YouRank: Let User Engagement Rank Microblog Search Results

Wenbo Wang[α], **Lei Duan**[β], **Anirudh Koul**[β], **Amit P. Sheth**[α]
[α]Kno.e.sis Center, Wright State University, Dayton, OH 45435 USA
[β]Microsoft Corp., Sunnyvale, CA, 94089 USA
{*wenbo, amit*}*@knoesis.org* {*lei.duan, akoul*}*@microsoft.com*

Abstract

We propose an approach for ranking microblog search results. The basic idea is to leverage user engagement for the purpose of ranking: if a microblog post received many retweets/replies, this means users find it important and it should be ranked higher. However, simply applying the raw count of engagement may bias the ranking by favoring posts from celebrity users whose posts generally receive a disproportionate amount of engagement regardless of the contents of posts. To reduce this bias, we propose a variety of time window-based outlier features that transfer the raw engagement count into an importance score, on a per user basis. The evaluation on five real-world datasets confirms that the proposed approach can be used to improve microblog search.

Introduction

The microblog is becoming an increasingly important source of information, because it complements the traditional Web with real time information. For example, several eye witnesses posted details about the tragic plane crash at SFO airport on Twitter on July 6th 2013, within the first few minutes of the crash. Due to the large volume of daily microblog posts, microblog search becomes an essential way for people to find relevant information. In this paper, we take Twitter, a popular micrblogging site, as our test bed for microblog search.

Twitter search is very challenging for a number of reasons including: i) ranking needs to be made on top of limited content: each tweet is limited to 140 characters in length; ii) the problem space is vast: recent traffic has been about 500 million posts per day . These challenges lead us to make two main observations: i) tweets are so short in length that *humans perform better at reading, understanding, assessing and ranking tweets than machines*, as humans can easily grasp the ideas, facts, and humor conveyed through these 140 characters, while a machine cannot. *Consequently, can we involve humans into the ranking process?* ii) due to the vast volume of newly posted tweets every day, there can be a large number of relevant candidate tweets for a given query. *How do we differentiate the most important candidates from others that are also relevant?*

In recent years, learning-to-rank has been the most popular framework for Twitter Search, where features are extracted from labeled tweets, and ranking models are automatically constructed on top of these features. So far, much attention has been paid to content-based features (Miyanishi et al. 2011; Metzler and Cai 2011). However, it is difficult to decide whether a tweet should be ranked higher than another tweet by comparing only their contents due to the 140 character limit. Some studies have explored incorporating raw counts of user engagement (i.e., retweets, replies) as features, but this may bring ranking bias, since popular Twitter users usually receive a large amount of user engagement, no matter what they tweet. For example, just a simple *'Good day'* tweet from Justin Bieber was retweeted almost 90 K times in total[1].

In this paper, we propose **YouRank**, a user engagement-based approach for ranking microblog search results. We call it YouRank because engagement from each user (**You**) plays an important role in deciding the ultimate **Rank**ing. We first study distributions of user engagement in one day's sample data from Twitter to give an overview of user engagement activities. Then, we focus on user engagement with tweets that were posted by the same author, and show that the volume of user engagement can reflect the importance of tweets, on a per user basis. We propose a variety of time window-based outlier features to capture the importance of tweets, and experiment these features on five real world datasets.

Related Work

We focus on existing studies that apply the learning to rank approach for twitter search, which is the most relevant to our paper. Features that have been explored in these studies can be broadly grouped into the following categories:

Content features: How much is the information overlap between the query and tweets (Miyanishi et al. 2011)? Are there many misspelled words (Duan et al. 2010)? What is the polarity of the tweet (Bermingham and Smeaton 2012)? Other content features that have been used, include whether a tweet has a URL/mention and the number of words/characters in a tweet (Miyanishi et al. 2011; Duan et al. 2010).

[1]https://twitter.com/justinbieber/status/308676221250723840

Time features: The time difference between a tweet's posting time and the query time (Metzler and Cai 2011).

Author features: The hypothesis is that active influential users tend to provide more trustworthy, high quality tweets, which can be captured by different author attributes, e.g., follower count, friend count, whether or not it is an account verified by Twitter, etc (Miyanishi et al. 2011).

