How to detect information outbreaks in Twitter

Thibault Lasserre
Illinois Institute of Technology
Computer Science
tlasserr@hawk.iit.edu

1. INTRODUCTION

Everyone is receiving a tons of information on an unparalleled scale. So, it becomes very difficult for everyone to extract the relevant information., the one that will be the most quickly helpful. One way to resolve this dilemma should be to keep one's eyes on some trusted influential people in the diffusion network. Thus, if a person can keep himself/herself updated with the information from these sources, he/she will be eager to know every ongoing hot topic in the world. So, the goal of this paper is to present a method to detect these influential nodes. In that way, we can predict which tweet will be widely spread by focussing on influential nodes and their content-related information. This paper examines how to perform topic modeling on tweets. Indeed, we give to a Twitter user a distribution of topic. However, as we have only 11M tweets and the social graph we study has 41M users, we assume that a tweet we find for a user will give to that user its topic distribution. Moreover this paper compares three different measures of influence: indegree, retweet and mention. It provides a better understanding of the different roles users play in social media. Indegree represents popularity of a user; retweets represent the content value of one's tweets; and mentions represent the name value of a user.

2. DATA

Different kind of data have been used in this project to answer the problem. First to perform topic modeling, a dataset provided by Dr Aron Culotta has been used. This dataset is composed of about 11 Millions tweets collected between September 2009 and May 2010 that covers a large number of topics. Before applying some algorithm to this tweets, stopwords are removed.

Moreover, in this project, the Twitter social graph data from H. Kwak, C. Lee and S. Moon is used. This directed graph contains 41 million users and 1.5 billion edges which represent social relations and a follower-following topology. It has been collected in July 2009.

We also use The Higgs dataset which has been built after monitoring the spreading processes on Twitter before, during and after the announcement of the discovery of a new particle with the features of the elusive Higgs boson on 4th July 2012. The messages posted in Twitter about this discovery between 1st and 7th July 2012 are considered.

The four directional networks made available here have been extracted from user activities in Twitter as

- ≅ re-tweeting (retweet network) (425008 nodes and 733647 edges)
- ≅ replying (reply network) to existing tweets (37366 nodes and 30836 edges)
- ≅ mentioning (mention network) other users (302975 nodes and 449827 edges)
- ≅ friends/followers social relationships among user involved in the above activities (456631 nodes and 14855875 edges)

It is worth remarking that the user IDs have been anonymized, and the same user ID is used for all networks. This choice allows to use the Higgs dataset in studies about large-scale interdependent/interconnected multiplex/multilayer networks, where one layer accounts for the social structure and three layers encode different types of user dynamics .

3. METHODS

This project implies to different kind of data, i.e. some tweets or some social graph. The tweets are used to perform topic modeling whereas the social graphs are used to analyze influence network in Twitter. In this section, we will analyze and describe these different methods used in this project, about topic modeling and Influence analysis.

3.1 Latent Dirichlet Allocation

Topic models provide a simple way to analyze large volumes of unlabeled text (11M tweets). A "topic" consists of a cluster of words that frequently occur together. Using contextual clues, topic models can connect words with similar meanings and distinguish between uses of words with multiple meanings.

Thus, to perform topic modeling in this project, Latent Dirichlet Allocation (LDA) is used. LDA is such a probabilistic model for uncovering the underlying semantic structure of a document collection based on hierarchical Bayesian analysis. The original idea of LDA is to model documents as if they arise from multiple topics, where each topic is defined to be a distribution over a fixed vocabulary of terms.

LDA assigns a document with a distribution of multiple topics. The basic pipeline of LDA is linearly scanning

through each word in the document, and initializes by randomly assigning some topics to each word. Afterwards, LDA goes through a iterative improvement process (like Expectation Maximization) until it converges. For each word, it computes the probability of topic given document and word given topic. Then we reassign each word to a new topic based on the maximum likelihood estimation from the previous iteration. Finally, this process converges and each document is assigned a distribution over the latent topic space.

