Understanding Factors That Affect Web Traffic via Twitter

Chunjing Xiao[1,2,3(✉)], Zhiguang Qin[2], Xucheng Luo[2], and Aleksandar Kuzmanovic[3]

[1] School of Computer and Information Engineering, Henan University, Kaifeng, China
chunjingxiao@gmail.com
[2] School of Information and Software Engineering, UESTC, Chengdu, China
{qinzg,xucheng}@uestc.edu.cn
[3] Department of EECS, Northwestern University, Evanston, USA
akuzma@cs.northwestern.edu

Abstract. Currently, millions of companies, organizations and individuals take advantage of the social media function of Twitter to promote themselves. One of the most important goals is to attract web traffic. In this paper, we study the problem of obtaining web traffic via Twitter. We approach this problem in two stages. First, we analyze the correlation between important factors and the click number of URLs in tweets. Through measurements, we find that the commonly accepted method, increasing followers by reciprocal exchanges of links, has limited effects on improving the number of clicks. And characteristics of tweets (such as the presence of hashtags and tweet length) exert different impacts on users with different influence levels for obtaining the click number. In our second stage, based on the analyses, we introduce the Multi-Task Learning (MTL) to build a model for predicting the number of clicks. This model takes into account the specific characters of users with different influence levels to improve the predictive accuracy. The experiments, based on Twitter data, show the predictive performance is significantly higher than the baseline.

Keywords: Popularity · Prediction · Web traffic · Twitter

1 Introduction

Web traffic is one of the key indicators of a website's success, and most of individuals and companies rank websites mainly on the basis of their web traffic, such as the well-known Alexa[1]. Thus, website owners constantly strive to increase their web traffic by implementing various strategies, such as advertisements or audience analyses. The popularity of Twitter provides a new mean of promoting websites. In fact, Twitter has become new influential media for information sharing [18]. Thus, millions of organizations, companies, and individuals register

[1] http://www.alexa.com/.

© Springer International Publishing AG 2016
W. Cellary et al. (Eds.): WISE 2016, Part II, LNCS 10042, pp. 105–120, 2016.
DOI: 10.1007/978-3-319-48743-4_9

accounts on that and publish their URLs to attract web traffic, and Twitter has been a beneficial platform for a number of the websites [3,21].

Although the capability of Twitter to generate web traffic is widely accepted, little work focuses on examining the factors that affect obtaining web traffic via Twitter, and a serial of questions in this field keep unknown. For example, the previous work shows that the number of followers does not necessarily reflect their influence in terms of retweets or mentions [8], however, the reason is still unknown. To increase the follower number, users may randomly follow others in the hope that they follow back [23], and this phenomenon is called reciprocal links by Ghosh *et al.* [12]. Whereas it is not clear whether these types of followers can enhance content diffusion. In addition, there is a need to understand how hashtags and mentions in tweets impact the click number of URLs and whether these factors have the predictive power of the click number.

Our approach to answering these questions begins with an extensive characterization of important affecting factors, such as the follower number, presences of hashtags and mentions, as well as tweet length. To understand the impact of followers, we analyze the correlation between the numbers of clicks and followers, and find their correlation is not as strong as expected, which is consistent with the finding in [8]. However, the difference in the numbers of followers and reciprocal links has an obviously higher coefficient of correlation with the number of clicks. Therefore, reciprocal links are a key reason why the number of user followers does not necessarily reflect their influence in terms of the click number. And our further analyses also show reciprocal links have limited effects on content diffusion, although it is widely used to increase the number of followers.

Besides, we exploit the effect of tweet characteristics on the click number, such as the presences of hashtags and mentions and tweet length. And we find that the correlation between the number of clicks and these characteristics exhibits different trends for users with different influence levels (Here the influence level is measured by the difference between the numbers of followers and reciprocal links). Specifically, in terms of hashtags, URLs in tweets with hashtags obtain more clicks for users with low influence, but less for users with high influence. And for tweet length, when tweets have 50 and 120 characters, their URLs attract a similar maximum number of clicks for users with low influence. However, it is hardly affected by tweet length for users with high influence.

The second part of work for answering these questions is to conduct prediction about the number of clicks. Because the above analyses show that hashtags, mentions and tweet length exert different effects on users with different influence levels for obtaining the number of clicks, the model should take into account these different effects to improve predictive performance. To this end, we cast the predictive problem as a Multi-Task Learning (MTL) problem.

Specifically, we build a SVM+MTL model to predict the number of clicks. In this model, users are placed into different groups based on their influence levels, and each group is treated as a task. The model considers both the common properties of all the users and specific characters of users with different influence levels to improve predictive performance. Based on the Twitter data, the experiment results show the accuracy of our model is significantly higher than the baseline.

2 Related Work

There is little work focusing on the number of clicks on Twitter, however, the number of clicks, to some extent, can be a measure of popularity. Therefore, our work is related to the fields of popularity, which mainly consist of two threads of work: analyzing factors that affect popularity and predicting popularity in social media.