User engagement features: The more users that engage with a tweet, the more popular a tweet is. To characterize engagement from users, the following features have been explored: the number of times the tweet has been retweeted, the sum of follower counts of people who retweeted the tweet and the number of times the author had been mentioned in other tweets, etc. (Duan et al. 2010).

Other features: A relevant tweet may be missed by simple word matching if it does not contain any of the keywords in a query. To solve this problem, query expansion (Metzler and Cai 2011) and Latent Semantic Indexing (Miyanishi et al. 2011) have been studied.

In this paper, we define and extract tweet importance out of user engagement from the perspective of outliters, which, to our knowledge, has not been studied so far.

User Engagement Analysis

On Twitter, if one likes a tweet and would like to engage with it, he/she can *reply* to the tweet or re-post it on his/her timeline (i.e., *retweet*). To study reply and retweet activities, we obtained a 1% sample of Twitter's data for a single day and grouped tweets into different buckets, based on raw count of retweets and replies separately. E.g., bucket [1,10) contains all the tweets whose retweet/reply count is greater than or equal to 1 and less than 10. Table 1 shows the size of each bucket in percentages. We have the following observations: i) despite the 500 million tweets were posted per day, it is instructive to see that a large number of tweets did not attract any engagement: 79.418% of tweets did not receive a single retweet and 99.139% of tweets did not get any reply. The fact that Twitter users show little or no interests in these tweets suggests that we may eliminate them from the provided search results. ii) there is more retweet engagement than reply engagement: 20.582% of tweets received at least one retweet, while 0.861% of tweets received at least one reply. We suspect this may be partly due to the fact that it takes more effort to reply than to click the retweet button.

User Engagement - An Example

Once an author posts a tweet, it will be displayed at the top of timelines of the author's followers. Because of this design, a large number of engagement is from the author's followers. Therefore, the raw count of engagement will be limited by the number of the author's direct followers. Hence, it provides little insight to directly compare the raw counts of user engagement between two tweets from different authors: Justin Bieber's tweets can easily get a half million retweets, while Tim Berners-Lee's tweets typically receive less than a few hundred retweets.

Figure 1 shows that if we focus on the same author, the comparison of user engagement with tweets shows an interesting pattern: the more user engagement, the more important a tweet is. The author, Josh Cox, is a marathon runner who posted one of the very first on-site tweets about the Boston Marathon explosion. Because of this, the top tweet received 292 retweets. In the middle tweet, he decided to give away some gear for free, which accumulated 72 retweets. At the bottom tweet, he called upon his followers to participate in a survey, and he got only 1 retweet. By checking his tweet history, we find that most of his tweets receive no more than five retweets. Hence, the 292 retweets (for the top tweet) and the 72 retweets (for the middle tweet) are outliers that received far more retweets than usual. Moreover, the further the outlier sits from its mean value in a positive direction, the more important the tweet is: as retweet count increases from the bottom to the top, the corresponding importance increases accordingly. Next, we will show how to statistically capture the importance of each tweet based on user engagement.

Table 1: User engagement analysis on 1% sample of one day Twitter data

Retweet count	0	[1,10)	[10,100)	[100,1000)	[1000,)
Tweets (in %)	79.418	17.985	2.162	0.406	0.028
Reply count	0	[1,10)	[10,100)	[100,1000)	[1000,)
Tweets (in %)	99.139	0.823	0.035	0.003	0.000

Figure 1: Tweets from same author with various user engagement: as user engagement (retweet counts) of each tweet decreases from top to bottom, the importance of each tweet decreases accordingly.

