TwitNews: Microblog Ranking Using Structured Social Data

Chris Moghbel, Lei Jin
Dept. of Computer Science, UCLA
cmoghbel@cs.ucla.edu, rickyjin@cs.ucla.edu

Abstract

The rise in popularity of Microblogs in the recent years, and in particularly Twitter, has contributed in generating vast amounts of raw data every day. This data includes real time breaking news in the midst of noisy spam and advertisements. However, Twitter provides various social mechanism that allows users to distinguish who and what they find important. Users can "follow" other users given a topical interest in their comments or tweets. They can also share information they find interesting by "retweeting" or mentioning a users tweets to their followers. Many users also explicitly include shortened URLs in their tweets to contribute additional information relevant to the given subject. In this paper, we will show how these relationships can be used to distinguish between tweets that are newsworthy and tweets that are not. We propose a ranking algorithm that takes advantage of these social mechanisms to surface relevant information from amongst the noise. In our preliminary experiments on real Twitter data, our proposed ranking algorithm gives promising results in delivering quality news from raw tweets.

1 Introduction

The recent rise in popularity of Microblogs has changed the way Internet users interact and vastly increased the speed at which data travels across the web. Twitter is the website that introduced the concept of microblogging and helped to make it popular. Originally, Twitter was intended as a way for users to share short 140 character snippets with their "followers", or other users who had agreed to see these "tweets" from the user. However, Twitter rapidly evolved to be a service about much more than just sharing inane details of people's lives.

One of the first indications of Twitter's evolution was its role in the Iranian election protests in 2009. With the Iranian government censoring more traditional forms of communication, citizens tweeted their opinions and images of the protest movement that spawned after many have thought incumbent President Mahmoud Ahmedinijhad had illegally rigged elections in his favor. In addition, Iranians used Twitter to coordinate new protests even after other means of communication had been cut off. Within minutes of events happening within the country, people across the world were aware of the actions happening in Iran. Twitter became so important to the protest movement that the site actually delayed preplanned service outages so it could stay operational for Iranian citizens.

Since the Iranian protests, the spread of news through Twitter has skyrocketed. Every major news outlet has a presence on Twitter, as do most celebrities. People often find out about breaking news from their Twitter streams before they read about it anywhere else on the web. Twitter even surfaces the current top 10 "trends", or most commonly tweeted about topics, as streams that users can access to see what is currently popular.

However, the Twitter stream has very few tools for aiding discovery of relevant information from these popular topics. Tweets in any stream are presented chronologically rather than ranked according to relevance or importance. Additionally, the majority of tweets are essentially random noise when viewed from the perspective of a user with the task of finding relevant information. Despite the myriad of meaningful ways in which people now use Twitter, the majority of tweets are still meaningless tidbits such as "going to sleep" or "heading to campus."

In this paper, we address the challenges of extracting useful information from the vast amounts of data created by Twitter and its users. How can we determine what topics are relevant or popular? Among these topics, how can we filter out the noise and surface the information that is most relevant? We propose TwitNews, a system for ranking tweets based on the structured so-

cial information that is inherent in all Twitter data.

We will start in section 2 by providing a brief background on the Twitter structure that is applicable to the TwitNews algorithm. We will then talk about our code framework and how we obtained our data set, before presenting the metrics upon which we will evaluate our work. In section 3, we will go into more detail about the TwitNews ranking algorithm. In section 4, we will present out experimental setup and results. Finally, we will present our conclusions along with related and possible future work in sections 5 and 6.

2 Framework

In this section, we will start by giving a brief background on Twitter terminology for those unfamiliar with Twitter. We then describe our code architecture followed by the download policies and implementations of our crawlers. We finish the section by outline the method of evaluation we will be using for our experiments.

2.1 Twitter Background

For those unfamiliar with Twitter, we will provide a brief run-through of its terminology in order to make further discussions more clear and precise.

Twitter **trends** are algorithmically generated topics that are determined to be tweeted more frequently in the current interval of time than previously. Essentially, trends are the breaking news unfolding around the world in real time. Not only is popularity a factor in determining a new trend, but also its novelty compared to the past history of tweets.

Tweets are simply 140 character long strings that allow users to express brief comments to the public whenever and wherever there is access to Twitter, whether from a smartphone, via S.M.S. or from the browser on a PC.

Retweets are a feature on Twitter that allows a user to share a tweet that the user deems interesting with their followers. Retweets are generally specified by the starting tag **RT**. While most retweets follow this guideline, some do not, making it hard to determine the exact number retweets for any given tweet. Also, Twitter's API currently only shows a retweet count with a maximum of 101 retweets per tweet.

Hashtags are "#" symbol initiated keywords that mark specific topics in a Tweet, in order to categorize messages. This feature allows users to more easily search for relevant tweets with such hashtags later if they wish. If a certain hashtag becomes frequently tweeted, it may turn into a trending topic. However, since there are no limits on how many and where hashtags are located in tweets, they can be spammed with impunity.

Twitter **verified users** is a new verification service provided by Twitter for well known figures or organizations. To become a verified user, Twitter manually verifies certain user accounts as authentic and trustworthy. Verified users are designated by a corresponding badge on their Twitter profile.

2.2 Code Architecture

In order to build a data set to run our algorithms on, we used a MySQL database as persistent storage for information on trends, tweets, users, and rankings. The various Twitter crawlers we implemented download tweet, user, and trend data, which they then persist to the database in a batch after 250 tweets are collected. We will discuss the Twitter API in further depth later on in this section.