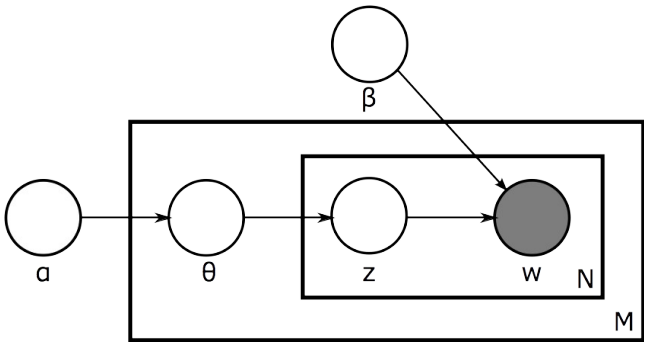


Figure 1:Latent Dirichlet Allocation diagram in plate notation

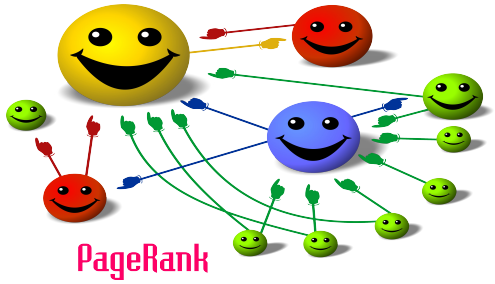
With the plate notation of Figure 1, the dependencies can be captured concisely. The boxes are "plates" representing replicates. The outer plate represents documents, while the inner plate represents the repeated choice of topics and words within a document. M denotes the number of documents, N the number of words in a document. Thus:

- α is the parameter of the Dirichlet prior on the per-document topic distributions
- β is the parameter of the Dirichlet prior on the per-topic word distribution
- Θ_i is the topic distribution for document *I*
- Φ_k is the word distribution for topic *k*
- Z_{ij} is the topic for the *j*th word in document *I*
- w_{ij} is the specific word.

3.2 Measuring Influence

Influence is "the power or capacity of causing an effect in indirect or intangible ways." Despite the large number of theories of influence in sociology, there is no tangible way to measure such a force nor is there a concrete definition of what influence means, for instance, in the spread of news. In this project, Twitter is modeled as a network, the mathematical term for which is a "graph". So, one of the first question to ask then is "which parts of my graph dataset are the most important?"> Before one can investigate how Twitter users become influential, one has to find who the most influential Twitter users are in the first place.

A well-known algorithm for finding the most important nodes in a graph is called PageRank. PageRank is a link analysis algorithm, named after Larry Page and used by the Google web search engine, that assigns a numerical weighting to each element of a hyperlinked set of documents, such as the World Wide Web, with the purpose of "measuring" its relative importance within the set. The algorithm may be applied to any collection of entities with reciprocal quotations and references. A PageRank results from a mathematical algorithm taking into consideration authority hubs. The rank value indicates an importance of a particular node in the graph. An edge to a node counts as a vote of a support. The PageRank of a node is defined recursively and depends on the number and PageRank metric of all nodes that link to it ("incoming edge"). So, a node that is linked to by many pages with highPageRank receives a high rank itself. If there are no edge to a node, then there is no support for that node.



