Twitter Trend Detection

T-61.5910 Research Project in Computer and Information Science

Zhe Xie
Aalto University
Espoo, Finland
zhe.xie@aalto.com

ABSTRACT
Many researches have been built up by analysing the textual contents in tweets. In this project, we are particularly interested in their topics and try to find out the most popular ones, namely trends. To meet this intention, a series of algorithms are developed including stop word generation, trends and spam tweet detection and representative tweet identification. In the end, evaluation of the results will be performed after applying these algorithms on the real tweet data downloaded from Twitter API.

1. INTRODUCTION
Twitter is an online social networking and microblogging service with which users can send and read text message limited to 140 characters, known as "tweets". Although the length of the message is limited, tweets provide some predefined symbols to enhance their utility. For example, symbol "@" can be used to communicate with other users by typing a user identifier after it and symbol "#" can be added in front of a phrase to emphasis a topic, known as a "hashtag".

Twitter has become a very popular social network since its born in 2006. According to an online statistic [1], by the time October 24, 2013, there are about 231.7 million active users on Twitter worldwide and 100 million of them log into the service on daily basis. Every second there are more than 5,000 tweets posted and over 400 million tweets are posted per day.

Given such an amount of information available on Twitter, we are interested in analysing their contents and finding out what are the hot topics in them, namely trends. In the end, trends are presented with the form of key words and its representative tweets.

A two-stage detecting procedure has been developed, It begins with trend detection algorithms which identify trends by tracking the frequency of words or phrases in tweets. Specially in this project, two sequential words in tweets or bigrams are considered to be main form of trends. This is because there is no formal rules to define trend explicitly and bigrams are usually easy to understand and explain. (In fact the algorithms applied on bigrams can also be applied on phrases with different lengths. If trends with different gram length are considered, we should then develop an algorithm for trend combination as well.)

As for next stage we will try to find their representative tweets based on the found trends in the form of bigrams. This is because bigrams may be ambiguous and we still want to know the events behind them. There are several ways to find the representative tweets. The most common way is to directly search for the trend using Twitter searching API, which is used by some twitter trend tracking website[2]. In this project we develop a scoring algorithm to find the most representative tweets for each trend. What's more, this algorithm also can remove spam trends from the detected trends of stage 1.

Twitter also generates and exhibits a list of trends at their website as user recommendations. This trend list can be tuned by users' preference or location. However, there is a drawback about its trend finding algorithm: it doesn't group similar trends[1]. In this project, we also develop a algorithm to fix the problem to some extent.

2. SETTING
In this section, we will present a general description of the dataset and provide some useful data statistics.

2.1 General Description of the Dataset
The data used in this project is downloaded from twitter public streaming API and consists of two parts.

The first part is collected from 2013-9-28 16:35:53 to 2013-10-02,12:45:40 and used to explore data and learn the algorithms. The second part is collected from 2013-11-27 09:24:46 to 2013-12-11 13:10:23 [3] and used in the experiment section to be the test set for trend detection by the

[1] http://socialmediatoday.com/irfan-ahmad/1854311/twitter-statistics-IPO-infographic

[2] http://trends24.appb.in/

[3] There are three interruptions during the collection precessing: from 2013-12-1 0:26:36 to 2013-12-2 10:33:38, from 2013-12-3 1:52:12 to 2013-12-3 10:41:5 and from 2013-12-8 0:40:26 to 2013-12-9 11:0:43

algorithms developed in Algorithms section.

During both of the data collections, we filter the tweets by setting the language option in order to get tweets in English. In total 4,782,894 tweets are collected in the first part and 16,425,973 tweets are collected in the second part.

Besides tweet contents, other information can also be found in the data set including the embedded hashtags, embedded link, user information, geographic information and tweet published time.

2.2 Useful Data Statistics

In this part we will use the data set part 1 and explore the single words in the dataset first.