We periodically save a snapshot of the database as a SQL script. This allowed us the the freedom to reload the data set at any time in case of data corruption or in order to re-run the algorithm. After a sizable chunk of data is stored, usually at least several thousand tweets spanning a dozen trends, we ran our ranker algorithms on the database to pull out the top ranked tweets given a particular trend. This preliminary ranking was initially used to test our algorithm's capability in avoiding spam tweets.

In addition to crawling trends and their corresponding tweets, we crawled news tweets from @breakingnews, a popular Twitter account dedicated to providing breaking news. We were able to reuse our persistence code and existing database in order to store and use this information. After the same process of downloading a large chunk of tweets and storing into MySQL, we used a Boolean model to generally cluster similar news topics, and ran our modified algorithm on the resulting set to pick out the "interesting" tweets. Our algorithms are explained further in section 3.

2.3 Crawler Download Algorithms

One of the major components of the TwitNews application is the crawler responsible for acquiring data from Twitter. To do this, it uses API calls made public by Twitter to download trend, tweet, and user information. For the first iteration of our crawler, we attempted to use the REST API provided by Twitter. The REST API[1] has the advantage of allowing for more precise and specific queries for data to be made. For instance, there exists an API call that returns current tweets for a specific

Twitter trend. However, the REST API is rate-limited by Twitter, allowing each registered application to make a maximum of 350 API calls per hour. Also, each call is limited in the quantity of data it returns. For example, the API call to return most recent tweets for a trend returns a maximum of 100 tweets. Thus, it is impossible to build a data set of significant size in a reasonable amount of time using the REST API.

For the next phase of the crawler, we decided to use the Twitter provided STREAM API. In contrast to the REST API, the STREAM API[1] is not rate limited. However, the functionality it provides is not as extensive as the REST API. Instead of having specific API calls, the STREAM API provides a single stream of tweets from Twitter. In addition, this stream can be filtered by number of keywords provided to the API. Thus, to implement the next phase of the crawler, we began by getting the current trends from Twitter via the REST API, which takes one API call. We then iteratively provided each trend as sole keyword to the STREAM API in order to get a stream of tweets matching that trend. After obtaining a certain amount of tweets for a trend, the stream was restarted with the next trend as a keyword. This method of crawling allowed us to obtain a much larger dataset than the solely REST based API. However, we were still only able to obtain a maximum of a few thousand tweets per trend in a reasonable period of time. Another problem with the STREAM API is that applications are by default provided "sample" level access to the stream. In essence, this means that the stream can only select tweets from approximately 10% of all tweets happening on Twitter. Access to higher levels (50% "Gardenhose" access, or 100% "Firehose" access) require special permission from Twitter in order to be used by an application. Thus, we were still left looking for a way to generate larger data sets.

Our next goal was to automate the crawler to run continuously, rather than for a set period of time. As with the previous phase, the first step retrieved the current trends from the REST API. However, we then set all 10 trends as filtering keywords for the STREAM API. Then we looked at each tweet crawled in that one hour time period, and parsed the keywords in order to determine which trend it belonged to before persisting the tweet to the database. After this one hour time period, we retrieved the updated trends from the REST API, and reset the STREAM with the new trends as the tracking API. Run in this fashion, the crawler could persist upwards of 1 million tweets in a 24 hour time span. We entitled this phase of the crawler the "Trend Crawler", and we will refer to it as such throughout the rest of this paper. This crawler was used to generate the data set of Twitter trends with corresponding tweets.

In addition to the Trend Crawler, we developed another crawler, which used the Twitter account @breakingnews as its seed instead of Twitter trends. We named this crawler the "News Crawler", and we will refer to it as such throughout the rest of this paper. Similar to the Trend Crawler, the News Crawler ran continuously, monitoring the Twitter STREAM API, and refreshing its keyword tracking set every hour. However, the keywords were generated by downloading the most recent tweets from the @breakingnews Twitter account, a verified Twitter account dedicated to collecting the most relevant world news stories. Each tweet was then parsed into keywords by extracting the proper nouns from the tweet text. These keywords were then used to filter the Twitter stream. Also, an inverted index was built, mapping each keyword parsed from the set of @breakingnews tweets to the set of tweets that contained that keyword. This inverted index was later used to match each tweet downloaded from the Twitter stream to the closest matching news story obtained from @breakingnews. This matching was done via a simple Boolean model that counted the number of matching keywords between the tweet obtained from @breakingnews, and the tweet downloaded from the Twitter stream. This News Crawler allowed us to build up a section test set, consisting of tweets we knew to be mostly related to actual news topics. In addition, many tweets seen in the tweet stream for the news crawler did not have matches with the original tweet set downloaded from @breakingnews. Thus, for the dataset obtained via the news crawler, it was necessary to combine our ranking algorithm with a more traditional information retrieval method, in this case the Boolean model.

2.4 Evaluation Methods

By the nature of the problem, determining relevance for a document is very subjective. This is especially true in the context of Twitter, as there are many factors in a very limited context of 140 characters that might make a tweet relevant to one person, but not another. Because of this, it was difficult to determine an objective method for evaluating the results of our algorithms.

One idea with which we experimented was ranking by looking at news stories on more traditional sources, such as CNN.com or Google News. However, there were a few problems with this approach. First, this would have been a difficult task to accomplish in an automated fashion. One of the important parts of breaking news on Twitter is that the news spreads before it reaches more traditional sources. This means that we would have had to crawl these sources asynchronously in order to find the matching stories. Also, we found that

many Twitter trends do not correspond specifically to a specific news story, but are instead random topics that become popular among Twitter users. Finally, many of these traditional sources do not provide an API, which would have made implementing a crawler more difficult. Due to these problems, we decided not to use this as a method of evaluation.

