Prediction of Re-tweeting Activities in Social Networks Based on Event Popularity and User Connectivity

Sayan Unankard(✉)

Information Technology Division, Faculty of Science,
Maejo University, Chiang Mai, Thailand
sayan@mju.ac.th

Abstract. This paper proposes an approach to predict the volume of future re-tweets for a given original short message (tweet). In our research we adopt a probabilistic collaborative filtering prediction model called Matchbox in order to predict the number of re-tweets based on event popularity and user connectivity. We have evaluated our approach on a real-world dataset and we furthermore compare our results to two baselines. We use the datasets crawled by the WISE 2012 Challenge (http://www.wise2012.cs.ucy.ac.cy/challenge.html) from Sina Weibo (http://weibo.com), which is a popular Chinese microblogging site similar to Twitter. Our experiments show that the proposed approach can effectively predict the amount of future re-tweets for a given original short message.

Keywords: Re-tweets · Prediction · Micro-blog · Social networks

1 Introduction

The prediction of message propagation is one of the major challenges in understanding the behaviors of social networks. In this work, we study that challenge in the context of the Twitter social network. In particular, our goal is to predict the propagation behavior of any given short message (i.e., tweet) within a period of 30 days. This is captured by measuring and predicting the number of re-tweets.

To model the re-tweeting activities, we use the datasets crawled by the WISE 2012 Challenge from Sina Weibo, which is a popular Chinese microbloging site similar to Twitter. In Sina Weibo, retweet mechanism is different from Twitter. In Twitter, users can only re-tweet a tweet without modifying the original tweet. However, in Sina Weibo user can modify or add information from other users' in the re-tweeting path in their own re-tweet.

The dataset that to be used in this challenge contains two sets of files. Firstly, Followship network, it includes the following network of users based on user IDs. Secondly, Tweets, it includes basic information about tweets (time, user ID, messages ID), mentions (i.e., user IDs appearing in tweets), re-tweet paths, and

© Springer International Publishing AG, part of Springer Nature 2018
P. Perner (Ed.): MLDM 2018, LNAI 10935, pp. 357–368, 2018.
https://doi.org/10.1007/978-3-319-96133-0_27

Table 1. Number of original messages re-tweeted in 30 days

Number of re-tweets	Original messages		Annotated with events	
	#messages	%	#messages	%
<10	42,551,891	94.749	882,191	2.073
10–99	2,171,214	4.835	65,809	3.031
100–499	173,803	0.387	5,464	3.144
500–999	10,283	0.023	400	3.890
1,000–4,999	2,838	0.006	158	5.567
5,000–9,999	26	0.00006	2	7.692
≥10,000	11	0.00002	1	9.091
Total	44,910,066	100.00	954,025	2.124

Table 2. Number of re-tweets in 10 levels within 30 days

Level	Number of re-tweets	%
1	107,025,967	56.056
2	49,401,724	25.874
3	16,934,845	8.869
4	8,045,285	4.213
5	4,196,992	2.198
6	2,315,732	1.212
7	1,294,638	0.678
8	746,494	0.390
9	428,158	0.224
10	240,606	0.126

whether containing links. User IDs and message IDs are anonymized. Content of tweets are removed, based on Sina Weibo's Terms of Services. Some tweets are annotated with events. For each event, the terms that are used to identify the event and a link to Wikipedia[1] page containing descriptions to the event are given. For the purpose of this challenge, 369 million messages and 68 million user profiles were extracted. The sizes of the followship dataset and the microblog dataset are 12.8 GB and 64.8 GB, respectively. It should be note that the dataset is not complete but it is sufficiently large to predict the re-tweeting behavior of users on Sina Weibo.

In preparation for the challenge, we further collected some statistical information for a better understanding of the available datasets. In particular, for the followship dataset (i.e., the who is following whom relationship), we found that the majority of users have less that 10 followers (approximately 91%) as shown in Fig. 1. Additionally, for the microblog dataset (i.e., whose tweets are

[1] https://wikipedia.org.

