Activity Prediction: A Twitter-based Exploration

Wouter Weerkamp
ISLA, University of Amsterdam
w.weerkamp@uva.nl

Maarten de Rijke
ISLA, University of Amsterdam
derijke@uva.nl

ABSTRACT

Social media platforms allow users to share their messages with everyone else. In microblogs, e.g., Twitter, people mostly report on what they did, they talk about current activities, and mention things they plan to do in the near future. In this paper, we propose the task of activity prediction, that is, trying to establish a set of activities that are likely to become popular at a later time. We perform a small-scale initial experiment, in which we try to predict popular activities for the coming evening using Dutch Twitter data. Our experiment shows the feasibility and challenges of the task, with a simple method resulting in human-readable activities. This exploration also identifies several issues (e.g., temporal phrases and activity classification) that need to be addressed in future work.

Categories and Subject Descriptors

H.3 [**Information Storage and Retrieval**]: H.3.1 Content Analysis and Indexing

General Terms

Algorithms, Theory, Experimentation, Measurement

Keywords

Activity prediction, life mining, microblogs, Twitter

1. INTRODUCTION

Social media platforms like blogs, social networks, and microblogs allow users to share messages with everyone else. For several years now, the popularity of these platforms has been increasing and there seems to be no end to it yet. Different social media platforms are used for different messages: blogs seem to be more suitable for news paper like messages in which users discuss experiences or convey opinions; multimedia platforms (like YouTube and Flickr) allow users to post comments that discuss the contents of a multimedia object; microblogs, like Twitter or Facebook status updates, are used to post quick updates on people's lives.

We focus on microblogs and specifically on Twitter. On this platform, people mostly post messages about what they did, what they are currently doing, or about their plans for the (near) future. Table 1 shows a set of typical tweets mentioning previous or current "events" in users' lives. In this paper we focus on a particular time-aware information extraction task: we are not interested in current or past activities of people, but in their future plans. We propose the task of *activity prediction*, which revolves around trying to establish a set of activities that are likely to be popular at a later time. More specifically, given a future timeframe (e.g., tonight, tomorrow, next week) and a stream of (microblog) messages, try to determine what the most popular activities will be for a given future timeframe. To give a few examples of messages (in this case, tweets) talking about future plans, see Table 2.

TAIA'12, August 16, 2012, Portland.
ACM xxx-x-xxxxx-xxx-x/xx/xx.

Table 1: Tweets referring to past or current activities.

Frustrated with a few things this morning. Most of all, that I set the wrong alarm and slept through run time... GRR
I did not sleep at all last night. I must be that excited for Peter and Chris tonight.
Working on my day off....I wouldn't have it any other way!!!
Bird watching with momma
just finished watching Wrath Of The Titans! #GoodMovie :)

Table 2: Tweets referring to a future activity.

im gonna wrestle a midget tonight... uberexcited
I'm kinda nervous for my date tonight!! #NeedToManUp
exited to dance with the girls tonight:) #wewillrockyou
Excited for bodypump class tonight! :D #gym #motivated #excited
come on germany tonight.. like to see u in the final.... watching tv tonight at home... have the dogs.. kids out public viewing

We are witnessing the emergence of a new type of time-aware information extraction that is perhaps best characterized as "life mining": extracting useful knowledge from the combined digital trails left behind by people who live a considerable part of their life online. Activity prediction is a special case of life mining.

There are several reasons why activity prediction is an interesting task. From an end user point of view, being able to predict popular activities within a user's network of friends, allows a system to recommend activities to this user. Imagine not knowing what to do tomorrow on your free Saturday and a recommender system coming up with a set of popular activities from your "friends" (going to the beach, exercising, attending a music festival). Taking the viewpoint of social media monitors (such as police, intelligence services), activity prediction could be used to identify locations or events where many people will gather and that require additional

resources for, for example, crowd control. A final reason is to look at it from a marketing point of view. Based on predictions, marketeers could decide to do additional advertising at certain events or during particular tv shows. Similarly, event organizers and tv stations could adjust their ad pricing and communication plans based on predicted popularity of activities.

Previously, researchers have looked at a variety of prediction tasks that make use of social media (see the survey paper by Yu and Kak [18] for more examples besides those below). Tsagkias et al. [15] try to predict the impact of news articles using user-generated comments on news paper websites and online news platforms before these articles are published. In a comparable setting, Szabo and Huberman [14] predict the popularity of online content on YouTube and Digg after observing these items for several hours (or days). Tsagkias et al. [16] perform similar online prediction, but then for news article impact. A popular prediction research area is movies. Initial work focuses on predicting movie revenues by counting and analyzing tweets about these movies [1, 9]. Similar work has previously been done on predicting book sales using blogs [8]. Moving from revenue to appreciation in the movie domain, Oghina et al. [11] try to predict IMDb ratings of movies using Twitter messages about the movies and statistics from their trailers on Youtube (e.g., views, likes, and dislikes). Encouraged by outcomes of "easier" prediction tasks, researchers also focus on more challenging domains. Stock market prediction is a potentially extremely profitable task and has received a fair bit of attention in recent years. Most notably, prediction work by De Choudhury et al. [5] (using blogs) and Bollen et al. [3] (using Twitter) are commonly cited as successful attempts. Another domain that is currently popular among researchers is politics and more specifically, elections. Both Balasubramanyan et al. [2] and Tumasjan et al. [17] try to predict election outcomes based on political messages in Twitter. The latter two prediction tasks (stock market and elections) received a lot of attention, both in mainstream news and in academia, since they make strong claims about very important fields. Recently, papers (e.g., by Gayo-Avello et al. [6]) and blog posts[1] focus on the flaws in these papers and cast doubts about their validity.