Eventually, due to the difficulties of determining a better, more objective method of evaluation, we decided to do manual evaluation for each policy. For the analysis of our unaltered ranking algorithm for ranking tweets for Twitter trends, we look at the top 10 tweets specified by our ranking algorithm, versus the top 10 tweets as ranked by most recent timestamp, which is how tweets are presented on Twitter.com. For the analysis of the data set obtained via the news crawler, we compare three scenarios. The first is again the top 10 tweets by most recent timestamp, as the tweets would be presented on Twitter.com. The second is the top 10 tweets as ranked by a simple Boolean matching model. The third is the top 10 tweets as ranked by a hybrid of the Boolean model, and our ranking algorithm. While we acknowledge that our analysis will be subjective, we believe that the results are drastic enough that most readers will at least agree with the general theme of our analysis.

3 Ranking Algorithms

In this section, we go into detail about the TwitNews ranking algorithm. We start by describing the stand alone algorithm, as it was used to rank the tweets we downloaded in correspondence with Twitter trends. We will then describe the modified algorithm we used to perform search using the TwitNews ranking algorithm.

3.1 Ranking Tweets for Twitter Trends

The end result of the TwitNews ranking algorithm is a score for each tweet that indicates how interesting or relevant that tweet is. The ranking score produced is an integer scaled score that ranges from 0 to positive infinity. The ranking algorithm for Twitter trends is comprised of two components, the TwitNews ranking algorithm and our spam score. Our ranker runs through only unique tweets that do not have identical text, with duplicate tweets filtered out from the database. Here, we mathematically define the formula we call the "TwitNews ranking algorithm":

$$R(T_i) = \alpha \cdot \log\left(\frac{\sum rt_i}{RT}\right) + \beta \cdot v_i + \gamma \cdot l_i + \delta \cdot \log\left(\frac{f_i}{F}\right) \quad (1)$$

rt_i = A retweet of T_i
f_i = # of followers of tweeter, at time of tweet
RT = max # of retweets, limited to 101 by API
F = # followers of most followed Twitter user
$\alpha + \beta + \gamma + \delta = 1$

$$v_i = \begin{cases} 1 & \text{if the tweeter is verified} \\ 0 & \text{otherwise} \end{cases}$$

$$l_i = \begin{cases} 1 & \text{if the tweet contains a link} \\ 0 & \text{otherwise} \end{cases}$$

The TwitNews ranking score is computed by the various positive factors that contribute to a healthy discussion of a certain topic. These factors include: retweet counts as votes of confidence, outside URL links inside tweets signifying other forms of image or video media, whether a user is trusted, and whether a user is popular with many followers. We normalize the retweet count by dividing by the actual number of retweets per tweet by the maximum retweet count as provided by the Twitter API, or 101. To note, this API limit hampers the accuracy of the ranking algorithm, as we cannot calculate different retweet scores between tweets with more than 101 retweets. A bonus is given if a link is present within the tweet. Also, a greater bonus is given if the author of the tweet is a verified user. The user popularity score is determined by the author's number of followers. Again, we normalize this by dividing by the follower count of the most popular user on Twitter. Additionally, in order to minimize the scale differences, we take the log of both the retweet score and the popularity score.

Each score is given a weight, denoted in the TwitNews ranking algorithm by the greek letters alpha, beta, gamma, and delta. Through our intuition, along with trial and error, we determined values for each of these weights to each of the raw scores accordingly: retweetFactor (alpha) = 0.3, linkFactor (gamma) = 0.1, trustFactor (beta) = 0.4, followerFactor (delta) = 0.2. We give higher weights to the retweet score as retweets have been shown as a good indicator of topic relevance [10]. We also give a high weight to tweets from verified users, a decision inspired by the TrustRank [7] variation of the PageRank [3] algorithm.

We also compute and factor in a spam score using a preloaded lists of adult swear words and common advertisement words that have been compiled by other researchers on the web [6] [9]. The Spam score value is essentially a penalty or negative weight that brings down a tweets rank. We keep track of the excessive number of "uninteresting" words such as non-related trend names in the tweet text, non-related hashtags in the tweet, and "bad words" from our lists of swear and spam words. We can figure out the proportion of spam

to tweet ratio by summing up the characters of the 3 types of uninteresting words, and dividing the uninteresting character sum by the total tweet character size. This translates to a direct spam score that we use to penalize a tweet that is composed of entirely uninteresting words.

3.2 Rankings Tweets related to News Topics

In order to examine how our ranking algorithm could be combined with existing IR techniques to provide improved search in microblog systems, we combined our ranking algorithm with a simple Boolean matching model. In order to facilitate this, we used tweets from the @breakingnews Twitter account as example queries, and parsed both these example queries, and tweets we downloaded, to produce a set of keywords. We defined these keywords as all proper nouns within a tweets text. We then computed a matching score that was simply the number of matching keywords between the tweet and the query tweet. Here, we mathematically define our combined algorithm:

$$R(T_i) = \epsilon \cdot \sum M \cdot \alpha \cdot \log\left(\frac{\sum rt_i}{RT}\right) + \beta \cdot v_i + \gamma \cdot l_i + \delta \cdot \log\left(\frac{f_i}{F}\right) \quad (2)$$

rt_i = a retweet of T_i

f_i = # of followers of tweeter, at time of tweet

RT = max # of retweets, limited to 101 by API

F = # followers of most followed Twitter user

M = # of keyword matches between query and tweet

$\alpha + \beta + \gamma + \delta + \epsilon = 1$

$$v_i = \begin{cases} 1 & \text{if the tweeter is verified} \\ 0 & \text{otherwise} \end{cases}$$

$$l_i = \begin{cases} 1 & \text{if the tweet contains a link} \\ 0 & \text{otherwise} \end{cases}$$

We modified our ranking algorithm as follows. First, we removed some of the spam detection that made sense mostly in the settings of Twitter trends. We also divided each of the weights in the previous algorithm by half, and gave a weight of 0.5 to the simple ranking algorithm. This allowed us to perform searches (using the example queries from @breakingnews) to find a set of matching tweets ranked by our TwitNews ranking algorithm.