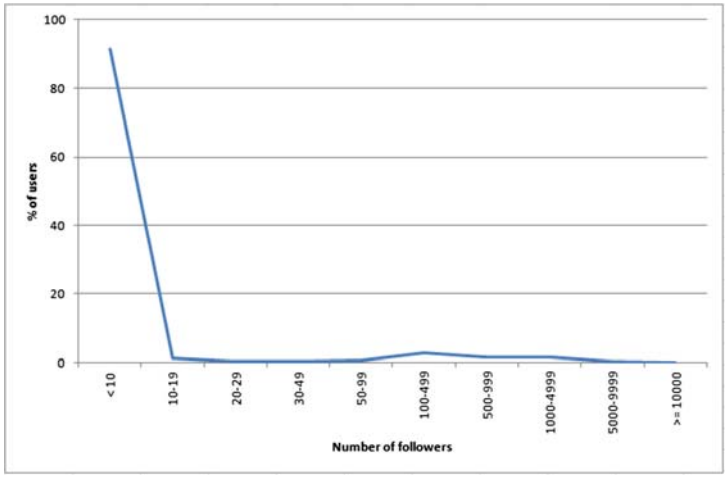

Fig. 1. User distribution based on numbers of followers

re-tweeted by whom), we ranked the distribution of the original tweets based on how many re-tweets they received within 30 days as shown in Table 1.

The table also shows the subsets of tweets that have been annotated with events. As the table shows, approximately 95% of the original tweets were re-tweeted less than 10 times, of which approximately 2% were annotated with events. In addition, most original tweets were re-tweeted in 3 levels within 30 days (approximately 91%) as shown in Table 2 and Fig. 2.

In order to understand the re-tweet activity, we also studied the re-tweet activity by day of the week and time of the day. We selected original tweets associated with events which have the number of re-tweets more than 100 for our study (6,934 messages). In Fig. 3, the graph shows the number of re-tweets per day of week. Based on a sample of tweets, Monday is the most popular day for re-tweet activity; followed by Tuesday and Friday. In Fig. 4, the chart shows the number of re-tweets per hour of the day. During the day, the most re-tweet activity happens from 10 a.m. to 12 p.m.

The contributions of this paper are summarized as follows: (1) An extensive statistical studies on the re-tweeting activities of users' behaviors in the widely used social network are provided. (2) The number of re-tweets is measured to understand the users' participation for spreading information in social network. (3) An approach to automatically predict the number of re-tweets over micro-blogs is proposed.

This paper is organised as follows, Sect. 2 is about related work. The proposed approach is presented in Sect. 3. In Sect. 4, we present the experimental setup and results, the conclusions are given in Sect. 5.

2 Related Work

Microblogging activities in social networks have been attracting growing attentions from researchers in Data Mining and Information Retrieval. One interesting