The number of words is 3,678,034 after data preprocessing including removing punctuations and Unicode characters

If we don't remove stop words, the top 10 most frequent words will be as shown in the left of the following figure.

	unigram	count	unigram	count
1	rt	1538667	stats	50910
2	i	1518992	tomorrow	49452
3	the	1319728	tonight	44663
4	to	1217182	game	44059
5	you	1097608	twitter	39181
6	a	963814	wait	38632
7	and	742280	real	38602
8	my	642099	bitch	38110
9	is	606229	sleep	38061
10	in	561775	friends	37932

As we can imagine, most of the frequent words will be stop words which refer to common but meaningless words. So if we remove all the stop words according to the stop word list generated from our stop word algorithm. (We also set the minimal acceptable word length to 3, which means we also remove all word with word length less than or equal 2. This is because those short word are mostly numbers or special characters), then the top 10 most frequent words will be as shown in the right part.

When we compare these two lists, we can get several interesting findings:

- Meaningful word has appeared in the latter list such as game, sleep and friends.
- The count numbers have dropped dramatically ,given that the highest frequency drop from over 1,400,000 to less than 60,000, which means the meaningless word took a great part of the dataset and the computation load would be more manageable in the reduced dataset.

Secondly we will take a look at some statistics about bigrams. We have calculated the frequency for all the bigrams on the tweet data set after data preprocessing including removing punctuations, Unicode characters and stop words.

The following list contains top 10 bigrams for the whole data set.

	bigram	count
1	stats follower	13881
2	posted photo	12608
3	# androidgames # gameinsight	11864
4	# android # androidgames	11270
5	gold coins	10944
6	photo facebook	8388
7	automatically checked	8081
8	# ipadgames # gameinsight	7996
9	government shutdown	7318
10	# ipad # ipadgames	7162

We also plot the histogram as shown in Figure 1 in which the x-axis presents the bigram frequency and y-axis presents the number of bigrams with certain frequencies.

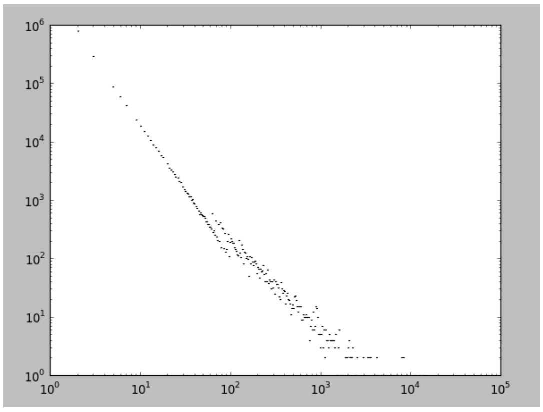

Figure 1: Histograms for bigrams

From the figure we can notice that the number of bigrams with frequency less than 10 is larger than 10,000 and some even over 10^5. So if we remove those bigrams with small frequencies , the total number of bigrams will be much smaller; we can also notice that there are few bigrams with frequency over 1000.Roughly speaking, we will be mostly interested in bigrams with frequency between 10 to 1000.

3. ALGORITHMS

In this section, we will describe all the algorithms in the trend detection task including the algorithm for stop words generation, trend identification, spam tweet detection and representative tweet identification.

3.1 Algorithm for stop word generation

We define stop words as function words or words with extra-high frequency in tweets. Function words may include pronouns (e.g. he, she, it), articles (e.g. a,an,the) and particles (e.g. if, then,well),etc. We remove function words because they don't contain any lexical meanings and remove extra-high frequent words because we basically can't get any valuable information from them since their high frequencies.

We will first detect all the extra-frequent words and then combine them with a general purpose stop list which consists of function words(It turns out that these two kinds of stop word lists have a lot of elements in common).

The whole stop word finding algorithm is as following:

1. Group tweets within a time window of 5 minutes and i total 100 groups are obtained which cover 500 minute
2. Mark the appearance of each word in the groups
3. Claim one word as a stop word if it appear in equal or more than 90 percent of all the groups
4. After all the stop words are detected, merge the stop word list with a general-purpose stop word list[4]

Algorithm 1: stop word generation

3.2 Algorithm for detecting trends

In this section, two algorithms are designed to detect trends. The first algorithm considers raw frequency and identify bigrams with high frequencies as trends and it works as in Algorithm 2.

1. Set up the length of the time window δ
2. Group all tweets within one time window and count the frequency of each bigram.
3. Sort the bigrams according to their raw frequency from largest to smallest.
4. Choose the top N bigrams as the trends within the time window.

Algorithm 2:trend detection by bigram frequency

In this algorithm, some bigrams are usually taking the top ranks because of their common high frequency. Sometimes we also would like to capture a trend with a abrupt increase in its frequency. In order to capture such an increase, we will define a ratio based on the bigram's current and historical frequencies. Details are as follows.