4 Results

In this section, we discuss the results of the experiments we ran on real Twitter data. First, we start with the data set composed of Twitter Trends and their corresponding tweets. We look at two specific trends from the set as examples, comparing our ranking algorithm against the timestamp method currently employed by Twitter. We then review our second experiment, in which we combined the TwitNews ranking algorithm with a simple Boolean matching model in order to perform search amongst tweets. This experiment was run against the data set downloaded by crawling tweets related to the most recent tweets from @breakingnews. From one specific example query from @breakingnews, we will compare the results of the timestamp method employed by Twitter, the Boolean model on its own, and our combined Boolean matching with TwitNews ranking algorithm.

4.1 Ranking Tweets for Twitter Trends

Our first experiment consisted of using the TwitNews ranking algorithm to rank the tweets for any given Twitter trend. The experiment was conducted on a data set consisting of approximately 1.5 million tweets collected through the Trend Crawler, distributed over 52 Twitter trends. This data set was crawled over multiple days. For this section, we will present the results for two trends in the data set.

4.1.1 Interesting Tidbit Trend

The first trend we will examine is “DO A BARREL ROLL”, which was chosen as an example of a trend that helped spread a simple tidbit of information that would be of interest to a user, but not generally be classified as a news story by more traditional sources such as Associated Press (AP) or the BBC. This trend was related to a Google.com Easter Egg where the Google.com search results page would do an animated spin if the user searched the query “Do a Barrel Roll”.

First, we look at the results as they would be displayed on Twitter. We do this by ordering the tweets by most recent timestamp. Table 1 on page 6 shows the top 10 tweets when ranked by this method.

Looking at the top 10 tweets as they would be presented on Twitter, it is hard to figure out what the trend is about. Many of the top tweets do not provide any useful information. For instance, the first tweet is really more about Justin Beiber than about DO A BARREL ROLL. The second tweet is just a repeat of the trend with a question mark. A user would have to look down to the 6th tweet to see any relevant information. In fact,

Rank	User	Verified?	Followers	Tweet	Link Domain
1	dmbmeg	No	440	I just realized "do a barrel roll" is trending and I just tweeted about Justin Bieber. Please don't unfollow me. I'll do better next time.	None
2	brunnoleonardo	No	724	Do a Barrel Roll! ?	None
3	jonnalammers	No	131	RT @lisajorinde: RT @NinaSylvie: hahahaa typ in bij google do a barrel roll , je schrikt je dood	None
4	xddlovatosmile	No	331	Ho digitato "do a barrel roll" su Google,google se ne va di testa ogni tanto.AHAHAHAHAHAH	None
5	vadro	No	23	RT @Nesttor77: Entren a Google y pongan do a barrel roll! Jajajaja mola mucho xD	None
6	JeromeParadis	No	2669	Nice Easter egg! Re: Do A Barrel Roll! In Google! Right Now! via @techcrunch http://t.co/mrgOAgMh	techcrunch.com/
7	DoweyBaothman	No	117	RT @MishaelSaad: THINGS YOU SHOULDNT DO: Drugs.	None
8	chellyybean	No	224	umm wtf...i want my google to do a barrel roll. WHY WON'T IT WORK?!	None
9	LizaCRVG	No	118	@JonataanW_OFC - Abra a pagina inicial do google, e procure: do a barrel roll e me diz o que achou =)	None
10	GlitterMonsterx	No	831	Going on a laptop JUST to search up " Do a Barrel Roll "	None

Table 1: Top 10 tweets for query "DO A BARREL ROLL" using the Timestamp method.

only 3 of the tweets presented really seem to be relevant to the topic. None of the tweeters is a Twitter Verified User. Also, most of the users have on the order of 10,000 followers, with one user having a follower count of over 100,000. Thus, a user looking at these tweets will not be confident that the information presented is trustworthy, or has had an impact on a large number of people. Finally, only 1 of the 10 tweets has a unique link, which points to techcrunch.com, a reputable news source. More telling, none of the top 10 tweets has a link to Google.com, the site where the Easter egg was actually available.

Next we look at the same data set of tweets, but as ranked by the TwitNews ranking algorithm. Table 2 on page 7 shows the top 10 tweets identified by this method.

In contrast to the previous results, the top 10 tweets presented by our ranking algorithm are all relevant to the DO A BARREL ROLL trend. All of the tweeters are Twitter Verified Users. Three of the 10 tweeters have over a million followers, while most have at least 200,000. Thus, the user could feel secure in knowing that the information provided is both trustworthy and impactful. Finally, 9 of the top 10 tweets include links to additional sources. Four of the 10 link to reputable tech news sites, including Mashable.com, Kotaku.com, and Techcrunch.com. More tellingly, 5 of the 10 link to Google.com, the site where the Easter egg was available. Interestingly, the 10th tweet, by @Econsultancy, even provides a link with the proper search term already encoded, so that the user would be presented with Google "doing a barrel role" upon clicking the link.

Comparing the two methods, we believe it is clear that our ranking algorithm provides significantly better results than those currently employed by Twitter. The tweets provided by Twitters method are essentially random, and do not consistently provide trustworthy, useful information. In contrast, The tweets provided by our algorithm consistently come from users verified by Twitter that have a large number of followers. Additionally, the tweets provide relevant information to the trend in question, and contain a large number of links to reputable sites where the user can find more information on the trend. Thus, we believe our ranking algorithm preforms well on trends of similar themes to the "DO A BARREL ROLL" example we just examined.