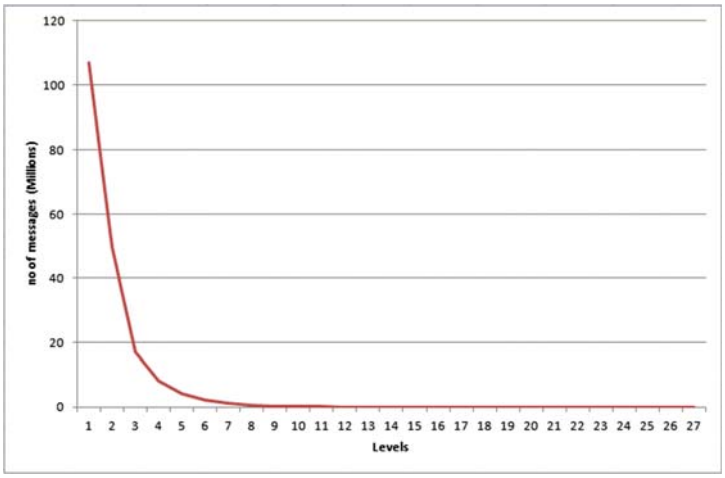

Fig. 2. Number of re-tweets in each level

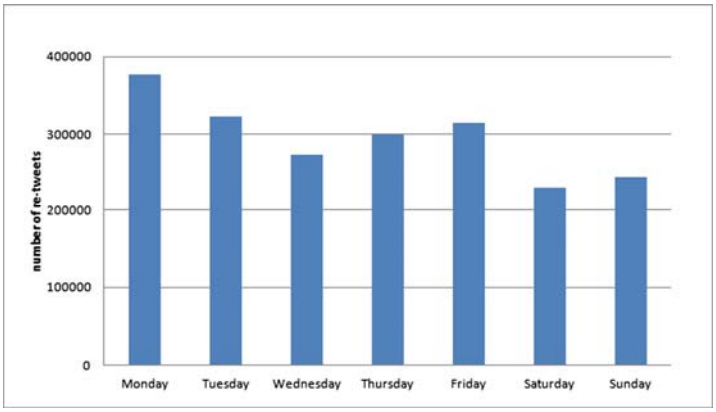

Fig. 3. Re-tweet activity by day of the week

problem is the study on the re-tweeting behaviours from an information diffusion perspective. Most works had focused on Twitter, a popular microblogging site. Insightful studies on re-tweeting behaviors can be seen from [1,2,4,9].

In [1], Boyd et al. studied the various aspects of re-tweeting. They conducted interviews with Twitter users and investigated the reasons why they re-tweet. Letierce et al. in [4] surveyed how researchers used Twitter to spread scientific messages. However, neither of them attempted to predict on whether a given message is to be re-tweeted. Galuba et al. in [2] focused on the URL propagation via re-tweets. In [9], Suh et al. gathered content and contextual features from Twitter and identified factors that impact re-tweeting. They found that URLs

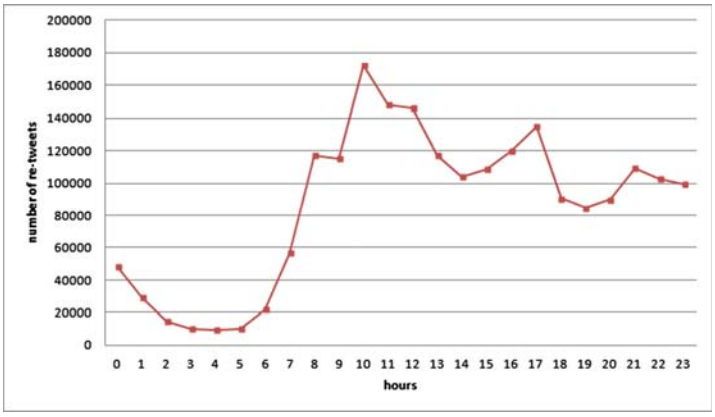

Fig. 4. Re-tweet activity by time of the day

and hashtags have strong relationships with re-tweetability and identified the number of followers and followees as important factors.

Zaman et al. in [12] adapted a probabilistic collaborative filtering model called Matchbox [8] to predict information spreading in Twitter based on features such as tweeter and re-tweeter information, and the tweet content. In [11], Yang et al. proposed a factor graph model based on users' re-tweeting history.

Recently, Petrovic et al. in [7] built a time-sensitive model based on the passive-aggressive algorithm (PA) to automatically predict re-tweets activities. Hong et al. in [3] trained a binary classifier to predict if a message will be re-tweeted or not and a multi-class classifier based on logistic regression to predict the volume of re-tweets for a given message. For the multi-class classification, they used four class labels (0: no re-tweet, 1: re-tweets less than 100, 2: re-tweets less than 10000, and 3: re-tweets more than 10000).

In [6], Peng et al. modelled the re-tweeting activities by using conditional random fields with three types of features, namely content influence, network influence and temporal decay factor. Naveed et al. in [5] argued that the tweet content is the key for re-tweeting prediction. They used logistic regression to compute re-tweet likelihood based on various interesting content features such as emotion positive/negative, exclamation/question mark, etc.

In our work, the tweet content has been removed from Sina Weibo microblog dataset pre-processed by WISE 2012 Challenge due to Sina Weibo's Term of Services.

3 Proposed Approach

3.1 Assumptions

Based on the given datasets, together with our statistical information presented in Sect. 1, we make the following assumptions:

- An event category is a group of similar events (manually grouped).
- The more popular the event category is, the more likely the tweet will be re-tweeted by a user.
- Similar events have similar re-tweet patterns.
- A user who has re-tweeted frequently in the past is likely to re-tweet in the future.
- Most users are only interested in tweets under certain event categories. Most followers are users who have similar interests.
- Users' interests and preferences are assumed to be stable.