We define $F_x(t)$ as the raw frequency of bigram x in time window t.

So the ratio for bigram x at time window t can be calculated as:

$$ratio_x(t) = \frac{F_x(t)}{w_1 F_x(t-1) + w_2 F_x(t-2) + ... + w_n F_x(t-n)}$$

in which the w_n denote the weight of each historical frequency in the past time windows and all the weights are summed up to one.

The parameter n controls the tracking-back length.When n equals to 1, it means we only take the last time window into consideration.

Particularly, there are some assumptions when we can apply this ratio formula.

1. As the time goes by, we have accumulated a list of historical frequency for each of the bigrams in the t time window. The frequency number of the list are sorted in chronological order and the length of the list is at least n
2. We only deem the bigrams with a increasing frequency are of interest.

We also have several parameters that can be tuned,such as the time window length δ, the tracking-back length n and the weight for each historical frequency. We will try different parameter settings in the experiment section. Here intuitively we will assign more weight to the frequency in the further time window in order to capture the increase trend: the ratio would be larger if in the long past the frequency is low but right now very high.

Finally Algorithm 3,using ratio scores of bigrams to detect trends is as follows:

1. Set up the length of the time window δ
2. Group all tweets within one time window and count the raw frequency of each bigram
3. Calculate the ratio for each bigrams with the current frequency and the historical frequencies according to the ratio formulae.
4. Sort the bigrams according to their ratio score from largest to smallest.
5. Choose the top N bigrams as the trends within the time window.

Algorithm 3:trend detection by bigram ratio

3.3 Algorithm for detecting spam tweets

After we have identified the top N bigrams as the trends in Algorithm 2, we notice that top trends usually come from tweets which are automatically generated by some online applications. Some examples are shown in figure 2.

Figure 2: Two groups of tweets: the first 3 tweets contain the trend "stats follower" and the rest contain the trend "posted photo". These tweets are generated by online applications

Clearly these trends should be regarded as spam trends and tweets which contain these trends should be removed from the tweet group. The algorithm for removing those spam tweets is developed as in Algorithm 4.

1. For each tweets calculate the edit distance with all the other tweets
2. Set a distance threshold δ, if the distance between two tweets is less than δ, we describe the two tweets as close neighbours.

[4] http://jmlr.org/papers/volume5/lewis04a/a11-smart-stop-list/english.stop

3. For each tweet, if the number of its close neighbour is larger than ρ, then it is marked as a spam

Algorithm 4: Algorithm for removing spam tweets

3.4 Algorithm for representative tweet identification

In this algorithm, we not only try to identify representative tweets based on trends detected in Algorithm 2 or 3 but also try to remove spam trends and group similar trends as mentioned in the introduction. The algorithm is shown in Algorithm 5.

1. *Find the trends within some time window(N is set much larger than that in Task 1, actually N is a function of minimal accepted frequency or ratio α)*

2. Find and list all the tweets which contain trends

3. Remove all replicating and spam tweets

4. Construct an inverse table where for each tweet all the trends it contains and their frequencies or ratios can be found

5. Sort tweets by the average frequency or ratio of its top three trends.(If the number of trends is less than three, we calculate the average frequency/ratio of all its trends)

6. Merge different trends together if they show in one tweet or refer to similar events

7. Group tweets by their trends so only one tweets will be preserved for one group of trends

8. Choose the top M tweets as the representative tweets for that time window

Algorithm 5: Algorithm for representative tweets identification

4. EXPERIMENTAL EVALUATION

In this section, data set part 2 is used to detect trends. Several time window lengths are considered. As for detection by bigram ratio, different tracking-back length n will also be applied. In the end representative tweets for corresponding trends will be presented.

In order to make the result easier to explain (especially for the time window length of 24 hours), we set the reference time or the current time to Dec 11 2013 00:00:00. This setting is convenient e.g. the last time window will be a whole day with time window length of 24 hours.

4.1 Trend detection by bigram frequency

In the detection of trends by frequency, several time window lengths are considered including 1 hour, 5 hours and 24 hours. The results are shown in Table 1 to Table 3.