Rank	User	Verified?	Followers	Tweet	Link Domain
1	mashable	Yes	2,568,170	"Do a Barrel Roll" on Google, and You Won't Be Disappointed - http://t.co/pqaDJ4Fi	mashable.com
2	Alyssa_Milano	Yes	1,828,268	Go to http://t.co/5QlfhQwX & type "do a barrel roll".The results will entertain you. / via @BuzzFeed	google.com
3	Kotaku	Yes	100,767	Google "Do a Barrel Roll" Right Now http://t.co/OJMZZMyj	kotaku.com
4	jennydeluxe	Yes	420,292	Everyone do this NOW rt @ryeclifton Google "do a barrel roll" via @lindseyweber	None
5	threadless	Yes	1,672,228	Type "do a barrel roll" into http://t.co/AGwbxxEv. When your screen stops spinning, request a reprint of this - http://t.co/FsISaFwT	google.com
6	toptweets_es	Yes	277,021	RT @jbrownridge: 1) Go to http://t.co/4fy4lVeH 2) Search for "do a barrel roll"	google.com
7	toptweets_es	Yes	277,020	RT @abrmorales: 1. Entra a http://t.co/ANAGfAOy 2. Escribe: do a barrel roll 3. Pulsa enter.	google.es
8	Scrobleizer	Yes	215,917	RT @mashable: "Do a Barrel Roll" on Google, and You Won't Be Disappointed - http://t.co/pqaDJ4Fi	mashable.com
9	arrington	Yes	99,944	RT @Techmeme: Do A Barrel Roll! In Google! Right Now! (@mjburnsy / TechCrunch) http://t.co/b2suSooM http://t.co/WZyncXwx	techcrunch.com
10	Econsultancy	Yes	74,588	A little fun! RT @lakey: The mother of all Google tricks... try entering the search query: "do a barrel roll" http://t.co/6Sr5LQMw	google.com

Table 2: Top 10 tweets for query "DO A BARREL ROLL" using the TwitNews ranking algorithm.

4.1.2 Newsworthy Trend

The second trend we will examine is “James Bond”, which was chosen as an example of a trend that reflected a news story. In particular, this trend was associated with the announcement of a new James Bond movie entitled Sky Fall. Stories about this topic appeared on other, more traditional sources, such as Google News.

First, we look at the results as they would be displayed on Twitter. As before, this is done by ordering the tweets by most recent timestamp. Table 3 on page 9 shows the top 10 tweets as ranked in this fashion.

Looking at these top 10 tweets, we can see that the results are rather hit or miss. There are two tweets that are retweets of more reputable news sources, which do provide some interesting information. There are also two tweets, by @digitalspybrk and @digitalspyent, that also provide some interesting information. However, the remaining 6 tweets display personal opinion. These opinions range from criticisms of the title and series, to comments about the physical attractiveness of the main actor, Daniel Craig. As a result, only 4/10 of the tweets in the top ten are likely to be relevant for someone looking to find more information about James Bond or the new James Bond film. In addition, none of the users that posted these tweets are Twitter Verified Users. Half of the users have a follower count less than 100. The user with the highest follower count, @MalcomIngram, had a follower count of 11,653. As a result, a user looking at these tweets would not be able to very certain about the truthfulness of any information presented. Finally, there are only 3 unique links presented, meaning any additional information the user could find from looking at these tweets might be limited.

Next, we will examine the same set of tweets, but ranked in descending order according to the score produced by our ranking algorithm. Table 4 on page 10 shows the top 10 tweets when ranked by this method.

The top 10 tweets presented by our ranking algorithm provide more relevant and more consistent information than the timestamp based method currently used by Twitter. All 10 tweets presented display relevant information to someone looking to find our more about James Bond or the new James Bond film. In comparison to the timestamp method presented, all the tweets in the top 10 presented by our ranking algorithm are from Twitter Verified users. Thus, a user could be confident that he is seeing trustworthy information. Four of these users have more than 1 million followers, and all but one user has at least 15,000 followers. As a result, the user would be sure to know that he is looking at the tweets and information seen by the largest number of people. Finally, 7 of the 10 tweets presented by our algorithm contain a unique link, meaning a user could find much more additional detail about the topic by following those links.

Again, we think that examining the difference between the two results shows that our ranking algorithm performs much better. Again, the tweets provided via the method used by Twitter are essentially random, and often do not contain trustworthy or relevant information. In contrast, the results presented using our algorithm consistently provide results from trusted and popular sources, and that the information presented is very relevant to the trend in question. Thus, we believe our ranking algorithm preforms well at surfacing the most relevant information in Twitter trends.

4.2 Finding and Ranking Tweets related to News

In addition to testing our ranking algorithm on Twitter trends, we wanted to examine how it could be integrated with a real search example. To do so, we decided to take the latest tweets from @breakingnews, a verified Twitter account dedicated to presenting world news stories, and considered those as example “queries”.

In order to modify our ranking algorithm to perform basic search, we implemented our simple Boolean matching algorithm, which calculated the number of matching keywords between a given tweet and each of our original query tweets. In doing this, we defined keywords as any proper nouns that appeared in the tweet. We then combined the scores determined by this simple Boolean model with the results of our ranking algorithm to produce a score for each tweet in the data set. In this section, we will refer to this combined rank as simply “rank” for the sake of brevity.