For the analyses of affecting factors, Suh *et al.* [24] examine a number of features that might affect the retweets. They find that URLs, hashtags and the numbers of followers and friends affect the retweets. Comarela *et al.* [10] identify factors that influence user response or retweet probability. They find that some basic textual characteristics, such as message size and the presence of hashtags, mentions and URLs, affect the replies or retweets. Liu *et al.* [20] evaluates eleven extrinsic factors that may influence the response rate in social question and answering from Sina Weibo. They show that the features, such as the number of followers, frequency of posting, hashtags and emotion, can be used to predict the number of responses. Apart from microblogs, Khosla *et al.* [16] and Bakhshi *et al.* [4] study the important factors that impact the popularity of images and quality of reviews respectively. Compared with these studies, we, beyond analyzing basic factors, explore the reason of existed phenomenons, and study whether tweet characteristics (such as hashtags, mentions and tweet length) exert different impacts on URLs in tweets of users with different influence.

For the popularity prediction, the studies fall into two main genres: conducting prediction before and after content publication. For the former, because the distribution of cascade sizes is very skewed, predicting the exact number of cascade sizes remain relatively unreliable [5]. Hence, rather than predicting exact integer values, most of the researchers define several categories to represent the popularity levels and predict which categories contents will belong to. For example, Hong *et al.* [13] define several categories to represent popularity of tweets and use logistic regression to predict the categories of tweets. Jenders *et al.* [15] predict whether a given tweet will be more frequently retweeted than a certain threshold. They firstly analyze the correlation between the retweet frequency and user features, and then they use the probabilistic models to conduct prediction. Vasconcelos *et al.* [27] categorize reviews into various popularity levels and predict the levels using multivariate linear regression and SVM models.

To achieve higher accuracy of prediction, many studies predict popularity after content publication. In this case, the early number of retweets or views within a short period after content publication can be used for prediction. Some work uses the early information to predict the exact integer values. For example, Szabo *et al.* [25] find the early number of retweets or views is strongly correlated with the later number on Digg and YouTube, and predict the popularity of content based on this finding. Kupavskii *et al.* [17] and Bao *et al.* [7] improve the performance of popularity prediction by exploiting the features of the cascade flow and structural characteristics respectively And Zhao *et al.* [28] develop a self-exciting Point Process Model to predict tweet popularity.

Other work still uses the early information to predict the categories which represent the popularity levels. For example, Gao *et al.* [11] predict whether a tweet will be popular based on temporal features of first 10 retweets using the bagged decision trees model. Given a cascade that currently has size k, Cheng *et al.* [9] predict whether it grow beyond the median size 2k by using the temporal and structural features. They use a variety of learning methods, including logistic regression classifier, naive Bayes and SVM for the prediction.

The method of popularity prediction after content publication generally achieves better performance than that of before content publication, but it is still crucial for the prediction before content publication. Because (*i*) publishers always want to know popularity of their contents before publication, (*ii*) and this method can clearly measure the importance of static factors in affecting popularity. Therefore, we conduct prediction before content publication. And our MTL-based predictive model is built based on our findings. To the best of our knowledge, we are the first to predict popularity using MTL.

3 Data Description

3.1 Background of URL Clicks

In this section, we present information about clicks of short URLs. Due to the limitation of tweet length, users tend to publish shortened URLs on Twitter. Therefore, the service of shortening long URLs is provided by many companies, and Bitly is among the most popular ones. Furthermore, Bitly APIs[1] provide the information about the click number of URLs in tweets. These number can be classified into two types: the *exact click number* referring to the number of clicks from a given tweet of the user; the *global click number* referring to the number of clicks from all the domains and platforms, including Twitter, Facebook and so on. For these two kinds of numbers, the exact click number can be considered as the ability of the tweet to attract web traffic. Therefore, the exact click number are used as the standard for analyzing factors that affect web traffic via Twitter. And the global click number can be used to reflect the popularity of the tweet content, and will be used as one of the features to predict the exact click number. Below the click number will refer to the exact click number for simplicity.

3.2 Twitter Data

As our goal is to analyze users who are aiming to attract web traffic via Twitter, we need to select users who tend to publish tweets with short URLs. In our study, we only select short URLs hosted by Bitly, because their exact click number can be obtained, and they are the most popular ones, taking about 50 % of all the URLs in Twitter [3].

[1] http://dev.bitly.com/api.html.