Table 1: trends for time window length of 1 hours

	$TrendInThePast1Hour$	$TrendInThePast2Hour$
1	fashion show	fashion show
2	@real_liam_payne liam	#musicfans #peopleschoice
3	victoria secret	twitter update
4	#musicfans #peopleschoice	posted photo
5	secret fashion	victoria secret
6	twitter update	paul walker
7	show tonight	show tonight
8	paul walker	#androidgames #gameinsight
9	stats follower	stats follower
10	posted photo	gold coins

	$TrendInThePast3Hour$	$TrendInThePast4Hour$
1	fashion show	@thevampscon @thevampstristan
2	#musicfans #peopleschoice	@thevampsbrad @thevampscon
3	posted photo	@thevampstristan @thevampsbrad
4	show tonight	@thevampsbrad @thevampstristan
5	victoria secret	@thevampscon @thevampsbrad
6	secret fashion	@thevampsjames @thevampstristan
7	photo facebook	@thevampsjames @thevampsbrad
8	phil jones	@thevampsjames @thevampscon
9	stats follower	@thevampstristan @thevampscon
10	paul walker	@thevampsbrad #thevampswildheart

	$TrendInThePast5Hour$
1	fashion show
2	#musicfans #peopleschoice
3	posted photo
4	paul walker
5	photo facebook
6	stats follower
7	#androidgames #gameinsight
8	nelson mandela
9	#android #androidgames
10	gold coins

Table 2: trends for time window length of 5 hours

	$TrendInPast1TimeWindow$	$TrendInPast2TimeWindow$
1	fashion show	#musicfans #peopleschoice
2	#musicfans #peopleschoice	posted photo
3	posted photo	photo facebook
4	@thevampscon @thevampstristan	paul walker
5	victoria secret	stats follower
6	show tonight	fashion show
7	@thevampsbrad @thevampscon	#androidgames #gameinsight
8	paul walker	#android #androidgames
9	@thevampsbrad @thevampstristan	gold coins
10	@thevampstristan @thevampsbrad	nelson mandela

	$TrendInPast3TimeWindow$	$TrendInPast4TimeWindow$
1	posted photo	stats follower
2	nelson mandela	posted photo
3	paul walker	drop gpa
4	stats follower	paul walker
5	photo facebook	fast #teamfollowback
6	@justinbieber #pray4philippines	#musicfans #peopleschoice
7	automatically checked	photo facebook
8	justin bieber	#androidgames #gameinsight
9	#musicfans #peopleschoice	#retweet retweets
10	#androidgames #gameinsight	#android #androidgames

	$TrendInPast5TimeWindow$
1	#musicfans #peopleschoice
2	chillin arnie
3	@zaynmalik chillin
4	fashion show
5	posted photo
6	stats follower
7	paul walker
8	arnie pk1nov9iwg
9	#androidgames #gameinsight
10	#android #androidgames

Table 3: trends for time window length of 24 hours

	$trendsfor2013.12.9$	$trendsfor2013.12.8$
1	#musicfans #peopleschoice	#musicfans#peopleschoice
2	posted photo	posted photo
3	paul walker	paul walker
4	stats follower	photo facebook
5	fashion show	#androidgames #gameinsight
6	photo facebook	stats follower
7	#androidgames #gameinsight	#android #androidgames
8	#android #androidgames	lose weight
9	gold coins	lovatics #musicfans
10	automatically checked	gold coins

	$trendsfor2013.12.7$	$trendsfor2013.12.6$
1	#mtvstars direction	#mtvstars direction
2	justin bieber	justin bieberl
3	#mtvstars justin	#mtvstars justin
4	paul walker	#musicfans #peopleschoice
5	posted photo	posted photo
6	lovatics #musicfans	stats follower
7	photo facebook	photo facebook
8	neon lights	#androidgames #gameinsight
9	demi lovato	nelson mandela
10	stats follower	#android #androidgames

	$trendsfor2013.12.5$
1	nelson mandela
2	#mtvstars direction
3	#musicfans #peopleschoice
4	justin bieber
5	#mtvstars justin
6	#femaleartist #peopleschoice
7	directioners #musicfans
8	demi lovato
9	#breakoutartist #peopleschoice
10	#musicvideo #peopleschoice

From the tables above, we can notice that some bigrams extra-frequently appear in the trend list such as #musicvideo #peopleschoice, stats follower and posted photo. These frequently shown-up trending bigrams will reduce the information in tread lists within the time window. Ideally we wish to see distinct trend items between different time windows.