To see how well our combined algorithm worked, we will examine one of the example queries given three separate ranking solutions. The example query we will look at is “Penn State said to be planning football coach Joe Paterno’s departure, sources say - @NYTimes http://t.co/1KAsWAZu.” The first of the three ranking algorithms we will examine will be the simple timestamp based ranking currently employed by Twitter (Table 5 on page 11). The second will be ranking first by only the simple Boolean model, with ties being broken by the timestamp method (Table 6 on page 12). Third, we will examine our combined Boolean model and ranking algorithm approach (Table 7 on page 13).

We can see that, as was the case when ranking tweets for trends, the timestamp model produced essentially random results. Again, much of the information is irrelevant to the original query, and comes from

Rank	User	Verified?	Followers	Tweet	Link Domain
1	KayBee0311	No	58	"@BreakingNews: Producers say new James Bond film will be called 'Skyfall,' star Daniel Craig, Javier Bardem - AP" Yay can't wait!!	None
2	MalcomIngram	No	11,653	James Bond is 007 in SKYFALL?!?! Worst title ever Better one off the top of my head James Bond is 007 in TIME EQUALS DEATH #BetterBondTitles	None
3	Diianamedez	No	59	Novo filme do James Bond :) Aiii Daniel deliiiciiiaaa :P	None
4	CajunBrook	No	357	New James Bond movie coming out, but not until November 2012. Can I buy tickets now?	None
5	NorbertKuehn	No	52	RT @HuffingtonPost: Breaking updates: All the new James Bond film details as they happen! http://t.co/6ViOiQ2K	huffingtonpost.com
6	digitalspybrk	No	3,887	'Skyfall' plot sees MI6 "under attack", James Bond loyalty to M tested http://t.co/dITRYmth	digitalspy.com
7	digitalspyent	No	5,013	'Skyfall' plot sees MI6 "under attack", James Bond loyalty to M tested http://t.co/GSCFgT1R	digitalspy.com
8	lincingcom	No	56	New James Bond Film is Officially Titled 'Skyfall'; Mendes and Producers Update Casting and Plot Details http://t.co/TugXWONC	showbiz.lincing.com
9	Phildaddy007	No	50	Well now Daniel Craig is now a trending topic over James Bond... That works, I'll go with that... he is the #BestBond	None
10	Khusomir	No	130	Enough with James Bond, such a boring series...	None

Table 3: Top 10 tweets for query "James Bond" using the Timestamp method.

untrusted and unpopular sources.

Looking at the top 10 tweets as ranked solely by the Boolean model, we find that it does a good job of finding information that is related to the original query. However, it fails to distinguish very well between information that matches the original query. As we can see, many of the tweets presented by the Boolean model, only top 10 are actually duplicate retweets of the same originally tweet. Consequently, while all ten tweets contain a link, there are actually only 3 unique links in the whole set. Also, the Boolean only model fails to take in to account any of the structured information present in the tweet, such as verified user status or number of retweets.

Examining the combined method, we find that the TwitNews ranking algorithm handles the flaws apparent in the Boolean model only very well. It takes into account the structured social information inherent in every tweet in order to rank amongst tweets with the same amount of matching terms. Looking at the top 10 tweets provided in table 4.2.3, we can see that 7 out of the 10 tweets come from verified users, and that each tweet comes from users with a high number of followers. Also, out of the 10 tweets, 7 unique links are presented. We believe these results show that the TwitNews ranking algorithm can be extended to enhance other, more sophisticated information retrieval techniques in order to improve search techniques in microblog networks.

5 Related Work

In TwitRank: Extracting and Ranking Relevant Tweets for News Events[4], Dhar explores the idea of finding real-time information about breaking news events from the Twitter platform, an idea similar to ours. However, his motivation is different, as Dhar uses a machine learning algorithm to rank tweets based on previously fed breaking news events from authoritative sources like nytimes.com, cnn.com, and wsj.com to find other, closely related news topics. Our algorithm is largely

Rank	User	Verified?	Followers	Tweet	Link Domain
1	BBCWorld	Yes	1,182,858	Next James Bond film will be titled Skyfall, producers announce. More soon http://t.co/jTItwN70	bbc.co.uk
2	AP	Yes	593,789	James Bond returns next year in "Skyfall," starring Daniel Craig and Javier Bardem. It's 007's 23rd movie: http://t.co/5nKhsM1C -EF	hosted.ap.org
3	BreakingNews	Yes	3,225,044	Google Producers say new James Bond film will be called 'Skyfall,' star Daniel Craig, Javier Bardem - AP	None
4	HuffingtonPost	Yes	1,345,737	Breaking updates: All the new James Bond film details as they happen! http://t.co/6ViOiQ2K	huffingtonpost.com
5	BBCNews	Yes	379,179	Next James Bond film will be titled Skyfall, producers announce. More soon http://t.co/S4kmCo2G	bbc.co.uk
6	holacom	Yes	157,398	Comienza el rodaje de 'Skyfall' el nuevo filme de James Bond, con Daniel Craig y Javier Bardem como protagonistas: http://t.co/g1LW82bl	noticias.hola.com
7	ParisMatch	Yes	53,248	James Bond tombera du ciel: Lors d une confÃľrence de presse, le titre des 23e aventures du cÃľlÃĺbre agent secret ... http://t.co/DXGY0ZTB	parismatch.com
8	NoticiasMVS	Yes	34,532	La nueva pelÃ■cula de James Bond se llamarÃą 'Skyfall' http://t.co/XXMDD0Jm	noticiasmvs.com
9	MetroUK	Yes	18,863	James Bond film title announced as Skyfall at launch with Daniel Craig http://t.co/EL9rePnJ	metro.co.uk
10	YahooOmgUk	Yes	6,555	The James Bond/Daniel Craig news WE'D been waiting for http://t.co/koAiVBFw	uk.omg.yahoo.com

Table 4: Top 10 tweets for query "James Bond" using the TwitNews ranking algorithm.