3.2 Event Category

In WISE 2012 Challenge, the given original tweets are annotated with some social events together with their corresponding keyword lists. It is difficult to automatically group events into different categories and it is neither in our focus in this report because some events are simply labelled by personal names or by location names. Moreover, their relevant keyword lists are arbitrary and do not show clear contextual information between the keyword list and the event title. To solve this problem, we manually divide the WISE 2012 provided 46 events that have links to Wikipedia pages into 12 categories such as Natural Disaster, Celebrities, Product Release, Sports, and etc. The examples of event categories are shown in Table 5.

In order to predict the number of re-tweets, we adopt a probabilistic collaborative filtering prediction model called Matchbox which is a probabilistic model for generating personalized recommendations of items to users of a web service. Matchbox is used for the prediction of rating that users are likely to assign to items. It uses content information in the form of user and item metadata to learn correlations between them. Details of the Matchbox model can be found in [8]. This model can be applied to cope with our problem by the prediction of re-tweeting probability instead of the prediction of rating.

Matchbox is a factor graph for Bi-linear rating model. Each user and item are represented by a vector of features. Each feature is associated with a latent trait vector and the linear combination of the trait vectors for a particular user or item. An existing implementations of the Matchbox Recommender can be found at this link[2]. We adopt this model to predict whether followers of user will re-tweet the message posted by user who has posted an original tweet. For our approach, each tweet is regarded as an item while re-tweeter is considered as a user.

3.3 Tweet and Re-tweeter Features

According to datasets which have been pre-processed by WISE 2012 Challenge, we have Followship network and Tweets data without content. Although keyword lists are provided, they are arbitrary and do not show clear contextual

[2] https://docs.microsoft.com/en-us/azure/machine-learning/studio-module-reference/train-matchbox-recommender.

Algorithm 1. *PredictRetweetviaMatchbox*

Input: *mid*:message id
Output: *num_r*:predicted number of re-tweets

```
1  tweets = GetPrevious100Messages(mid); //Get the latest 100 messages of the
   user before the predicted message (mid) has been posted.
2  users = GetRetweeters(tweets);
3  retweets = GetRetweetHistory(tweets);
4  tw_vectors = CreateTweetFeatures(tweets);
5  usr_vectors = CreateUserFeatures(users);
6  model = TrainModel(tw_vectors, usr_vectors, retweets);
7  foreach u ∈ usr_vectors do
8  |   predict = model.predict(u, mid);
9  |   if predict.getProbTrue() ≥ threshold then
10 |   |   num_rt = num_rt + 1;
11 |   end
12 end
13 return num_rt;
```

information between the keywords and the event. For our approach, each tweet is regarded as an item while re-tweeter is considered as a user to train the model.

Tweet features consist of tweet id, user id who posted the original tweet, number of followers, number of followees, day of the week, time of the day and event category. Re-tweeter features include user id who re-tweeted the tweet, number of followers and number of followees. Re-tweeters are extracted from all users who have re-tweeted in the past of each tweet. The binary feedback is 1 if the re-tweeter re-tweeted the tweet within 30 days and 0 otherwise. The output of the model will be the probability of a re-tweet of the tweet by the re-tweeter.

3.4 Training Data

In order to train the model, it is required the positive binary feedback and also negative feedback. The positive feedbacks are from all re-tweet action in the past of each tweet in the same event category. For a given tweet, the negative feedbacks are from all followers in the re-tweet network who did not re-tweet a given tweet. For each test event, we train the model by random select 1,000 original tweets in the same event category as items and extract re-tweeters from re-tweet history of each tweet.

3.5 Prediction

To predict the number of re-tweets, for given original tweet and set of users if user has the high probability of a re-tweet greater than threshold, the user is likely to

re-tweet the original tweet. In order to find the most suitable value for threshold, we did the prediction on different threshold values. When threshold = 0.4 it render the best performance. The algorithm is shown as Algorithm 1.

4 Experiments and Evaluations

4.1 Baselines

The two baselines were compared with our results.

Baseline 1: Regression based on Popularity and Connectivity. It is a model to predict re-tweet activities based on event popularity and user connectivity by using a naïve approach. The intuition is that a tweet is more likely to be re-tweeted if it is about a popular event and its author is highly connected with others. The prediction will be the estimation of the probabilities of these two parameters in the space (connectivity of the user and category popularity). The formula for re-tweet prediction is shown as Eq. 1.

$$NumberOfRTs = 19.950(0.024C(uid) + 0.976P(uid, category)) \quad (1)$$

where function $C(uid)$ is to find how many re-tweets a uid (user ID) may have based on the number of followers she has, function $P(uid, category)$ is to predict how the event category popularity influences a tweet being re-tweeted. More details can be found in [10].