To select targeted users, we firstly extract domains hosted by Bitly based on a random sample of public tweets (around 790 million) collected by Twitter streaming APIs. And we obtain 6,524 domains hosted by Bitly, including many well-known companies and organizations, such as nyti.ms (New York Times), wapo.st (Washington Post) and es.pn (ESPN). Secondly, from these 790 m tweets, we extract the users whose language is English and whose tweets include at least one short URL hosted by Bitly. Base on this, we further select users who tend to publish Bitly URLs and tend to increase their websites via Twitter. According to these rules, we select users whose ratios of Bitly URLs are more than 50 %, and whose domain focuses are more than 50 %. Here the domain focus is defined as the degree of short URLs redirecting to the same domain, and can be calculated as follows: $D_i = \frac{1}{V_i} \max_k v_{ik}$, where V_i refers to the summary of URLs of user i, and v_{ik} refers to the number of URLs with the domain k of user i. If all the URLs published by a user redirect to one domain, its domain focus will be 1. Finally, 214,293 users are selected as our targeted users.

For these selected users, by Twitter APIs, we download their profiles, followers, and friends, as well as their tweets during June 2014, as shown in Table 1. And by Bitly APIs, we collect the click information of short URLs extracted from these tweets.

Table 1. Summary of Twitter data

Number of users	214,293
Number of follower links	1,261,721,039
Number of friend links	180,803,547
Number of tweets	46,286,824
Number of short URLs	34,338,613

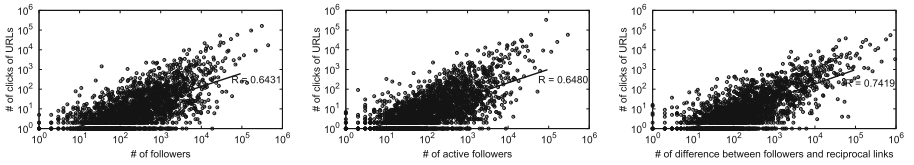
4 Analyses of Affecting Factors

We firstly describe the effect of user followers and tweet characteristics on the click number. The results in the section are the foundation for the predictive method, which is presented later.

4.1 The Role of User Followers

The number of followers is frequently used to gauge influence or reputation of users [14,23], and compare to other criterions, such as the number of retweets and mentions [8,18]. Therefore, we first analyze how the number of followers is correlated with the number of clicks received by URLs in tweets.

Figure 1(a) shows the correlation between the numbers of followers and URL clicks. The X-axis is the number of user followers, and the Y-axis is the sum of



(a) Followers (b) Active followers (c) Difference in followers and reciprocal links

Fig. 1. The correlation between the numbers of followers and clicks

the number of clicks. This figure shows that the coefficient of linear correlation, 0.64, is not as high as expected. This finding is consistent with the previous work, which shows that popular users with a high number of followers do not necessarily have high influence in terms of retweets or mentions [8] and the global click number of short URLs [22].

This observation raises the question why the number of followers is not very strongly correlated with the number of clicks. To address this question, we conduct analyses from two perspectives. (*i*) How do inactive followers affect the relationship between the numbers of followers and clicks? Thomas *et al.* [26] show that numerous accounts on Twitter have been suspended because of spamming issues or similar reasons. Moreover, some users tend to register multiple accounts but use only a part of them or stop using Twitter. We, therefore, attempt to evaluate the correlation between the numbers of active followers and clicks to analyze the effect of inactive followers. (*ii*) How do reciprocal links affect the relationship of the number of followers and clicks? On Twitter, a part of users randomly follow other users in the hope that they will follow back, whereas, some users join groups in which each member agrees to follow all of the other members in that group [23]. This phenomenon, which is called the reciprocal links by Ghosh *et al.* [12], is a way to increase one's number of followers, and users are recommended to increase their followers through this way to gain more web traffic [1]. However, whether these reciprocal links increase the diffusion effect of content remains unclear. Therefore, we attempt to explore the correlation between the numbers of reciprocal links and clicks to answer these questions.

To analyze the effect of inactive followers, we first identify whether a user is active. In general, Twitter regard users who log in at least once a month as active ones [2]. However, considering that we cannot obtain information about logging in activities, we regard users who publish at least one tweet, including any kind of tweets such as retweets and replies, within the last two months as active ones. After collecting the recent tweets by Twitter APIs, we can compute the active followers for each user. Further, the correlation between the numbers of active followers and clicks is plotted in Fig. 1(b). The coefficient of correlation, 0.6480, is nearly the same as that of the numbers of followers and clicks. We also analyze the correlation between numbers of active followers and all followers, and find that a strong linear relationship exists between them. These results suggest that inactive followers are not the main reason behind the moderate relationship between the number of followers and clicks.

For reciprocal links, we first collect the follower and friend list of each user, and then compute the intersection between the follower set and the friend set. This intersection is regarded as the reciprocal links. Based on this data, the correlation between the number of clicks and the difference in followers and reciprocal links is calculated, as shown in Fig. 1(c). Compared with Fig. 1(a) and 1(b), the points in Fig. 1(c) are centered around the straight line and a stronger correlation exists between the number of clicks and the difference in followers and reciprocal links. The coefficient, 0.7419, is approximately 10 % higher than that of followers and clicks. These results indicate that reciprocal links considerably affect the correlation between numbers of followers and clicks. And when reciprocal links are removed, the number of followers becomes more strongly correlated with the number of clicks.