Rank	User	Verified?	Followers	Tweet	Link Domain
1	mdulian	No	786	ESPN is reporting that it looks like Joe Paterno will be gone! RT @Elliott_Sadler: What is Penn State goin... (cont) http://t.co/HRXHgBOI	None (Points to Tweet Longer service, entry no longer exists)
2	1DSheeranBiebs	No	1251	@AwwMusic when I have credit and I'm with Joe next I'll ring you and you can talk to him hahahhaaa	None
3	LordFestusIII	No	52	RT @speakz: RIP Joe Fraizer.Ali called you an uncle tom & a paper champion in build up to your 1st fight...You responded by beating his ...	None
4	NewsDangler	No	52	I'm at Nyack Beach State Park (698 N Broadway, Upper Nyack) w/ 2 others [pic]: http://t.co/HXEJ5Dho	https://foursquare.com/
5	jOleary20	No	170	Hope everything works out for Penn State @ItsMeSnitchez10	None
6	gonzalezmam	No	2128	Penn State's Paterno cancels briefing amid scandal http://t.co/ajjYsbZ9	http://www.reuters.com/
7	freehan11	No	542	This scandal at Penn State is truly disgusting. No one ever learns that coverups always come uncovered. God Bless the victims.	None
8	avalam00	No	21	RT @mark_may: The next AD at Penn State should be Matt Millen	None
9	Bojnadles	No	34	Penn State should get the death penalty.. Sick!	None
10	DatDudeNickT93	No	48	Guys, Real Joe Pa has been dead now for some time. They simply are propping his corpse up in the booth on GameDays #WeekendAtBernies Style	None

Table 5: Top 10 tweets for query "Penn State said to be planning football coach Joe Paterno's departure, sources say - @NYTimes http://t.co/1KAsWAZu" using the Timestamp method.

Rank	User	Verified?	Followers	Tweet	Link Domain
1	kimtee	No	103	"@BreakingNews: Penn State said to be planning football coach Joe Paterno's departure, sources say - @NYTimes http://t.co/sOCyq3pgâĂİ WHAT.	http://www.nytimes.com
2	JustRizz	No	165	RT @SportsCenter: Breaking News: The New York Times: Penn State is planning head coach Joe Paterno's exit http://t.co/K5AfLhqQ	http://www.nytimes.com
3	MsMoo_Pow	No	141	RT @SportsCenter: Breaking News: The New York Times: Penn State is planning head coach Joe Paterno's exit http://t.co/K5AfLhqQ	http://www.nytimes.com
4	BrwnEyeGemini	No	151	@2_shay_my_nigga The New York Times: Penn State is planning head coach Joe Paterno's exit Âż http://t.co/9DigGXjmâĂİ « that ish right there	http://www.nytimes.com/
5	WrnrG	No	26	Report: Penn State planning Joe Paternos exit amid sexual-abuse scandal - College Football News | FOX Sports on http://msn.foxsports.com/collegefootball/story/Penn-State-planning-Joe-Paterno-exit-amid-sexual-abuse - MSN: http://t.co/lQSzSoBm	http://msn.foxsports.com
6	nthartness6	No	605	RT @SportsCenter: Breaking News: The New York Times: Penn State is planning head coach Joe Paterno's exit http://t.co/K5AfLhqQ	http://www.nytimes.com/
7	jcr4522	No	60	âĂIJSportsCenter: Breaking News: The New York Times: Penn State is planning head coach Joe Paterno's exit http://t.co/rn7Yfz8MâĂİ	http://www.nytimes.com
8	DHarguth	No	155	RT @SportsCenter: Breaking News: The New York Times: Penn State is planning head coach Joe Paterno's exit url-http://t.co/K5AfLhqQ	http://www.nytimes.com
9	D2Lo	No	41	RT @SportsCenter: Breaking News: The New York Times: Penn State is planning head coach Joe Paterno's exit http://t.co/K5AfLhqQ	http://www.nytimes.com
10	final118	No	64	RT @USATODAY: Penn State cancels Joe Paterno's scheduled news conference http://t.co/QuOhFXr8	http://content.usatoday.com

Table 6: Top 10 tweets for query "Penn State said to be planning football coach Joe Paterno's departure, sources say - @NYTimes http://t.co/1KAsWAZu" using only the Boolean model.

Rank	User	Verified?	Followers	Tweet	Link Domain
1	USATODAY	Yes	189,806	Report: Penn State planning Joe Paterno's exit http://t.co/zVEhIm3s	http://content.usatoday.com
2	USATODAY	Yes	189,804	Penn State cancels Joe Paterno's scheduled news conference http://t.co/QuOhFXr8	http://content.usatoday.com
3	michaelombardi	Yes	133,191	According to the NY Times Penn State is planning for Joe Paterno's exit. http://t.co/gT3voIwC These is a horrendusly sad story...	http://www.nytimes.com
4	YahooSports	Yes	40,747	Paper: Penn State planning Joe Paterno's exit http://t.co/KmBM17y2	http://rivals.yahoo.com/
5	ESPNNewYork	Yes	33,045	Penn State Nittany Lions call off Joe Paterno's news conference http://t.co/Z0sVeYRT	http://espn.go.com
6	vleach44	Yes	16,125	RT @SportsCenter: Breaking News: The New York Times: Penn State is planning head coach Joe Paterno's exit http://t.co/DGsPU6zT	http://www.nytimes.com/
7	CBSi	Yes	10,712	Report: Penn State planning Joe Paterno's exit - http://t.co/BRkBG5WD	http://eye-on-collegefootball.blogs.cbssports.com
8	OpieRadio	No	105,719	YAWN! knew this 18hrs ago! BELIEVE RT @BreakingNews Penn State planning football coach Joe Paterno's departure @NYTimes http://t.co/9VaXtLXk	http://www.nytimes.com
9	evertus	No	50,069	Penn State president cancels Joe Paterno's press conference (morningcall): Share With Friends: | | Top News - ... http://t.co/7KbKSxFy	http://www.mcall.com
10	hackhype	No	29,468	Report: Penn State planning coach Paterno's exit: Joe Paterno's tenure as head coach of Nittany Lions is likely ... http://t.co/lyXotQWN	http://nfl.com