Baseline 2: Classification based on User Preferences. User preferences are used to train a classifier to predict the possible number of re-tweets in 30 days for a given original tweet. Given an original tweet, the authors need to compute how possible a user will re-tweet the original tweet in the category. The candidate users are extracted from re-tweet history in a form of "who-retweet-who". The authors use $P(r, u, c)$ to denote the interestingness of candidate re-tweet user r to original user u on category c. The function is defined as Eq. 2.

$$P(r, u, c) = \sum RT(r, u, c) / \sum T(u, c) \quad (2)$$

where $RT(r, u, c)$ returns the number of re-tweets by user r from user u on category c; $T(u, c)$ returns the total number of u's tweet on category c. More details of this algorithm can be found in [10].

Table 3. Average prediction error scores

Methods	Error scores
Baseline 1 : Regression based on Popularity and Connectivity	0.700
Baseline 2 : Classification based on User Preferences	0.666
Our approach : Probabilistic collaborative filtering prediction model	**0.627**

Table 4. The 33 predicted re-tweets of our approach and baselines

Mid	Ground truth	Baseline 1	Baseline 2	Our approach
Death of steve jobs				
8872263516485596	165	228	127	428
8872961090747701	3550	135	128	312
8872983825828431	154	184	137	128
8872990233170214	121	126	140	156
Fuzhou bombings				
2700059958269443492	798	476	152	185
2700117991448817596	242	93	132	303
2700176673306864228	686	223	140	624
2701374467440601577	384	418	222	449
2701431322360449433	1271	10	148	488
Japan earthquake				
51000180083282169	576	68	157	138
51000180083492814	187	46	142	169
51000180091104384	188	46	172	42
55000180091534860	2119	43	147	463
55000180527027036	1068	5	134	40
58000180083553705	699	30	740	114
Li Na win French open tennis				
2709258383303085289	620	3	260	281
2709864654666932643	13638	33	117	52
2709870697693881414	417	25	114	246
2709871713230486085	1383	53	132	232
2709893077170155796	163	33	130	403
Xiaomi release				
8896800636296312	1230	20	119	83
8896822338137478	114	95	257	101
8896858839607761	1681	23	136	555
8896889634186199	808	4	178	185
8896952812610010	249	12	129	154
Yao Jiaxin murder case				
2243526721410152330	700	232	160	141
2243578214587694822	129	142	142	159
510001856830842390	534	170	182	159
510001856834367317	121	39	298	152
510001904903643837	1001	946	143	128
510001908564754698	3474	9	616	106
5100019107401880	1126	609	170	187
550001906873838396	4900	31	184	164

Table 5. The 12 event categories in *WISE 2012* dataset

Category	Event
Natural disaster	Earthquake of Yunnan Yingjiang
	Japan earthquake
	Yushu earthquake
	Zhouqu landslide
Product release	iPhone 4s release
	Windows Phone release
	Motorola was acquisitions by Google
	Xiaomi release
Sports	Yao Ming retirement
	Spain Series A League
	Li Na win French Open in tennis
Famous people	The death of Muammar Gaddafi
	The death of Steve Jobs
	Family violence of Li Yang
	Tang Jun educatioin qualification fake
	The death of Kim Jongil
	The death of Osama Bin Laden
Social problem	Anshun incident
	China Petro chemical Co. Ltd.
	Foxconn worker falls to death
	Guo Meimei
	Incident of self-burning at Yancheng, Jangsu
	Shanghai government's urban management officers attack migrant workers in 2011
	Yao Jiaxin murder case
	Yihuang self-immolation incident
	The death of Wang Yue
	Case of running fast car in Heibei University
Public security	Bohai bay oil spill
	Foxconn bombing in Chengdu
	Fuzhou bombings
	Shanxi
Protests	Chaozhou riot
	Mass suicide at Nanchang Bridge
	Protests of Wukan
	Qianxi riot
	Zhili disobey tax official violent
Development projects	Line 10 of Shanghai-Metro pileup
	Shenzhou-8 launch successfully
	Tiangong-1 launch successfully
Economy	House prices
	Individual income tax threshold rise up to 3500
Human right	Qian Yunhui
	Deng Yujiao incident
Accident	Gansu school bus crash
	Wenzhou train collision
Crime	Chongqing gang trials