To further evaluate the effect of reciprocal links in improving the number of clicks, we analyze the correlation between reciprocal links and clicks, as well as the correlation between reciprocal links and friends. the coefficient of the former, 0.1632, indicates that reciprocal links are not significantly correlated with clicks. The coefficient of the later is 0.9125, suggesting that most of the friends originate from reciprocal links.

Therefore, based on these analyses, we conclude that although reciprocal links are widespread to be used to increase the number of followers, they have limited effects on improving the number of clicks. And the difference of followers and reciprocal links can be a better measure of user influence. Hence, below this difference is regarded as the measure of user influence (levels), and user followers refer to this difference except Sect. 5.2.

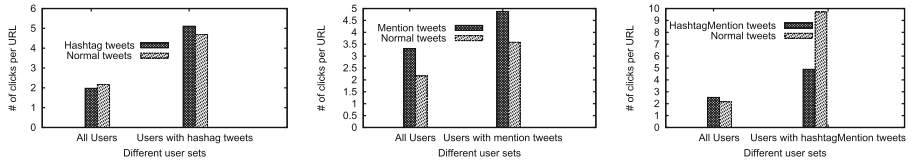
4.2 The Role of Tweet Characteristics

In this section, we analyze the impact of two kinds of tweet characteristics on the click number of URLs: tweet types (i.e., the presences of hashtags and mentions in tweets) and tweet length.

Tweet Types. On Twitter, tweets contain two widely used objects: hashtags and mentions. The former is used to mark keywords or topics in a tweet and to categorize messages, whereas the latter is a form of conversation on Twitter. Users are often encouraged to include hashtags to increase the click number of URLs on some web pages, such as [1]. Therefore, we explore how tweets that contain hashtags or mentions affect the number of clicks.

For this purpose, we group the tweets into four types: *hashtag tweets*, which are tweets that include at least one hashtag; *mention tweets*, which are tweets that include at least one mention; *hashtagMention tweets*, which are tweets that include both hashtags and mentions; and *normal tweets*, which are tweets without hashtags and mentions. To avoid any preference for users who tend (not) to publish more hashtag or mention tweets, we also analyze users with at least one hashtag, mention, or hashtagMention tweet.

Figure 2 shows the number of clicks per URL in different tweet types for different user sets. The Y-axis presents the average number of clicks for a given



(a) Hashtag tweets (b) Mention tweets (c) HashtagMention tweets

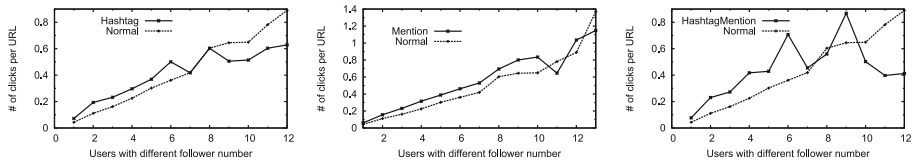
Fig. 2. Tweet type vs. number of clicks

user set. For hashtag tweets, shown in Fig. 2(a), the average number of clicks of hashtag tweets is lower than that of normal tweets for all users; however, the values are reversed for users with at least one hashtag tweet. Therefore, we cannot fully ascertain how tweets that contain hashtags correlate the number of clicks. For mention tweets, depicted in Fig. 2(b), the average number of clicks of mention tweets is always higher than that of normal tweets for both user sets. This result suggests that a positive correlation exists between tweets containing mentions and the number of clicks. For hashtagMention tweets, presented in Fig. 2(c), the trends are also inconsistent for different users.

Considering the unclear results about the effect of hashtags, we further explore whether hashtags and mentions exert the different effect on the number of clicks for users with different influence levels. For this purpose, we place users into buckets according to an interval of 200 followers. We use numbers to denote the buckets, i.e., the bucket 1 represents users with 0–200 followers, bucket 2 represents users with 200–400 followers, and so on. For each bucket, we group the tweets into four types: *hashtag tweets*, *mention tweets*, *hashtagMention tweets*, and *normal tweets*, and compute the average number of clicks for each type.

We compare the click number of the first three types of tweets with that of normal tweets, and the results are shown in Fig. 3. The figures do not show all of the buckets because of space constraints. The X-axis shows the bucket number. The Y-axis denotes the average number of clicks per URL for the particular bucket.

The results of the hashtag tweets are shown in Fig. 3(a). For the bucket 7 and 8 (referring to users with 1200–1400 and 1400–1600 followers respectively), the click numbers of the hashtag and normal tweets are very close. While, for the bucket 1 to 6, the hashtag tweets obtain a higher number of clicks than the normal tweets. However, the reverse is true for bucket 9 and beyond. These



(a) Hashtag tweets (b) Mention tweets (c) HashtagMention tweets

Fig. 3. Tweet type vs. number of clicks for users with different influence levels

results indicate that tweets with hashtags do not always achieve additional clicks, i.e., they can obtain more clicks for users with lower influence but not for that with higher influence.