With regard to social media and people's lives, Golder and Macy [7] use Twitter to monitor moods and activities during the day, while Ritter et al. [12] use the same platform to extract (future) events and create an open-domain calendar. The latter is related to our work, but focuses on large events several days in the future, whereas we focus on individual activities for the nearby future.

We perform an initial exploration of the activity prediction task based on Dutch messages from Twitter. We focus on the introduction of the task and assessing its feasibility by applying simple methods (Section 2) to it and observing the outcomes (Section 4). Based on the experiments and outcomes we suggest various directions for future work in the field of activity prediction (Section 5).

2. METHODS

We propose two naive approaches to predicting activities and future work should focus on developing more intelligent methods for this task. Both methods proposed here start by selecting a set of tweets that require analysis. In the next step we select key terms from this set of tweets, by comparing them to a background corpus, and we try to combine terms that belong together. Finally, we summarize the tweets belonging to one "topic" to allow for easy evaluation and interpretation.

Tweet selection. Tweet selection is necessary to create the set of tweets over which we calculate statistics and from which we extract activity terms. In this paper we use two ways of selecting tweets. First, we simply take all tweets up to a certain time that mention the future timeframe and use this set for our analysis. Second, we combine the timeframe selection with predetermined activity words to create a more focused set of tweets around both the future timeframe and an activity word.

[1] http://sellthenews.tumblr.com/post/21067996377/noitdoesnot

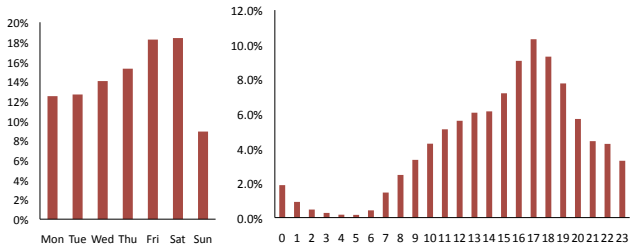

Figure 1: Percentage of tweets containing "vanavond" for each (Left): weekday and (Right): hour.

Term extraction and matching. To select terms from the sets of tweets we use the log-likelihood of a term, which compares the term frequency in the set of tweets to a background corpus and selects those terms that appear unexpectedly frequent. Our method selects 30 terms at 6pm and terms can only be selected if they occur in at least 40 tweets. To match terms that refer to the same activity, we use a naive co-occurrence metric, without any optimization.

Summarization. We use the method proposed by Sharifi et al. [13] for automatically creating summaries from a set of tweets related to the same topic. The method takes a set of tweets and a topic (term) as input and creates paths (sequences of terms) from the topic in both directions. The paths with the highest value (based on the number of tweets using this path) are selected to be the summary of this topic. We alter the method slightly, by introducing negative scores for stop words, to minimize the chance of starting or ending the summary with a stop word. If an activity consists of more than one term, we select the summary that belongs to the most common term as representation for that activity.

3. EXPERIMENTAL SETUP

For activity prediction we need to make one important decision: for which future timeframe are we predicting activities? In this paper we focus on one timeframe, tonight. To select the tweets referring to tonight, we issue a running query against the Twitter Streaming API using the terms "vanavond" (tonight) and "vnvnd" (2nite). Since these terms exist only in Dutch, there is no need for language identification. We have been collecting tweets since early August 2010 and have collected 7,076,021 tweets.

Besides the data set with tweets referring to tonight, we require a background corpus. To this end we use one year (2011) of Twitter Spritzer data and use language identification (based on work by Carter et al. [4]) to select only the Dutch tweets. This method gives us 8,317,184 (representative) Dutch tweets. For both sets of tweets we remove, on a per-day basis, duplicates with more than 50 characters, since these often relate to marketing stunts or news announcements. The numbers reported here are after de-duplication.

The plots in Figure 1 give an impression of the distribution of tweets referring to tonight. On the left we plot the weekdays and the percentage of all tweets referring to tonight that have been posted on each day. Similarly, we plot these numbers per hour of the day in the right plot. (We only use tweets posted before the evening (i.e., 6pm at latest) when predicting activities for that day.) A large number of tweets containing tonight is posted after 6pm, which is

Table 3: Judgments of extracted activities by two annotators for all summaries and for only proper summaries.