There are no formal rules to evaluate the effectiveness of each item in the trend lists. The evaluation process may be subjective w.r.t users' preference. However, we can develop a measure to determine the overall quality of the whole lists and we call it diversity. It is defined as follows:

$$Diversity(K) = \frac{\#distinct\ trends\ in\ K\ time\ windows}{K * N}$$

where N is the length of the trend list and K is the number of considered time windows.

The value of $Diversity(K)$ is ranged from 0 to 1. Larger $Diversity(K)$ means that more distinct trends are captured, which accords to our expectation.

With K and N equal to 5 and 10 respectively, the diversities for time window length of 1 hour, 5 hours and 24 hours are shown below:

Time Window Length	$Diversity(K)$
1 hour	0.48
5 hours	0.50
24 hours	0.46
Average	0.48

From the table we can see that $Diversity(K)$ doesn't change much with the increasing time window lengths. As we expected, the number of trends detected by frequency is limited.

4.2 Trend detection by bigram ratio

In this section, we will present the trends found by the bigram ratio. Here the time window is fixed to 1 hour but several tracking-back lengths are applied including 1,5 and 10. The results are shown in the Table 4 to 6.

Table 4: trend detection with tracking-back length of 1

	$TrendInThePast1Hour$	$TrendInThePast2Hour$
1	@real_liam_payne liam	roy keane
2	@ashton5sos christmas	@edsheeran family
3	@real_liam_payne christmas	losing weight
4	@itsnaashgrier retweet	send pictures
5	@real_liam_payne hiii	thought losing
6	christmas spirit	lose weight
7	@5sos slappe	keane viera
8	list thankyou	product kick
9	thankyou pati	send nudes
10	retweet list	twitter update

	$TrendInThePast3Hour$	$TrendInThePast4Hour$
1	phil jones	@thevampscon @thevampstristan
2	james milner	@thevampsbrad @thevampscon
3	bill bloombield	@thevampstristan @thevampsbrad
4	kimbleyd bill	@thevampsbrad @thevampstristan
5	bloombield dinner	@thevampscon @thevampsbrad
6	@niallofficial kimbleyd	@thevampsjames @thevampstristan
7	dinner gojtmlkgvw	@thevampsjames @thevampsbrad
8	shakhtar donetsk	@thevampsjames @thevampscon
9	tom cleverley	@thevampstristan @thevampscon
10	cleverley shoots	@thevampsbrad #thevampswildheart

	$TrendInThePast5Hour$
1	mack brown
2	@raf_ceng #rafceng
3	brown step
4	scores tonight
5	@britneyspears #perfumevideopremiere
6	@mileycyrus cough
7	cough sucks
8	britney spears
9	google play
10	high level

Table 5: trend detection with tracking-back length of 5

	$TrendInThePast1Hour$	$TrendInThePast2Hour$
1	@trevormoran #buythedarksideonitunes	losing weight
2	open honest	thought losing
3	spree retweet	product kick
4	tonight chicago	send nudes
5	homework classwork	weight pounds
6	classwork homework	recommend lose
7	@wiz_khaliifea sluts	made twitter
8	sluts ei2kdsowkk	amazing diet
9	#buythedarksideonitunes @trevormoran	results amazing
10	clingy disturbingly	impossible trid

	$TrendInThePast3Hour$	$TrendInThePast4Hour$
1	@arianagrande promised	tweet limit
2	spree til	@nuteila watching
3	til limit	cristiano ronaldo
4	promised spree	sir alex
5	twitter update	air jordan
6	@itsnaashgrier retweet	player score
7	@arianagrande santa	michael jordan
8	@michael5sos michael	@zaynmalik_x1 perrie
9	wasn born	page u1cik6i4d9
10	@arianagrande ariana	pregnant tesxpuvdiq

	$TrendInThePast5Hour$
1	@taylorswiftc3 fucked
2	fucked 3c2zk0ecwu
3	@ashton5sos ash
4	tits ntnxmhobd8
5	@istonergirl photos
6	photos tits
7	head coach
8	facebook likes
9	bynes 3v9dshsxim
10	@kimkardashlan_x amanda