For the mention tweets, presented in Fig. 3(b), from the bucket 1 to 10, the mention tweets generate a higher number of clicks than the normal tweets. While, for other buckets, the click numbers of both are interlaced with each other. That is, when users have less than roughly 1,800 followers, their tweets with mentions can attract additional clicks; however, when users have a higher number of followers, mentions do not contribute to improving the number of clicks. Affected by both hashtags and mentions, the hashtagMention tweets, presented in Fig. 3(c), exhibit a similar trend to hashtag tweets. The average number of clicks shows a small fluctuation because of their small number.

These results indicate that contrary to what people commonly assume, tweets with hashtags cannot always obtain more clicks. In fact, the hashtags and Mentions exhibit a different effect on users with different influence levels for obtaining the number of clicks.

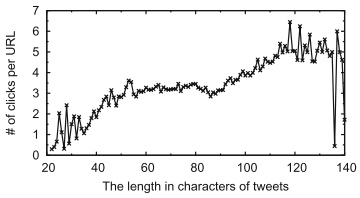


Fig. 4. All the users

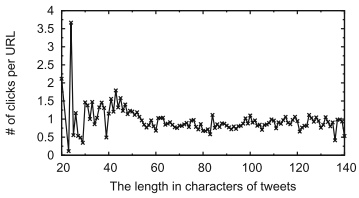
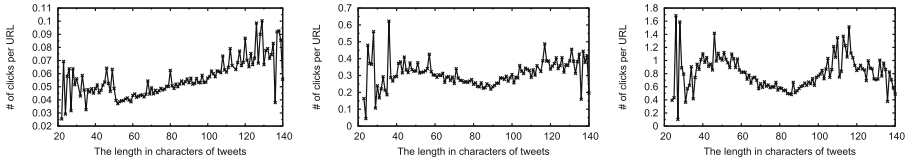


Fig. 5. Users with 2000–2200 followers

Tweet Length. Here, we explore the correlation between the tweet length and number of clicks. We first analyze this correlation for all users, and the results are shown in Fig. 4. The X-axis denotes the length of the tweets. The minimum length is 20 because the tweet contain the short URL with no less than 20 characters. The Y-axis refers to the average number of clicks with a particular length. From the figure, we can see that the number of clicks generally increases with the tweet length. And short URLs in tweets with approximately 120 characters tend to attract more clicks.

We further explore how the effect of tweet length differs for users with different influence levels. As in the previous section, we place the users into buckets according to an interval of 200 followers. For each bucket, we plot the correlations between the tweet length and number of clicks. By observing the trend of each figure, we find that these figures can be divided into two categories: users with 0–2,000 followers and users with more than 2,000 followers. For the former, all of the buckets exhibit a similar trend. In view of space constraints, we present the figures of three buckets: users with 1–200 followers, users with 600–800 followers and users with 1600–1,800 followers, as shown in Fig. 6. This category has the similar trend that the number of clicks exhibits a double hump phenomenon,



(a) users with 0-200 followers (b) Users with 600-800 followers (c) Users with 1600-1800 followers

Fig. 6. Tweet length vs. number of clicks for users with 0 to 2000 followers

and this trend becomes even more significant with the rise in the number of followers. For example, when the number of followers reaches 1,600–1,800, this trend becomes the most significant, and the two peaks of the click number are twice the minimum number of clicks. For users with more than 2,000 followers, we present the figures of users with 2,000–2,200 followers in Fig. 5, and all of the other buckets exhibit a similar trend. The number of clicks fluctuates because of the small amount of tweets when the tweet length is near 40, but it remains stable when the tweet length exceeds 50.

Basing on these results, we can conclude that the effect of tweet length on the number of clicks differs for users with different influence levels. Specifically, users with low influence, such as those with 0–2,000 followers, can be affected by tweet length, and URLs in tweets with around 50 to 120 characters tend to obtain more clicks. However, users with high influence, such as those with more than 2,000 followers, can hardly be affected by tweet length.

5 Methodology

5.1 Method of Prediction

The above analyses indicate that hashtags, mentions and tweet length place the different impact on users with different influence levels for obtaining the number of clicks. Therefore, the predictive model should take into account this different impact to achieve higher accuracy. However, a global model, such as logistic regression and SVM, will ignore this different impact. One way to address this challenge is to create and apply numerous models to the user sets with different influence levels. However, the data of some user sets, especially for user set with high influence levels, is very sparse and cannot build model accurately. Hence, to overcome this problem, we introduce the Multi-Task Learning (MTL) to predict the click number of URLs. MTL seeks to simultaneously learn the commonality as well as the differences between the multiple tasks. Therefore, we divide users into different groups based on their influence levels and treat prediction of each group as a task. And the MTL model is used to improve the performance by considering both the common properties of all users and specific characters of users with different influence levels. Here, we introduce an extension of SVM+ approach to multi task learning called SVM+MTL [19] to build the model.