4. EXPERIMENTS

4.1 Topic Modeling

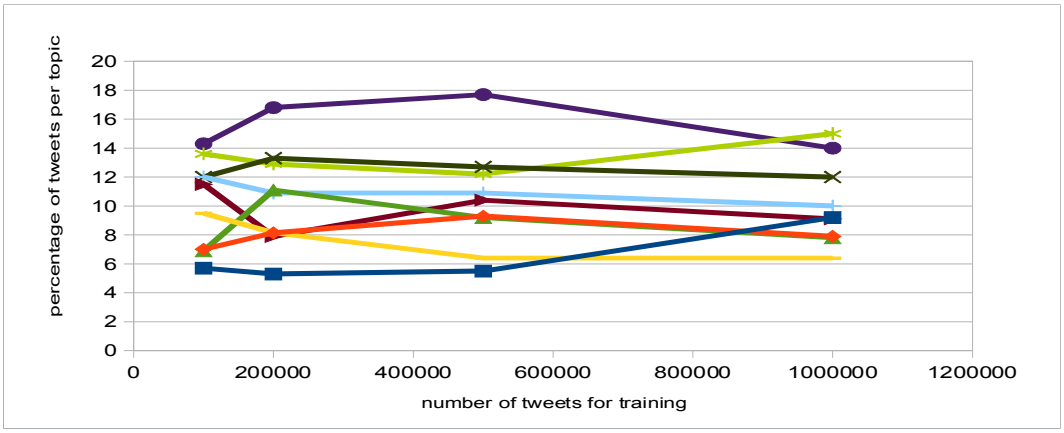
First, I processed the tweets data from Dr Aron Culotta with LDA from MALLET package, which assigns each documents (tweets) a topic distribution. Only 10 topics were selected. To train the model, four tests have been done with 100 000, 200 000, 500 000 and 1 000 000 tweets. It was difficult to go further because of the computational resources. Indeed, one test with 2 000 000 tweets did not finish after 5 hours of computation.

As result of these tests, I obtained a list of ten topics. Below is some example of topics with the list of the most important words:

topic 1: love show watching game watch wait awesome music song big tv playing amazing play movie guys ll hard tonight

topic 2: good day today time work great night home back tomorrow morning hope week fun long weekend house ready tonight

topic 3: make happy life world year money won give birthday making car football times found book buy season part team



We can verify the efficiency and also the good distribution of the topics with the graph below:

Each color represents one cluster. So, for example, for the purple line, with 100 000 tweets for train, around 14% of the total tweets were in this cluster.

This graph shows an uniform repartition of topics among the tweets.

4.2 How to analyze influence?

In this paper, the Twitter network is analyzed as a news spreading medium and study the types and degrees of influence within the network. Focussing on an individual's potential to lead others to engage in a certain act, three "interpersonal" activities on Twitter are highlighted. First, users interact by following updates of people who post interesting tweets. Second, users can pass along interesting pieces of information to their followers. This act is popularly known as retweeting. Finally, users can respond to (or comment on) other people's tweets, which we call mentioning. So, These three activities represent the different types of influence of a person:

1. The number of followers of a user directly indicates the size of the audience for that user.
2. The retweet influence, measured through the number of retweets containing one's name, indicates the ability of that user to generate content.
3. Mention influence, measured through the number of mentions containing one's name, indicates the ability of that user to engage others in conversation.

4.2.1 Finding the top influentials

First, the PageRank algorithm has been run on the Twitter social graph to find the 20 most influential users. Below is the result of the algorithm:

813286 => @BarackObama
14224719 => @Number10gov
15131310 => @WholeFoods
31567254 => @RiotNome
16409683 => @britneyspears
14075928 => @TheOnion
7040932 => @tonyhsieh
12687952 => @BJMendelson
5380672 => @threadless
26784273 => @charitywater
17850012 => @Radioblogger
5741722 => @nprpolitics
6149912 => @LiveEarth
5210841 => @ScotMcKay
20385404 => @
14389132 => @nansen
6449282 => @JetBlue
12589972 => @xavierlur
14669398 => @BradHoward
14994465 => @jimmyeatworld

We can see that all these users are among the users with more than 10 000 followers. The most followed users span a wide variety of public figures and news sources. They were news sources, politicians (Barack Obama), celebrities (Britney Spears) and bloggers. As the list suggests, indegree measure is useful when we want to identify users who get lots of attention from their audience through one-on-one interactions, i.e., the audience is directly connected to influentials.

4.2.2 Comparing three measures of user influence

To make some correlation between the 3 ways of influence, the PageRank algorithm has been applied to the 4 graphs of the Higgs dataset i.e. the social, retweet, reply, mention graphs.

The tables below show the percentage of the same nodes which are in the 20 most influential users of two graphs. But not essentially at the same rank.