Table 6: trend detection with tracking-back length of 10

	$TrendInThePast1Hour$	$TrendInThePast2Hour$
1	@trevormoran #buythedarksideonitunes	send nudes
2	open honest	made twitter
3	spree retweet	twitter update
4	tonight chicago	banga banga
5	@wiz_khaliifea sluts	crush texts
6	sluts ei2kdsowkk	doesn change
7	#buythedarksideonitunes @trevormoran	naked shoot
8	personality varies	@rihrarana fully
9	varies unbearably	twas finals
10	unbearably clingy	makes close

	$TrendInThePast3Hour$	$TrendInThePast4Hour$
1	twitter update	tweet limit
2	@michael5sos michael	@nuteila watching
3	@arianagrande ariana	air jordan
4	wasn born	michael jordan
5	000 times	kim pregnant
6	ready czbjtvhb46	pregnant tesxpuvdiq
7	@tatted_chicky ready	@kanyre_west kim
8	retweet #wcw	cyrus butt
9	@arianagrande leave	butt naked
10	put christmas	naked ul8p4nzvkp

	$TrendInThePast5Hour$
1	@taylorswiftc3 fucked
2	fucked 3c2zk0ecwu
3	@ashton5sos ash
4	tits ntnxmhobd8
5	photos tits
6	@istonergirl photos
7	head coach
8	josh mccown
9	facebook likes
10	nick foles

With the bigram ratio, we do discover some new interesting trend, e.g. "air jordan" in the fourth hour. We assume it is related to the coming new Air Jordan basketball shoes.

In order to measure the overall quality of the whole lists, $Diversity(K)$ is calculated again with the setting of $(K, 5)$ and $(N, 10)$. The result is shown below.

Tracking-back length n	$Diversity(K)$
1 time window	1
5 time windows	1
10 time windows	1
Average	1

We can see that all the trend lists in the 5 time windows are different from each other and we have got a maximum diversity.

4.3 Representative tweet presentation

In the section we will present the representative tweets for their corresponding trends. The time window length is set to be 1 hour and the trends are found by Algorithm 2. In table 7, representative tweets are presented for the trends in the last 1 hour.

Table 7: representative tweets for trends in last 1 hour

	trend	representative tweet
1	victoria secret @real_liam_payne liam fashion show:	@Real_Liam_Payne LIAM ARE YOU GOING TO WATCH THE VICTORIA SECRET FASHION SHOW
2	#musicfans #peopleschoice	Little Monsters #musicfans #PeoplesChoice
3	twitter update	RT @z4ny_: New twitter update.
4	paul walker	RT @RockOfficiaI: R.I.P My Brother Paul Walker http://t.co/t9DYwjtlFa
5	posted photo	RT @selenagomez: Just posted a photo http://t.co/Pb3hzvsVzC
6	justin bieber #musicfans #peopleschoice	"RT @ButeraBenzo: RT for Justin Bieber Fav for One Direction RT or Fav? #musicfans #PeoplesChoice http://t.co/DV08b6G36j"
7	#ipad #ipadgames #ipadgames #gameinsight	My team has won on 'Arizona - Ranch'! Join us! http://t.co/WArGpigcxP #iPad, #iPadGames, #GameInsight
8	#android #androidgames #androidgames #gameinsight	I can go Alice's Room! Have you already discovered this location? http://t.co/PTX4JRpRp3 #Android #AndroidGames #GameInsight
9	gold coins	Over the moon I've won a other coupon, more gold coins
10	@youtube video	RT @360MrJamz: I liked a @YouTube video from @assassin_sc http://t.co/784jPColjN Battlefield 4 VS Cod Ghosts

5. CONCLUSION & FUTURE WORK

In this project, we have detected trends in Twitter and also found their representative tweets. However, only bigram is considered to be the form of trends. In the future, other forms would be considered such as unigrams or trigrams. If we take them into consideration at the same time, we also need to develop a algorithm to rank and combine all detected trends from unigrams, bigrams and trigrams.

On the other hand, only frequency is considered in the project. In the future, we could directly identify the topic of each tweet and group tweets by their topics. In the end we can identify the topics with largest number of tweets as trends.

Reference

[1] Haewoon Kwak ,Changhyun Lee,Hosung Park,Sue Moon What is Twitter, a social network or a news media? Proceedings of the 19th international conference on World wide web Pages 591-600