In SVM+MTL, the training set T is the union of task specific sets $T_r = \{x_{ir}, y_{ir}\}_{i=1}^{l_r}$. For each task the learned weights vector is decomposed as $w + w_r, r \in (1, 2, ..., t)$ where w and w_r respectively model the commonality between tasks and task specific components. The optimization problem of SVM+MTL is formulated as follows:

$$\min_{w,b} \frac{1}{2}(w, w) + \frac{\beta}{2} \sum_{r=1}^{t} (w_r, w_r) + C \sum_{r=1}^{t} \sum_{i=1}^{l_r} \xi_{ir} \tag{1}$$

$$st : y_{ir}((w, \phi(x_{ir})) + b + (w_r, \phi_r(x_{ir})) + d_r) \geq 1 - \xi_{ir} \tag{2}$$

$$\xi_{ir} \geq 0, i = 1, ..., l_r, r = 1, ..., t \tag{3}$$

Here, all w_r's and the common w are learned simultaneously. β regularizes the relative weights of w and w_r's. ξ_{ir}'s are slack variables measuring the errors w_r's make on the t data groups. y_{ir}'s denote training labels while C regulates the complexity and proportion of nonseparable samples.

The goal of SVM+MTL is to find t decision functions $f_r(x) = (w, \phi(x)) + b + (w_r, \phi_r(x)) + d_r, r = 1, ..., t$. Each decision function f_r comprises two parts: the common weights vector w with bias term b, and the group-specific correction function w_r with bias term d_r.

5.2 Feature Spaces

In this section, we introduce the features which are used in the predictive model, including the attributes of user influence, publishing behavior, and short URLs.

Features of user influence describe the characteristics of the social topology of users. Based on the user profiles we can download by Twitter APIs, we use the metadata relative to user influence as the features, such as the number of followers, friends, lists and son on. Further, based on our analyses, we exploit the features related to influence: the active followers and differences between followers and reciprocal links, which can more accurately reflect user influence. The features are detailed in Table 2.

Features of publishing behavior are composed of the items which users can control when publishing tweets. The tweet characteristics, such as the presences of hashtags and mentions as well as tweet length, are also placed into this set, because users can determine whether their tweets include hashtags or mentions and how long their tweets are.

Features of short URLs describe the information collected by Bitly APIs. Among these features, the global click number of URLs can reflect the popularity of the tweet content, because the global click number is the sum of clicks from all the domains and platforms, and URLs in tweets are generally the key points of the tweets. The referrer number can also be a measure of popularity for URLs, because it means the sum of resources where clicks originated, i.e., the higher referrer number is, the more popular the URL is. Therefore, in the experiments later, we can evaluate whether the popularity of content has the predictive power of the exact click number by using the features about the global click number and referrer number.

Table 2. Summary of features

Feature sets	Name	Description
User influence	Followers	The number of followers
	Friends	The number of friends
	Lists	The number of lists including this user
	Active-followers	The number of active followers
	Diff-followers	Difference between followers and reciprocal links
Publishing behavior	Hashtags	The presence of hashtags in tweets
	Mentions	The presence of mentions in tweets
	Tweet length	The length of tweets
	Published time	The published time of tweets
	Average tweets	Average number of Tweets per day in our dataset
	Ratio of URLs	Ratio of numbers of tweets with URLs and all tweets
Short URLs	Global number	The global click number from all the domains and platforms
	Created time	Difference of tweet published time and URL created time
	Referrer number	The number of resources where clicks originated
	Domain ranking	Ranking in Alexa.com of the domain of expanded URLs

6 Prediction Results

Based on the method and features, we predict the click number of URLs in tweets. We describe the experiment setup and compare the results of SVM+MTL with the original SVM.

6.1 Experiment Setup

We conduct prediction before tweets publication, because compared with prediction after tweets publication, this kind of prediction can more clearly measure the factors that affect the number of clicks. As in [6,13,27], we define several categories to represent the levels of the click number and predict which categories the URL will belong to, instead of predicting the exact number. Because the latter is harder, particularly given the skewed distribution of popularity [5], and the former should be good enough for most purposes. Specifically, we divide URLs into five categories depending on the click number. That is we put URLs with 0, 1∼10, 11∼100, 101∼1,000, and more than 1,000 clicks into the category 1, 2, 3, 4 and 5 respectively. We select the same number of URLs for each category randomly, because the URLs in category 1 are dominant, accounting for around 70 % of all the URLs. When considering all the URLs for the experiments, the accuracy of prediction will reach 70 % even if we label all URLs as category 1.

