Mainstream Media vs. Social Media for Trending Topic Prediction – An Experimental Study

Aleksandre Lobzhanidze, Wenjun Zeng, Paige Gentry and Angelique Taylor

Department of Computer Science
University of Missouri
Columbia, MO 65211

***Abstract*— In the recent years, we have witnessed social networks blossom. Social networking reshaped worldwide communication significantly increased the speed of news spread, and connected the world stronger than ever. Although social networking has been such a revolutionary invention for the society, and many researchers have turned towards social media to explore trending topics, mainstream media still remains as the origin of the majority of the news discussed in social networking sites. Social stream mining to make video recommendations based on the trending topics has been an active direction in the research community. Understanding the trending topics and its impact on video sharing sites is very interesting for network traffic engineers. Quality of service can be significantly improved if we can predict what kind of video content will generate large traffic. The focus of this paper is to study which type of media, mainstream or social, can contribute better towards identifying trending topics. We present the experimental study of the story development process in mainstream and social media based on the real-world data. The study helps us properly identify which media source is more appropriate for the video recommendation and network traffic prediction systems. Through our findings, we discovered mainstream media could significantly improve the trend detection.**

***Index Terms*— Mainstream media, social network, topic model, popularity prediction, video recommendation.**

I. INTRODUCTION

The aim of this paper is to study the influence of mainstream media over Internet users and provide input for video traffic prediction systems. In the digital age, human interaction with computers and browsing various news articles and short videos on the Internet is a daily business. Given the number of Internet users worldwide reached 2.27 billion. Internet reach-out to the public is becoming more and more significant. It is almost undisputed mainstream media greatly influences its audience and shapes their interests, world vision, political, and philosophical beliefs. Recent years have witnessed the blossom of social networks; people are more connected these days than ever, news can travel with the speed of light. Social networking websites speed up the news spread. Social network indeed possesses some unique features, for example we can extract public opinion about different controversial issues, occurrence of unexpected events, the degree of interest regarding different political parties, reality shows, movies, and many others. Due to the unprecedented growth of social networking popularity, people often overlook conventional media. In this paper, we would like to explore the opportunities provided by social media, and compare it with the mainstream media. Both types of media have their strengths and weaknesses. We weigh each and conclude which one can be utilized more effectively for a number of applications listed below. There are a number of reasons to extract trending topics from the media. Social stream mining to make video recommendations based on the trending topics has been one of the active directions in the research community [2][3][4]. Another interesting application is predicting network traffic based on the trending topics. It is common for people to browse videos about interesting topics they read in the news. Such a trend creates a wave of data traffic in the network. Being prepared for such waves could greatly help with providing quality of service.

In this paper, we try to understand which type of media is more appropriate to use for the purpose of video recommendation or network engineering, what kind of difficulties they expose under different circumstances and how we can address those issues. It is interesting to study which media has a greater influence over its audience, how they are inter-dependent, and which media is in the driving seat. In the paper we explore the strength of influence of mainstream media over Internet users, and compare it to the influence of social networks.

The rest of this paper is organized as follows. In section II we present some related work. Section III discusses the differences of social and mainstream media. Section IV outlines methodology of data collection and analysis. In Section V, we present experimental data to provide answers to our questions, and lastly we conclude the paper with Section VI and discuss some future work.

II. RELATED WORK

Mining textual data to predict the popularity of visual data is a relatively new field in the research community. [3] is an attempt to design a social video replication scheme based on traces from Weibo re-shares. This approach proposes to explore the geographic location of the Weibo users, and replicate the videos in the nearby local cache servers. [2] is another approach based on the Youku and Weibo traces to predict video popularity. [4] builds a common topic space between twitter and YouTube to propose a cross-domain video recommendation system. The majority of the works in this area rely on the data that might not always be publically available. Our study solely depends on publically available data.

978-1-4673-3133-3/13/$31.00 ©2013 IEEE

III. Motivation

Our work is based on a number of experiments that we conducted, and several interesting conclusions have been drawn. As mentioned in Section I, it is crucial to have first-hand access to trending topics, if we would like to understand which videos would become popular in the near future. Several prior works [3][4] have claimed social media hits the web faster than mainstream media. Throughout our experiments, we extracted data from social and mainstream media, and compared which one is ahead in terms of timing to detect news as trending. We also identified the origins of the news that attracted interest of large public.

Social media has many useful features, but it also imposes a number of difficulties that need to be dealt with. It is challenging to extract relevant information from social networking websites (Twitter, Facebook, Myspace, etc.) and correlate social media across different domains. The data from the social stream tends to be very noisy (unstructured, grammatically incorrect, misspelled) and significant efforts need to be made for data polishing. Another characteristic of social media is the short length of the messages, whether its a tweet, or status update, they often tend to be short. In fact, Twitter restricts users to 140 characters only. Given documents are short and noisy, large amounts of data are required to hatch something meaningful out from the social stream. Data arriving in high volume creates scalability issues, which additionally need to be addressed. Lastly, when speaking of social data, we have to realize not every social networking website is a broadcasting tool, e.g. Facebook, LinkedIn, YouTube provide security mechanisms that will restrict user profile page to be viewed only by selected users.