Outlier-based User Engagement Features

For the sake of simplicity, we take retweet engagement as an example to show how to extract user engagement features; similar procedures can be done to extract reply features. Let $X = \{x_1, x_2, ..., x_n\}$ be a random variable measuring retweet engagement received by all the tweets from user u_i during a given period in the past, where $x_k (1 \leq k \leq n)$ measures user engagement of the k_{th} tweet. Then, the mean and standard deviation of engagement X will be:

$$\mu_X = \frac{x_1 + x_2 + ... + x_n}{n}, \sigma_X = \sqrt{\frac{\sum_{k=1}^{n}(x_k - \mu_X)^2}{n}} \tag{1}$$

where n is the number of user engagement observations. Given a new user engagement x, outlier feature o_x is defined by how many standard deviations away is x from mean engagement μ_X in history:

$$o_x = \frac{x - \mu_X}{\sigma_X} \tag{2}$$

The outlier feature is 0 when retweet engagement x just reaches mean user engagement μ_X in history; the outlier feature becomes a large positive number when x far exceeds μ_X, which is the case for the first two tweets in Figure 1. Later, we will propose three different ideas (x_{eq}, x_{fo} and x_{ra}) to quantify user engagement x, where each user engagement will be assigned different weights accordingly.

User u_i created a new post m_j at time t_c. Prior to query time t_q, a set of users $U = \{u_1, u_2, ..., u_{|U|}\}$ retweeted m_j. For user $u_p \in U(1 \leq p \leq |U|)$, let f_p be the number of followers, and g_p be the number of friends that u_p is following. If we assume each user has equal weight, user engagement x can be measured by the number of users who retweeted m_j *(equal weight feature)*:

$$x_{eq} = \sum_{p=1}^{|U|} 1 = |U| \tag{3}$$

A retweet from a more influential user (i.e., one with more followers) is usually a more important endorsement than that from a less influential user. So instead of assigning equal weights to every user, we assign higher weights to users with more followers *(follower weight feature)*:

$$x_{fo} = \sum_{p=1}^{|U|} \log(f_p + 1) \tag{4}$$

where we apply $+1$ to deal with the boundary case of a user having 0 followers. The larger the number of followers that retweeting users have, the higher the retweet engagement. In practice, we observe that some users are not influential, but they have many followers: they actively follow other users and other users tend to follow them back out of courtesy. So, we assign higher weights to users based on the ratio of the number of followers fp to the number of friends g_p *(ratio weight feature)*:

$$x_{ra} = \sum_{p=1}^{|U|} \log\left[\frac{f_p + 1}{g_p + 1}\right] \tag{5}$$

where we apply $+1$ to deal with cases of a user having no followers or is following no one.

Experiments

Data collection We filtered Twitter firehose data by predefined keywords for five events. We performed undersampling of tweets that receive little to no engagement so that the number of tweets that need to be labeled becomes manageable. In the end, we got: 1,669 tweets for *SFO plane crash*, 1,645 tweets for the *birth of UK royal baby*, 1,026 tweets for *Snowden's asylum in Russia*, 991 tweets for *the Moto X phone release* and 1,274 tweets for *the Castro Kidnap Trial*. To simulate the way people read tweets online, we showed full tweet information to annotators (similar to screen shots of tweets in Figure 1), including: text, author picture, screen name, retweet count, reply count, timestamp, etc. Each tweet was labeled with a number between 0 and 3 based on its importance to a query: the higher the score, the more important a tweet is considered. We simulated a user's search for up-to-date information about these events by repeatedly querying at an interval of 15 minutes. Assuming that users prefer the freshest possible results, we decreased the label of a tweet based on the time gap between its creation time and query time. The label was decreased by 1 if the gap was between 30 and 150 minutes, by 2 if the gap was between 150 and 310 minutes, and by 3 if the gap was larger than 310 minutes.

Base features We measured a tweet's freshness $t_f = t_q - t_c$ (in minutes) by the gap between its posting time t_c and query time t_q. We used the number of words l_w and the number of characters l_c to measure the length of a tweet.

Time window In practice, instead of using all engagement that occurred between $[t_c, t_q]$, we take user engagement in the *begin window* (first v minutes after m_j was posted on t_c: $[t_c, t_c + v]$) and *recent window* (past s minutes prior to query time t_q: $[t_q - s, t_q]$). Because we believe that engagement in the begin window can be a good indicator of overall engagement. Also, we want to check whether the tweet drew attention prior to query time (recent window). We use b and r to indicate begin window and recent window, respectively. In this paper, we set the lengths of both begin window (v) and recent window (s) to 10 minutes.