Table 7: Top 10 tweets for query "Penn State said to be planning football coach Joe Paterno's departure, sources say - @NYTimes http://t.co/1KAsWAZu" using the combined TwitNews ranking and Boolean model algorithm.

focused on filtering away the firehose of largely redundant and sometimes irrelevant tweets returned by Twitter from a user query. Additionally, our implementation is very different. Dhar uses machine learning techniques in his attempts. On the other hand, we use a pre-computed ranking algorithm designed for finding authoritative tweets combined with other existing IR techniques for search. Our advantage lies in fast results with no time or effort spend on training, as well as avoiding the need for trainers to initially judge Twitter results for machine learning to function. Also, our approach is more applicable to applications besides finding newsworthy tweets, as we focus on the concepts of trustworthiness, relevance, and authority conveyed by the social data provided along with each tweet.

In Ranking Approaches for Microblog Search, authors Nagmoti, Teredesai, and Cook examine ways to rank microblogs in real time. However, the mostly focus on the number of tweets a user has posted, and number of followers of a user as the main ranking mechanisms[8]. However, we feel that these factors are relatively unimportant in determining the relevance of a tweet.

In their paper Topical Semantics of Twitter Links[10], authors Welch, Schonfeld, He, and Cho argue that topical relevance is preserved better over retweet links. They provide the intuition that followers do not necessary agree with every topic that the "followee" tweets. This idea is one of the key factors that we exploit in our algorithms. Another factor that we took into careful consideration is the URL presence inside a tweet. This concept is explored in papers Time is of the Essence[2] and An Empirical Study on Learning to Rank of Tweets[5].

Before incorporating the Boolean model, as a heuristic to reduce the noise such as advertisement tweets from spam bots, we preloaded and checked tweets against various swear-filter and common bad words lists available on the web. However, much of the work done on spam detection on Twitter involves detection on the user level, where we were more interested with being able to detect spam based solely upon information provided with a tweet. Incidentally, our TrustRank algorithm does a fairly decent job of detecting spam with simple word filters.

6 Conclusions and Future Work

In this paper, we examined how we can rank tweets, or microblogs, using the structured information provided with each tweet. We then examined how this ranking algorithm could be combined with existing information retrieval techniques to improve search in microblogging networks. We believe that the TwitNews ranking algorithm does a good job of surfacing information that is trustworthy, relevant, and diverse. We now present a few areas where we believe further work could be done to improve or expand upon the ideas discussed in this paper:

- There are various limitation in place with the current Twitter API that limit the effectiveness of the TwitNews algorithm. In particular, being limited to 101 as the retweet count severely hampers the effectiveness of the algorithm in ranking tweets that obtain this maximum retweet count. It would be desirable to configure a system that could manually retrieve the retweet counts in order to improve the algorithm.
- We believe there is much more work that could be done in spam detecting on a tweet level basis. While we implemented some basic spam filtering mechanisms in TwitNews, we believe this is an area in which more research could be done.
- In a similar vein, we believe that another improvement that could be made is better handling of duplicate tweets. In TwitNews, we mostly shield away from attempting to remove duplicates, due to the difficulty in detecting duplicate tweets. For example, many tweets are identical, but differ because they use two different link shortening services.
- We would also be interested in seeing the TwitNews algorithm adapted to other social networks that provide microblogging services, such as Facebook (via the News Stream). We would also like to see the algorithm combined with other, more sophisticated, information retrieval techniques to gain further insight into how the TwitNews ranking algorithm can impact search in microblogging networks.

References

[1] Twitter api. http://dev.twitter.com.

[2] Dong Anlei, Ruiqiang Zhang, Pranam Kolari, Jing Bai, Fernando Diaz, Yi Chang, Zhaohui Zheng, and Hongyuan Zha. Time of the essence: Improving recency ranking using twitter data. 2010.

[3] Sergey Brin and Lawrence Page. The anatomy of a large-scale hypertextual web search engine. 1998.

[4] Nitin Dhar. Twitrank: Extracting and ranking relevant tweets for news events, May 2011.

[5] Yajuan Duan, Long Jiang, Tao Qin, Ming Zhou, and Hueng-Yeung Shum. An empirical study on learning to rank of tweets. 2010.

[6] Matt Facer. Swear filter. http://www.mattfacer.com/swear-filter/.

[7] Zoltan Gyongyi, Hector Garcia-Molina, and Jan Pedersen. Combating web spam with trustrank. Technical report, March 2004.

[8] Rinkesh Nagmoti, Ankur Teredesai, and Martine de Cook. Ranking approaches for microblog search. 2010.

[9] Alejandro Urbano. Bad words list. http://urbanoalvarez.es/blog/2008/04/04/bad-words-list/.

[10] Michael Welch, Uri Schonfeld, Dan He, and Junghoo Cho. Topical semantics of twitter links. February 2011.