whether the majority of users are first involved in this topic in Twitter and then go to YouTube for more details or vice-versa when some topic emerges. Therefore, we first collect the users who have referred to the topics in our trending topic list both in Twitter and YouTube (the available number of the users can be found in Table 2). Then we analyze the users' behavior data to find when the users first referred to the topics in Twitter and YouTube respectively. We use the same method to identify the topic as in section 2.2 and if the topic indicator occurs in user's tweets or video tags, we consider that the user refer to the topic. We further get the date when users begin to pay attention to the topic for the first time and judge in which network the user pay attention to the selected topic earlier. We aggregate the votes from all the users who have referred to the topics both in Twitter and YouTube and calculate the ratio between the Twitter votes and YouTube votes. The result is shown in Table 3.

From Table 3, we can see that the number of user votes for "Twitter is earlier" is far larger than that for "YouTube is earlier" almost on all topics. In other words, a larger proportion of users tend to first focus on some trending topics in Twitter and then they will go to YouTube for more details on this topic. As a result, the local temporal analysis based on each single user also meets the global patterns demonstrated in section 3.1 indicating that information emerging and spreading in Twitter is faster than that in YouTube.

Table 3. The number of user votes for "Twitter is earlier" and "YouTube is earlier" and their ratio on the topics in our trending topic list

	T1	T2	T3	T4	T5	T6	T7	T8	T9	T10
#Twitter earlier votes	352	414	58	80	50	181	135	141	140	40
#YouTube earlier votes	169	188	24	35	28	169	84	80	52	22
The ratio	2.08	2.20	2.42	2.29	1.79	1.07	1.61	1.76	2.69	1.82
	T11	T12	T13	T14	T15	T16	T17	T18	T19	T20
#Twitter earlier votes	480	155	177	48	107	45	181	61	48	42
#YouTube earlier votes	249	34	69	15	70	30	88	56	34	43
The ratio	1.93	4.56	2.57	3.2	1.53	1.5	2.06	1.09	1.41	0.98

3.3 Reaction to Certain Topics is Somewhat Topic-sensitive

In section 3.2, we also find the user vote ratio is different for different topics and for the topic "Skyfall" even the number of YouTube earlier votes is a little larger than that of the Twitter earlier votes. So it naturally comes to our mind that whether the temporal order between Twitter and YouTube has something to do with the topic category. Therefore, we choose 5 categories which have obvious category labels, i.e. celebrity, technology, movie, game, sport. We further calculate the average user vote ratio in each category and the result is shown in Table 4. We can see that the ratio on technology category is relatively large while the ratio on movie category is small instead. It may be partly due to the fact that users are more likely to discuss and get to know full comments on electronic products before they search for the related videos, meanwhile a lot of users may share their opinions on the newly-released movies after they have completely viewed it in YouTube. Therefore, the user reaction to different topic categories also follows different temporal patterns and we should take topic category into consideration when analyzing the user behavior on certain topics.

Table 4. The user vote ratio between Twitter and YouTube on different categories

Category	Celebrity	Technology	Movie	Game	Sport
The ratio	1.87	3.27	1.31	2.48	2.35

4 Applications

The user-oriented temporal analysis across different social media sites can facilitate a variety of applications, including but not limited to:

1. **Cross-network User Collaboration.** Since we have found that for certain kinds of topics users are very likely to first talk about them in Twitter and then go to YouTube to better understand the evolution of the topics, we can further design some time-aware cross-network collaboration applications based on the temporal analysis. For example, we have been working towards the topic of the personalized real-time video recommendation from Twitter to YouTube. We first capture the temporal interests of certain users and

what topics they currently pay attention to in Twitter, then we can utilize the user interest learnt in Twitter to help recommend the relevant videos to the same user for his YouTube account as the user attention in YouTube lag behind that in Twitter.

2. **Trend-aware Prediction of Content Popularity.** As we can identify the trending topics and track the temporal dynamics of the topic popularity in Twitter precisely, we can well capture the trend of the corresponding topics in real world. In general, the topic trend in YouTube somewhat follows a similar pattern but with some delay than that in Twitter. Therefore, We can utilize the topic popularity trend obtained from Twitter to help predict the trend of the same topic in YouTube.

5 Conclusion

In this paper, we present a measurement study of users' temporal behavior between Twitter and YouTube. In particular, we explore the temporal patterns of users' attention to certain topics on user level. It is shown that most users' attention to certain topics is earlier in Twitter than YouTube and the situation is topic-sensitive. Further work may include a quantitative analysis on the users' temporal behavior patterns across different social media networks and some novel applications based on our social behavior analysis.

References

1. Ahn, Y.Y., Han, S., Kwak, H., Moon, S., Jeong, H.: Analysis of topological characteristics of huge online social networking services. In: Proceedings of the 16th International Conference on World Wide Web, pp. 835–844. ACM (2007)
2. Lerman, K., Ghosh, R.: Information contagion: An empirical study of the spread of news on digg and twitter social networks. In: Proceedings of 4th International Conference on Weblogs and Social Media, ICWSM (2010)
3. Carmagnola, F., Cena, F.: User identification for cross-system personalisation. Information Sciences, 16–32 (2009)
4. Narayanan, A., Shmatikov, V.: De-anonymizing social networks. In: 2009 30th IEEE Symposium on Security and Privacy, pp. 173–187. IEEE (2009)
5. Abel, F., Araújo, S., Gao, Q., Houben, G.-J.: Analyzing cross-system user modeling on the social web. In: Auer, S., Díaz, O., Papadopoulos, G.A. (eds.) ICWE 2011. LNCS, vol. 6757, pp. 28–43. Springer, Heidelberg (2011)
6. Abel, F., Gao, Q., Houben, G.-J., Tao, K.: Semantic enrichment of twitter posts for user profile construction on the social web. In: Antoniou, G., Grobelnik, M., Simperl, E., Parsia, B., Plexousakis, D., De Leenheer, P., Pan, J. (eds.) ESWC 2011, Part II. LNCS, vol. 6644, pp. 375–389. Springer, Heidelberg (2011)
7. Roy, S.D., Mei, T., Zeng, W., Li, S.: Socialtransfer: cross-domain transfer learning from social streams for media applications. In: ACM Multimedia, pp. 649–658 (2012)
8. Becker, H., Iter, D., Naaman, M., Gravano, L.: Identifying content for planned events across social media sites. In: Proceedings of the Fifth ACM International Conference on Web Search and Data Mining, pp. 533–542. ACM (2012)

9. Osborne, M., Petrovic, S., McCreadie, R., Macdonald, C., Ounis, I.: Bieber no more: First story detection using twitter and wikipedia. In: SIGIR 2012 Workshop on Time-Aware Information Access (2012)
10. Yang, J., Leskovec, J.: Patterns of temporal variation in online media. In: Proceedings of the Fourth ACM International Conference on Web Search and Data Mining, pp. 177–186. ACM (2011)