Engagement features Subscripts $eq/fo/ra$ indicate equal weight-based (Formula 3), follower weight-based (Formula 4) and ratio weight-based (Formula 5) features, respectively. For example, x_{eq} is the raw retweet count, accumulated before query time; x_{eq}^b is the raw retweet count occurring in begin window after a tweet was posted; $\mu_{X_{eq}^b}$ is its mean value in history; $\sigma_{X_{eq}^b}$ is its standard deviation in history; $o_{X_{eq}^b}$ is the corresponding outlier feature, measured how many standard deviations away is x_{eq}^b from the mean $\mu_{X_{eq}^b}$. Similarly we have $\mu_{X_{eq}^r}$, $\sigma_{X_{eq}^r}$, $o_{x_{eq}^r}$ to extract retweet engagement in recent window.

Besides retweet engagement features, we used similar features for reply (Y) engagement, e.g., $\mu_{Y_{eq}^b}$, $\sigma_{Y_{eq}^b}$, $o_{y_{eq}^b}$, etc. We scanned the one-week history of tweets for all the users to calculate mean and standard deviation of different engagement measures.

Learning algorithm We applied a proprietary boosted decision trees-based ranking algorithm. We applied dataset-level five-fold cross validation: each time, we used four datasets for training and held the remaining one for testing, then picked four different datasets for training. This was continued until every dataset had been held out for testing once. We calculated average score of Normalized Dis-

Table 2: Retweet features in recent and begin windows. For retweets, ratio weight-based outlier features from recent window perform best, suggesting that retweets from recent window reflect up-to-date user interests.

Features		Averaged NDCG			
		@1	@3	@5	@10
t_f l_w l_c	$\mu_{X^r_{eq}}$ $\sigma_{X^r_{eq}}$ $o_{x^r_{eq}}$	0.5938	0.6071	0.6049	0.6182
	$\mu_{X^r_{fo}}$ $\sigma_{X^r_{fo}}$ $o_{x^r_{fo}}$	0.6082	**0.6284**	0.6335	0.6438
	$\mu_{X^r_{ra}}$ $\sigma_{X^r_{ra}}$ $o_{x^r_{ra}}$	**0.6174**	0.6227	**0.6352**	**0.6465**
t_f l_w l_c	$\mu_{X^b_{eq}}$ $\sigma_{X^b_{eq}}$ $o_{x^b_{eq}}$	0.5283	0.5467	0.5608	0.5904
	$\mu_{X^b_{fo}}$ $\sigma_{X^b_{fo}}$ $o_{x^b_{fo}}$	0.5439	0.5489	0.5666	0.5884
	$\mu_{X^b_{ra}}$ $\sigma_{X^b_{ra}}$ $o_{x^b_{ra}}$	**0.5802**	**0.5930**	**0.6010**	**0.6228**

counted Cumulative Gain (NDCG) at ranking k (k=1,3,5,10) on five datasets for evaluation purposes.

Retweet features in recent and begin windows: Table 2 compares the effectiveness of retweet engagement features extracted from recent and begin windows. The top half shows results of applying three variations of retweet features extracted from recent window, while the bottom half shows results of applying corresponding features from begin window. In both windows, applying ratio weight-based outlier features usually gives the best results, which indicates that ratio weight-based outlier features can better reflect tweet importance from retweet engagement. Moreover, retweet features extracted from recent window give better results than those extracted from begin window, which indicates that recent retweet engagement can provide the latest user interests and thus better results.

Reply features in recent and begin windows: Table 3 compares the effectiveness of reply engagement features extracted from recent and begin windows. The top half shows results of applying three variations of reply features extracted from recent window, while the bottom half shows results of applying corresponding features from begin window. In both windows, applying equal weight-based outlier features gives the best results, which indicates that equal weight-based outlier features can better reflect tweet importance from reply engagement. Since replying is a far more time consuming engagement than retweeting, most users do not reply unless they are highly engaged by a tweet, thus simple reply count itself is a good indicator. Similar to retweet features, reply features extracted from recent window give better results than those extracted from begin window. The reason is that recent reply engagement reflects up-to-date user interests.