4.2 Evaluations

For evaluation our approach, we predicted 33 test tweets and the ground truth of 33 tweets are provided by WISE 2012 Challenge[3]. For each tweet we compute the prediction error score (PE).

$$PE_i = \frac{|A_i - P_i|}{A_i} \tag{3}$$

where A_i is the actual value for tweet i and P_i is the predict value for tweet i. For each approach, the average of prediction error scores is computed.

$$Average_j = \frac{\sum_{t=1}^{N} PE_t}{N} \tag{4}$$

where N is the number of test tweets. The small number is the better prediction result. Table 3 shows the performance of our approach against baselines. Table 4 lists the predictions for the given 33 original tweets over 6 given events. In Table 3, our approach shows a better performance than others on the prediction number of re-tweets.

5 Conclusions

In this paper, we proposed an approach to automatically predict the number of re-tweets over micro-blogs. Our contributions can be summarized as: (1) We proposed a solution to estimate the volume of re-tweets for understanding the behaviors of social networks. (2) We adopt probabilistic collaborative filtering prediction model named Matchbox by the prediction of re-tweeting probability instead of the prediction of rating. (3) We provide an evaluation for the effective re-tweet prediction on a real-world dataset. Our experiments show that the proposed approach can effectively predict the number of re-tweet over the baselines. In future work, we will retrospectively study the assumptions that we have made on the given datasets and develop a hybrid approach to integrate the proposed methods.

References

1. Boyd, D., Golder, S., Lotan, G.: Tweet, tweet, retweet: conversational aspects of retweeting on twitter. In: HICSS, pp. 1–10 (2010)
2. Galuba, W., Aberer, K., Chakraborty, D., Despotovic, Z., Kellerer, W.: Outtweeting the twitterers-predicting information cascades in microblogs. In: Proceedings of the 3rd Conference on Online Social Networks, p. 3 (2010)
3. Hong, L., Dan, O., Davison, B.D.: Predicting popular messages in twitter. In: WWW (Companion Volume), pp. 57–58 (2011)

[3] http://content.wuala.com/contents/imc_ecnu/wise_challenge/A4_T2GTruth.zip?dl=1.

4. Letierce, J., Passant, A., Decker, S., Breslin, J.G.: Understanding how twitter is used to spread scientific messages. In: ACM WebSci Conference 2010, pp. 1–8 (2010)
5. Naveed, N., Gottron, T., Kunegis, J., Alhadi, A.C.: Bad news travel fast: a content-based analysis of interestingness on twitter. In: ACM WebSci Conference, pp. 1–7, June 2011
6. Peng, H.-K., Zhu, J., Piao, D., Yan, R., Zhang, Y.: Retweet modeling using conditional random fields. In: ICDM Workshops, pp. 336–343 (2011)
7. Petrovic, S., Osborne, M., Lavrenko, V.: RT to Win! predicting message propagation in twitter. In: ICWSM (2011)
8. Stern, D.H., Herbrich, R., Graepel, T.: Matchbox: large scale online bayesian recommendations. In: WWW, pp. 111–120 (2009)
9. Suh, B., Hong, L., Pirolli, P., Chi, E.H.: Want to be retweeted? Large scale analytics on factors impacting retweet in twitter network. In: SocialCom/PASSAT, pp. 177–184 (2010)
10. Unankard, S., Chen, L., Li, P., Wang, S., Huang, Z., Sharaf, M.A., Li, X.: On the prediction of re-tweeting activities in social networks – a report on WISE 2012 challenge. In: Wang, X.S., Cruz, I., Delis, A., Huang, G. (eds.) WISE 2012. LNCS, vol. 7651, pp. 744–754. Springer, Heidelberg (2012). https://doi.org/10.1007/978-3-642-35063-4_61
11. Yang, Z., Guo, J., Cai, K., Tang, J., Li, J., Zhang, L., Su, Z.: Understanding retweeting behaviors in social networks. In: CIKM, pp. 1633–1636 (2010)
12. Zaman, T.R., Herbrich, R., Stern, D.H.: Predicting information spreading in twitter. In: Computational Social Science and the Wisdom of Crowds, vol. 55, pp. 1–4 (2010)