The SVM+MTL takes into account both the common properties and specific characters of users with different influence levels. Hence, we place users into buckets according to an interval of 200 followers, and treat prediction of each bucket as a task. And the SVM is used as the baseline.

We use the classification accuracy and F-score to measure the performance. And the accuracy is defined as the proportion of true results in the population,

and the F-score combines recall and precision with an equal weight. And to evaluate the predictive performance, we randomly divide the URLs of each user into two sets: 50 % for training and 50 % for testing.

6.2 Results and Discussion

The accuracy and F-score of the SVM and SVM+MTL predictors are presented in Table 3 for the combination of the different feature sets. The best results (biggest accuracy) for each model are emphasized in boldfaced numbers. The first observation is that although the SVM model can perform reasonably well with around 69 % accuracy using all features, the performance of SVM+MTL, 81.77 %, is significantly higher than that of SVM. Besides, no matter which feature sets are used for prediction, the accuracies of SVM+MTL are always approximately 10 % higher than that of SVM. This indicates that grouping users based on user influence levels is appropriate for SVM+MTL, and by considering both the common properties of all users and specific characters of users with different influence levels, SVM+MTL can achieve expected predictive results.

In addition, we proceed to the feature set level to determine the importance of features in predicting the click number. Unsurprisingly, for the SVM+MTL model, the accuracies of using the influence feature set and behavior feature set arrive at 74.35 % and 72.46 % respectively, which suggest that both feature sets play an important role in predicting the levels. Interestingly, the features of short URLs cannot perform as better as that of user influence and behavior. Among the short URL features, both the global click number and referrer number can, to some extent, reflect the popularity of the URL content. But they fail to have a predictive power of the exact click number. This indicates that not every user can achieve more clicks by publishing popular URLs. We also compute the coefficient of correlation between the global click number and exact click number. The lower coefficient, about 0.38, also provides support for this point.

Table 3. The predictive results

Feature sets	SVM		SVM MTL	
	Accuracy (%)	F-score (%)	Accuracy (%)	F-score (%)
Influence	65.11	64.32	74.35	73.41
Behavior	62.33	61.48	72.46	72.14
URLs	57.68	58.13	68.74	69.85
Influence + behavior	66.04	66.84	75.26	74.11
Influence + URLs	65.3	65.91	76.18	77.14
Behavior + URLs	60.27	61.05	71.49	70.28
All features	**69.49**	69.74	**81.77**	81.37

7 Conclusions

In this paper, we conducted analyses and predictions about the click number of URLs in tweets. Through the analyses, we showed that the correlation of the click numbers and followers is not as strong as expected. This is due to reciprocal links, not inactive followers. And our further analysis suggested reciprocal links have limited effects on content diffusion, although it is widely used to increase the number of followers. We also found that hashtags and tweet length place different impacts on users with different influence levels for obtaining the number of clicks. Specifically, in terms of hashtags, URLs in tweets with hashtags achieve more clicks for users with low influence, but less for users with high influence. And for tweet length, URLs in tweets with 50 and 120 characters attract a similar maximum number of clicks. However, users with higher influence are hardly affected by tweet length. Based on these analyses, we built a SVM+MTL model to predict the click number. In this model, users with different influence levels are treated as different predictive tasks, and the commonality of all users and differences of users with different influence levels are learned simultaneously. The experiments, based on Twitter data, showed our predictive performance is significantly higher than the baselines.

Acknowledgment. This work is supported by the National Natural Science Foundation of China (No. 61402151 and 61272527), Science and Technology Foundation of Henan Province of China (No. 162102410010), and Open Research Fund of Network and Data Security Key Laboratory of Sichuan Province of China (No. NDS2015-02).

References

1. How to use twitter to increase web traffic. http://www.wikihow.com/Use-Twitter-to-Increase-Web-Traffic
2. Twitter announces 100 million active users. http://mashable.com/2011/09/08/twitter-has-100-million-active-users
3. Antoniades, D., Polakis, I., Kontaxis, G., Athanasopoulos, E., Ioannidis, S., Markatos, E.P., Karagiannis, T.: we.b: the web of short URLs. In: Proceedings of the 20th international conference on World Wide Web, pp. 715–724 (2011)
4. Bakhshi, S., Kanuparthy, P., Shamma, D.A.: Understanding online reviews: funny, cool or useful? In: Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work, pp. 1270–1276 (2015)
5. Bakshy, E., Hofman, J.M., Mason, W.A., Watts, D.J.: Everyone's an influencer: quantifying influence on Twitter. In: Proceedings of the Fourth ACM International Conference on Web Search and Data Mining, pp. 65–74 (2011)
6. Bandari, R., Asur, S., Huberman, B.A.: The pulse of news in social media: forecasting popularity. In: The Sixth International AAAI Conference on Weblogs and Social Media, pp. 26–33 (2012)
7. Bao, P., Shen, H.W., Huang, J., Cheng, X.Q.: Popularity prediction in microblogging network: a case study on Sina Weibo. In: Proceedings of the 22nd International Conference on World Wide Web, pp. 177–178 (2013)