Mainstream media- Major news companies post articles related to the popular topics/events daily. In the age of social network, the news can spread very fast through the web and reach the farthest corners of the globe. Major credible news companies (CNN, BBC, AP, etc.) have always carried the slogan "give people what they want". Mainstream media is interested in reporting stories that attract readers; therefore, it is reasonable to assume news articles will reflect what's popular in the society. Unlike social data, news articles are structured documents, written with high grammatical accuracy. We also have to keep in mind some newspapers might only follow a limited number of topics (showing emphasis only in sports, showbiz, technology, etc), and present the story from a bias angle.

As we have drawn the differences and identified some strengths and weaknesses for both types of media, we have to decide which one provides better input for video popularity prediction schemes. The majority of prior works have proposed that social media is the tool to use. Regardless of its noisy nature, we can still assess some interesting information from the social stream. We would like to know how better, or worse it performs against mainstream media. Intuitively, social media has the ability of detecting trends, however in the mainstream media there are a number of trained professional journalists who specialize in identifying and reporting top news. We can certainly take advantage of this trend. Typically, videos related to stories discussed in mainstream media generate high traffic [4]. To the best of our knowledge, there is no real data available that would learn and compare the story development process in the mainstream and social media. Nor have we found a study that would suggest which type of media provides better input for predicting video popularity. The purpose of our experiments is to have a complete understanding of the popularity development cycle of the news in the media. For the research community, social media has been a very popular tool to design cross-platform video recommendation systems [4]. We feel mainstream media has been overlooked in this case.

IV. Methodology of Data Collection and Evaluation

Comparison between mainstream and social media is not straightforward; therefore, we have collected data in a similar fashion and used some normalization methods to keep things fairly balanced.

A. Data from Social Media

Twitter is one of the most popular micro blogging platforms. All data in Twitter is publically available, thus it was a natural choice to use this system as a data source. Twitter provides API services that allow downloading and searching tweets. Data polishing is an important piece in the process of mining twitter data. On average, about 140 million tweets are posted per day. We have selected 30 million random users from the Twitter website, the ones who have over 100 tweets, tweet at least 1.4 times per day, follow at least 10 other users, and have specified their location. Users are spread all over the world. . Data has been collected in a similar fashion to [3]. We have implemented crawler to extract data from the twitter website.

B. Data from Mainstream Media

In order to collect news articles from mainstream media, we have used Really Simple Syndication (RSS) feeds. The majority of news websites support RSS, thus the document in this case is a nicely formatted XML file that contains a short summery of the article, publication date, title, and the link to the original article. Unlike Twitter data, there's no need for data polishing. Most documents do not contain misspelled words or grammatically incorrect sentences. Similar to [1] we pile all necessary feeds into a text file and feed it to a topic modeler. From the selected news sources, on average 20 different sources might generate up to 1000 feeds per day. Selected RSS sources [7] vary from global news corporations to local scope newspapers. According to our collected data, websites such as CNN, BBC on average post 171 RSS feeds per day, while local scale news websites (such as columbiamissourian.com, columbiatribune.com) produce 10 to 15 feeds per day. We have collected all available RSS feeds for the year of 2012 starting January 1st ending August 27th. We selected global scope, as well as local scope news sources. Each source has dedicated RSS URL to a certain category.

There is a great difference in the process of composing a tweet and the news article. Besides the obvious differences (grammatical correctness, properly spelled words, length, etc.), news articles are written in a certain format. In the domain of journalism, the story typically opens up with the news, mid section presents important details, and the last passage summarizes the development of the story [8]. Tweets have no

pre-defined format; majority of them contain hash-words, and/or links (via URL shortener).

C. Trending Topic Identification

To identify popular topics in the social and mainstream media, we use latent Dirichet allocation (LDA) [1]. LDA is a generative model that allows sets of observations to be explained by unobserved group of variables that explain why some parts of the data are similar. The goal of topic modeling is to automatically discover the topics from a collection of documents. We feed RSS and Tweets to LDA to detect trending topics in the mainstream and social media. From the LDA popularity list we classify top 20% as trending [9]. News sources are broadly covering different topics/events throughout the world. Similarly the selected Twitter users are spread all over the world, discussing a large array of topics.

V. Experimental Results

In this section we present findings through experimental results. Data was downloaded and tests were produced on a hardware having Intel Xeon 2.93GHz processor, with 8GB RAM.