Combine outlier-based features: Table 4 combines retweet count and the best outlier-based features from Tables 2 and 3. It shows that applying retweet count (1st row) gives very good results. We suspect social influence (Zhu, Huberman, and Luon 2012) to be the reason. When annotators were labeling tweets, it was possible that they could get influenced by displayed retweet count. Still, applying outlier-based features on top of base features and retweet count gives the best result, which shows that outlier-based features are complimentary to retweet count features.

Table 3: Reply features in recent and begin windows. For replies, equal weight-based (instead of ratio weight-based) outlier features from recent window perform best. Replying takes much more time than retweeting and users do not reply to a tweet unless it is important, thus simple reply count is a good indicator.

Features		Averaged NDCG			
		@1	@3	@5	@10
t_f l_w l_c	$\mu_{Y^r_{eq}}$ $\sigma_{Y^r_{eq}}$ $o_{y^r_{eq}}$	**0.5687**	**0.5708**	**0.5762**	**0.5952**
	$\mu_{Y^r_{fo}}$ $\sigma_{Y^r_{fo}}$ $o_{y^r_{fo}}$	0.5420	0.5613	0.5724	0.5949
	$\mu_{Y^r_{ra}}$ $\sigma_{Y^r_{ra}}$ $o_{y^r_{ra}}$	0.5229	0.5465	0.5554	0.5806
t_f l_w l_c	$\mu_{Y^b_{eq}}$ $\sigma_{Y^b_{eq}}$ $o_{y^b_{eq}}$	**0.5537**	**0.5565**	**0.5653**	**0.5904**
	$\mu_{Y^b_{fo}}$ $\sigma_{Y^b_{fo}}$ $o_{y^b_{fo}}$	0.5213	0.5321	0.5448	0.5766
	$\mu_{Y^b_{ra}}$ $\sigma_{Y^b_{ra}}$ $o_{y^b_{ra}}$	0.4926	0.5204	0.5391	0.5709

Table 4: Combine outlier-based features

Features		Averaged NDCG			
		@1	@3	@5	@10
t_f l_w l_c	x_{eq}	0.6578	0.6612	0.6687	0.6806
	x_{eq} $\mu_{X^r_{ra}}$ $\sigma_{X^r_{ra}}$ $o_{x^r_{ra}}$	0.6691	0.6652	0.6692	0.6826
	x_{eq} $\mu_{X^r_{ra}}$ $\sigma_{X^r_{ra}}$ $o_{x^r_{ra}}$ $\mu_{Y^r_{eq}}$ $\sigma_{Y^r_{eq}}$ $o_{y^r_{eq}}$	**0.6817**	**0.6731**	**0.6809**	**0.6941**

Conclusion

We explored ranking microblog search results by leveraging user engagement with tweets relative to each author. We proposed a series of time window-based outlier features that capture tweet importance out of user engagement. Our experiments show that ratio weight-based retweet features and equal weight-based reply features in a recent time window can reflect the latest user interests and improve microblog search results.

Acknowledgement

The work was primarily done during the first author's internship at Bing Social. It was also supported in part by US National Science Foundation grant IIS-1111182.

References

Bermingham, A., and Smeaton, A. F. 2012. An evaluation of the role of sentiment in second screen microblog search tasks. In *ICWSM*.

Duan, Y.; Jiang, L.; Qin, T.; Zhou, M.; and Shum, H.-Y. 2010. An empirical study on learning to rank of tweets. In *ACL*, 295–303.

Metzler, D., and Cai, C. 2011. Usc/isi at trec 2011: Microblog track. In *TREC*.

Miyanishi, T.; Okamura, N.; Liu, X.; Seki, K.; and Uehara, K. 2011. Trec 2011 microblog track experiments at kobe university. In *TREC*.

Zhu, H.; Huberman, B.; and Luon, Y. 2012. To switch or not to switch: understanding social influence in online choices. In *SIGCHI*, 2257–2266. ACM.