8. Cha, M., Haddadi, H., Benevenuto, F., Gummadi, K.P.: Measuring user influence in twitter: the million follower fallacy. In: Proceedings of International AAAI Conference on Weblogs and Social Media (2010)
9. Cheng, J., Adamic, L., Dow, P.A., Kleinberg, J.M., Leskovec, J.: Can cascades be predicted? In: Proceedings of the 23rd International Conference on World wide web, pp. 925–936 (2014)
10. Comarela, G., Crovella, M., Almeida, V., Benevenuto, F.: Understanding factors that affect response rates in Twitter. In: Proceedings of the 23rd ACM Conference on Hypertext and Social Media, pp. 123–132 (2012)
11. Gao, S., Ma, J., Chen, Z.: Effective and effortless features for popularity prediction in microblogging network. In: Proceedings of the 23rd International Conference on World Wide Web, pp. 269–270 (2014)
12. Ghosh, S., Viswanath, B., Kooti, F., Sharma, N.K., Korlam, G., Benevenuto, F., Ganguly, N., Gummadi, K.P.: Understanding and combating link farming in the Twitter social network. In: Proceedings of the 21st International Conference on World Wide Web, pp. 61–70 (2012)
13. Hong, L., Dan, O., Davison, B.D.: Predicting popular messages in Twitter. In: Proceedings of the 20th International Conference Companion on World wide Web, pp. 57–58 (2011)
14. Hutto, C., Yardi, S., Gilbert, E.: A longitudinal study of follow predictors on Twitter. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, pp. 821–830 (2013)
15. Jenders, M., Kasneci, G., Naumann, F.: Analyzing and predicting viral tweets. In: Proceedings of the 22nd International Conference on World Wide Web, pp. 657–664 (2013)
16. Khosla, A., Das Sarma, A., Hamid, R.: What makes an image popular? In: Proceedings of the 23rd International Conference on World Wide Web, pp. 867–876 (2014)
17. Kupavskii, A., Ostroumova, L., Umnov, A., Usachev, S., Serdyukov, P., Gusev, G., Kustarev, A.: Prediction of retweet cascade size over time. In: Proceedings of the 21st ACM International Conference on Information and Knowledge Management, pp. 2335–2338 (2012)
18. Kwak, H., Lee, C., Park, H., Moon, S.: What is Twitter, a social network or a news media? In: Proceedings of the 19th International Conference on World Wide Web, pp. 591–600 (2010)
19. Lichen, L., Cherkassky, V.: Connection between SVM+ and multi-task learning. In: Proceedings of the International Joint Conference on Neural Networks, pp. 2048–2054 (2008)
20. Liu, Z., Jansen, B.J.: Factors influencing the response rate in social question and answering behavior. In: Proceedings of the 2013 Conference on Computer Supported Cooperative Work, pp. 1263–1274 (2013)
21. Rodrigues, T., Benevenuto, F., Cha, M., Gummadi, K., Almeida, V.: On word-of-mouth based discovery of the web. In: Proceedings of the 2011 ACM SIGCOMM Conference on Internet Measurement Conference, pp. 381–396 (2011)
22. Romero, D.M., Galuba, W., Asur, S., Huberman, B.A.: Influence and passivity in social media. In: Proceedings of the 2011 European Conference on Machine Learning and Knowledge Discovery in Databases, pp. 18–33 (2011)
23. Stringhini, G., Wang, G., Egele, M., Kruegel, C., Vigna, G., Zheng, H., Zhao, B.Y.: Follow the green: growth and dynamics in Twitter follower markets. In: Proceedings of the 2013 Conference on Internet Measurement Conference, pp. 163–176 (2013)

24. Suh, B., Hong, L., Pirolli, P., Chi, E.H.: Want to be retweeted? Large scale analytics on factors impacting retweet in Twitter network. In: Proceedings of the 2010 IEEE Second International Conference on Social Computing, pp. 177–184 (2010)
25. Szabo, G., Huberman, B.A.: Predicting the popularity of online content. Commun. ACM **53**, 80–88 (2010)
26. Thomas, K., Grier, C., Song, D., Paxson, V.: Suspended accounts in retrospect: an analysis of Twitter spam. In: Proceedings of the 2011 ACM SIGCOMM Conference on Internet Measurement Conference, pp. 243–258 (2011)
27. Vasconcelos, M., Almeida, J.M., Goncalves, M.A.: Predicting the popularity of micro-reviews: a foursquare case study. Inf. Sci. **325**, 355–374 (2015)
28. Zhao, Q., Erdogdu, M.A., He, H.Y., Rajaraman, A., Leskovec, J.: SEISMIC: a self-exciting point process model for predicting tweet popularity. In: Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 1513–1522 (2015)