First, we explore how big the influence of mainstream media is over the social media. We have defined three types of events: **scheduled** event (e.g. sport competition, political debate, concert, etc.), **breaking news** (unexpected events), and news **first announced through social media** (e.g. Rafael Nadal tweeting he's not fit for the US open). The third types of events mostly come from the well-established celebrities. We have looked into the accounts of most followed users and manually extracted events that became trending topics later.

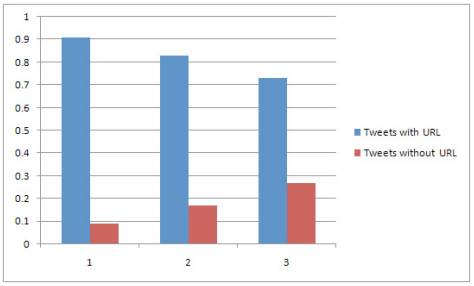

Figure 1. Percentage of tweets with URL vs. Percentage of tweets without URL

The experiment presented in Fig. 1 attempts to study the influence of mainstream media over social media. We select 3 types of events, with 20 different stories for each type. The tweets are extracted within 7 hours after the event occurs. We check how many tweets contain reference URL links to some news websites and plot the results. Due to the large amount of Twitter data, we have only extracted 1500 random tweets for each story, and identified how many of them contained URL. The findings suggest the majority of the tweets are referring to the news sources.

Our second experiment studies the speed of spread. We chose a trending topic from the London 2012 Olympic games. US Women's soccer team won the 4th Olympic gold against Japan (Japan had defeated US in the FIFA 2011 Woman World cup). The event provoked many tweets and articles. We extracted tweets from the users who have shown interest in soccer and Olympics.

The set of users were selected from the pool of 30-million user database that we created as described in section IV. Similarly, we have selected RSS feeds from the news sources that specialize in sports. From Fig. 2, we can observe a sudden jump in RSS feeds. The first feed was observed at the beginning of the first half of the game. Tweets won't appear until the middle of the first half. With a 2 minute interval, we check how many videos related to the event were uploaded to YouTube and plot results with a red dotted line.

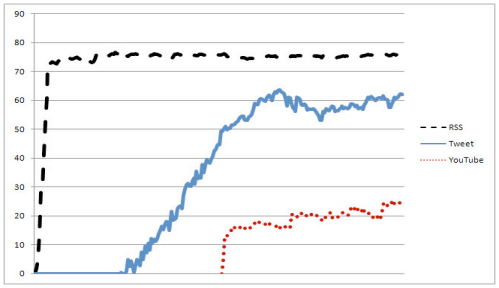

Figure 2. Percentage of documents related to the topic (US Women's soccer team winning the 4th Olympic gold) vs. Time (2 mins per unit, local kick-off time 19:45. 2-hour stretch)

The next experiment presents the process of a story development in the mainstream and social media. We chose breaking news "Empire state building shooting" [5] that took place August 24, 2012. Such events are interesting to observe, since they become a big hit on social video sharing websites.

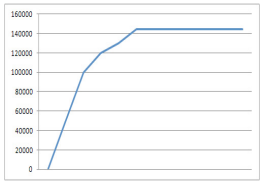

Figure 3 (a). Empire state building gunmen shot. View-count statistic with 5 hour stretch. View-count vs. Time

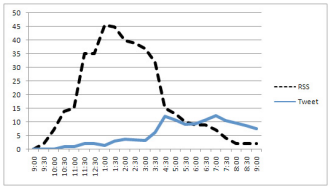

Figure 3 (b). Percentage (Empire state building shooting) vs. Time

We have been monitoring video view-count development, with a ten-minute interval after it has been uploaded to YouTube, as shown in Fig. 3(a). The video shows raw footage of the gunmen shot dead by police after the Empire State Building shooting. The largest jump was observed within the first 5 hours after the video was uploaded. Since then, the view-count grows very slowly. This event could qualify as a "Mushroom" story, where most of the attention is grabbed within first few hours, and the popularity is lost as fast as it has been gained. We would like to compare the view-count development of a story, to the popularity development in the media. Figure 3(b) lays out the details. As we can observe from Figure 3(b) there is a pyramid shape development of the story in mainstream media. We have evaluated RSS feeds from 9:00 AM August 24, 2012, until 9:00 PM the same day. [7] shows the list of all RSS sources used for this experiment. In Twitter we have a database of 30 million users, all tweets

posted by these users on August 24, 2012, between 9:00 AM and 9:00 PM have been collected. We feed the data from the news sources to the topic modeler, and plot the pattern of popularity. Similarly, with Twitter data, we send all tweets collected to LDA in order to compare the story development process with mainstream media. The percentage of the story is drawn against time. In the twelve-hour stretch, we can see the story climaxes in mainstream media sometime in the afternoon, while in the social media the story becomes popular later. The pattern observed hints that mainstream media would identify a topic as trending sooner.