	Social	Retweet	Reply	Mention
Social	-			
Retweet	15.00%	-		
Reply	15.00%	10.00%	-	
Mention	10.00%	30.00%	10.00%	-

We can see thanks to this table that by comparing the 20 most influential nodes of two ways of influence, we find a few nodes in common. Indeed the maximum of nodes in common are between the mention and retweet graphs with 30% of the nodes are the same in the two graphs. Otherwise , the results are about 10-15%. In the table below, we try to merge the graphs to see if there are better correlation:

	Retweet + Reply	Retweet + Mention	Reply + Mention	Retweet + Reply + Mention
Social	15.00%	5.00%	10.00%	10.00%

As we can see, merging the graph implies to have worst correlation. So, we can think social graph is not really the best way to analyze influence in Twitter, because the similarities between these ways of influence are very low. Retweets are trackers of trending topic and knowledge-able people in different fields, whom other users decide to retweet. Unlike indegree, retweets represent influence of a user beyond one's one-to-one interaction domain; popular tweets could propagate multiple hops away from the source before they are retweeted throughout the network. Mentions represent a public response to another user's tweet—the focus of a tweet is on content for retweets, while the focus is on the replied user for mentions. Thus, one way to regroup these graph and have better correlation should be to analyze distribution of topics of each user. Those who have the same distribution of topic and are linked by the friend/follower relationship are more likely to spread the same information.

5. FUTURE WORK

This project is not really "finished". The LDA has been used for topic modeling. The 20 most influential nodes of Twitter have been found tanks to the PageRank algorithm and 3 influence ways have been analyzed to explain why it is relevant to integrate distribution of topics in the diffusion model. But the way to integrate the distribution of topics in the diffusion model has not been implemented, but it will be explain in this part.

To take into consideration the distribution of topic, the idea was to modify the Twitter social graph. Indeed, in this graph, all the edges represent a friend/follower relationship. The idea was to remove all the edge between two users who do not have the same distribution of topic. In that way, in this graph, a node will be only linked by his followers with the same distribution of topic as himself. So, running the PageRank algorithm on this graph will compute the top influential nodes taking into consideration topic distribution. We can also think that the result will be closer than the 3 graphs of retweet, mention and reply.

An other way to improve the model is an other way to cluster the users once the topic modeling is done. In this project, a user was put in the cluster of his main topic. For example, if node A was related to football and node B to baseball, node A and node B were in two different cluster. But, the LDA affect to a user a distribution of topic. So, one other way to consider the clustering is to apply the K-means algorithm to all the user with as dimension, the number of topic. So that, a user will be considered as a point in a N dimension space (N number of topic), with as coordinates the value of the distribution of each topic. This will allow our to nodes A and B to be in the same topic maybe. This algorithm has been implemented but not tested.

Moreover running LDA necessitates a lot of computational resources. We could not train our model with more than 2 million tweets because of the execution time. One idea is to use Amazon Web Service to use more data.

6. RELATED WORK

Several recent effort have been made to track influence on Twitter. The paper [1] presents learning to detect outbreaks. It studies how to incorporate content-related information into normal outbreak detection. However, it uses two basic diffusion models: independent cascades and linear threshold. [3] measures dynamics of user influence across topic and time. Finally, [4] determines a grounded approach for measuring social networking of individual Twitter user.

7. CONCLUSION

In this project, topic modeling and influential analysis with three different measures were performed. This last shows that an analysis only with the Twitter social graph is not enough to measure influence because a few top users of the social graph are among top users of the other graphs. But we can think that adding topics of users will improve the way to find the top influential nodes and to detect information outbreaks.

8. REFERENCES

[1] Jiayuan Ma, Xincheng Zhang. Lerning to detect information outbreaks in social networks

[2] David Kempe, Jon Kleinberg, Eva Tardos, Maximizing the spread of influence through a social network.

[3] Meeyoung Cha, Hamed Haddadi, Fabricio Benevenuto, Krishna P. Gummadi, Measuring user influence in twitter: The Million Follower Fallacy

[4] Isabel Anger, Christian Kitti, Measuring influence on Twitter