We observed another unexpected/unforeseen event on August 19, 2012. Controversial comments on rape and pregnancy from Republican Party member Todd Akin [6] provoked many discussions among Internet users. In YouTube, the video was a big hit, a number of users even composed new songs with lyrics dedicated to this event. We feel such unexpected events are good examples for observing the story developing process in the mainstream and social media.

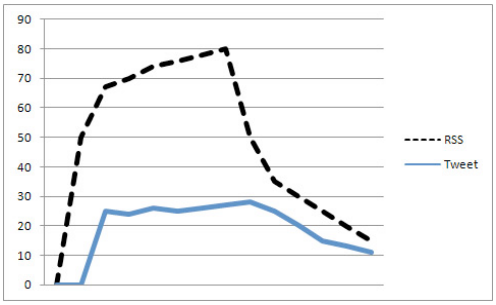

Figure 4. Percentage of story in media (Todd Akin's remarks) vs. Time (one week stretch)

Figure 4 represents a seven-day stretch, starting August 19, 2012. Mainstream media was the first to report the story. If we were to identify this event as quickly as possible, mainstream media would contribute better than social media. However, we believe social network has greatly boosted the view-count increase of YouTube videos that were related to this story.

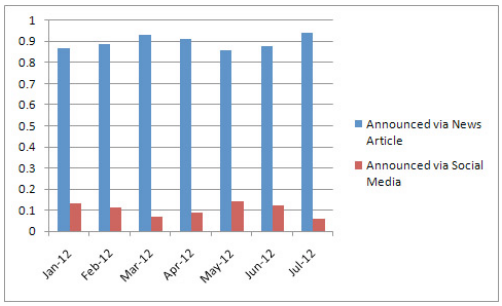

Figure 5. Origins (%) of news

Lastly, we investigate the origins of the news. We ran LDA topic modeler on the Twitter dataset only, and identified ten popular topics for every month from January to July in the year of 2012. We plotted the average percentage of the news origins. The majority of the events discussed in social networks have been first announced in the mainstream media, as shown in Fig. 5. From the experimental results presented above, we can conclude if timing in detecting the news as trending is important, traditional news sources might be a better choice than social media. Certainly there are cases where social network is irreplaceable, however we are only targeting video recommendation, or network traffic prediction systems.

From the charts, we observe pyramid shape in mainstream media meaning trend detection might be a lot faster. News eliminates large amounts of pointless babble that is very typical for social media. The amount of news articles required for processing is relatively small compared to tweets. In addition, we can skip data polishing, since the majority of articles are well structured and follow certain format.

VI. Conclusion

In this paper, we present experimental results based on the real data that show the story development process in the mainstream and social media. It has not been deeply studied how to make a choice between these two when it comes to cross-domain application systems that learn trending topics from textual data to make a prediction for the video portals. Social networks have significantly reshaped the news consumption among web users by speeding up the news spread. Even though social network has gained unprecedented popularity, it is still reasonable to consider the mainstream media as the primary source to make predictions for video popularity. Data from social networks tends to be noisy, and comes in large volume. Mainstream media is mostly structured. On top of that, journalists are trained to learn, explore, and present information in a laconic way. By using mainstream media, we can remove data filtering, and use a smaller amount of data for trend detection. Our future work includes designing an application for a video recommendation system that will learn trending topics from mainstream media and make predictions of video popularity for the social video portals.

Acknowledgment

This research is supported in part, by the National Science Foundation under the award CNS1004606, and in part, by a Mizzou Advantage seed fund.

References

[1] M. Hoffman, D. Blei, and F. Bach, "Online learning for latent dirichlet allocation," *Nature*, vol. 23, pp. 1–9, 2010.

[2] Z. Wang, L. Sun, C. Wu, and S. Yang, "Guiding Internet-Scale Video Service Deployment Using Microb log-Based Prediction," *IEEE INFOCOM* 2012.

[3] Z. Wang, L. Sun, X. Chen, W. Zhu, J. Liu, M. Chen, and S. Yang, "Propagation-Based Social-Aware Replication for Social Video Contents," *ACM Multimedia* 2012.

[4] S. D. Roy, T. Mei, W. Zeng, and S. Li, "Empowering Cross-domain Internet Media with Real-time Topic Learning from Social Streams," *Proc. of* 2012 *IEEE Inter. Conf. on Multimedia and Expo*.

[5] http://en.wikipedia.org/wiki/2012_Empire_State_Building_shooting

[6] http://en.wikipedia.org/wiki/Todd_Akin

[7] https://sites.google.com/site/ccnc2013

[8] www.nychsj.com/resources/General+Rules+to+Follow.pdf

[9] Asur S, Huberman BA, Szabo G, Wang C (2011) "Trends in social media: persistence and decay" *Technical Report HP Laboratories 2011*