	All summaries	Proper summaries
Qualitative summary?		
Not at all	13–17	
Proper	39–43	
Activity yes/no?		
Yes	26–29	22–23
No	27–30	11–12
Popular activity?		
Yes	17–17	14–16
No	39–39	18–20

due to the ambiguity of the term tonight. Besides referring to a future timeframe, "tonight" can also refer to current time ("great weather tonight") and to the past ("I had a great time tonight").

To explore the feasibility of the activity prediction task, we perform a small-scale experiment. Given the setup described above, we select three days (June 3–5, 2012) for which we extract activities at 6pm. We then present the extracted activities (represented by their summaries) to two (Dutch) assessors, who are asked whether the suggested activity (i) is properly summarized, (ii) really is an activity and (iii) could be a popular activity for the evening following the prediction time. An activity is defined as *something you* (the assessor) *could actually do tonight, assuming you can move to any location instantaneously.*

To simplify evaluation, assessments are on a three-point scale. This suffices to explore the feasibility of the task. For assessment (i) we use a scale, ranging from *not at all*, via *somewhat*, to *perfectly*, and for (ii) and (iii) we use *no*, *yes*, and *unknown*.

For the selection of tweets we need an activity word that indicates a future activity. Here, we experiment with the word "kijken" (to watch), as it is one of the most popular verbs in tweets referring to tonight. In Section 5 we discuss the issue of activity classification and activity indicators further.

4. RESULTS

We first explore the inter-annotator agreement of our two annotators on the three assessment tasks. We observe that Cohen's kappa is fairly low for assessments on the summary quality ($\kappa = 0.25$) and activity popularity ($\kappa = 0.26$), whereas it is fair for whether or not the summary presents an activity ($\kappa = 0.36$). To facilitate the analysis of the results, we decide on recoding the assessments by merging *no* and *unknown* decisions into one (*no*), and to do the same for *somewhat* and *perfect* for summary quality (*proper*). After recoding the agreement rises substantially and good agreement is obtained for assessments on activity yes/no ($\kappa = 0.68$) and activity popularity ($\kappa = 0.67$), and fair agreement for summary quality ($\kappa = 0.37$). Results in Table 3 show the number of extracted activities in each class for both annotators. We present numbers over all summaries/activities and over only proper summaries (i.e., after removing summaries judged *not at all* by one or both annotators).

The results show us that summary quality is very important in this task. After selecting only proper summaries, the agreement on deciding whether something is really an activity rises to $\kappa = 0.80$. Besides that, about 66% of the proper summaries is judged to be an activity. Looking at all summaries, about 50% of those refer to an activity according to the annotators.

Finally we look at the difference between activities predicted using all tweets (21 activities) and those referring to the activity word "to watch" (13 activities). Focusing on only the proper summaries, we find that agreement is high for the latter set of activities (only one disagreement) and the percentage of summaries annotated as activity is also higher than for all tweets (62% vs. 70–75%). We also observe a large increase in the percentage of highly popular activities when using the "watching" filter: 60–70% are considered popular, whereas this is only 30% when using all tweets.

Table 4: Examples of extracted activities and their annotations (proper summary, activity or not, popular or not). Quoted text are names or Dutch tv shows or events.

	Summ.	Act.	Pop.
Sunday June 3			
I'm going to bed early tonight	y	y	y
the final episode of "peter r de vries"	y	y/n	n
tonight to "wtt"	n/y	y	n/y
I tonight pizza	n/y	n/y	n
watching a movie	y	y	y
Monday June 4			
first evening of the "avondvierdaagse"	y	n	-
the finals of "in love with sterretje"	y	n	-
the first episode of "vioranje"	y	n	-
tonight squad training	y	y	n
tonight to "guns n roses"	y	y	y
I have to watch "gtst" tonight	y	y	y
final episode of	n	n	-
Tuesday June 5			
tonight to the fun fair	y	y	n/y
to the "tros" music festival	y	y	y
tonight working out with	y	y	n/y
tonight at 8.30pm it'll be raffled off	n/y	n/y	n
tonight "gtst"	y	n	n
tonight once again watching at practice	y	y	n

Examples and observations. Table 4 shows several examples of extracted activities and their annotations. The summaries are literally translated, so mistakes can easily be identified.

From the examples and the annotations we observe the following. Predicted activities should contain either a verb or "to" (as in: going to), as the examples clearly show that various perfect summaries ("the first episode of..," "tonight gtst[2]," "the finals of..") are activities, but lack the proper verbs (to watch). In case the activity contains "to," it is apparently not necessary to also add a verb (e.g., "going"). One of the annotators also mentioned that "a sentence like 'tonight gtst' is not really a sentence, let alone something you can do, even though 'watching gtst tonight' would be."

Other observations include the facts that (i) spam (or marketing stunts) can be mistaken for activities ("tonight at 8.30pm it'll be raffled off"), (ii) properly summarizing the activities is very important, but hard due to a relatively small number of tweets, and (iii) popularity is hard to estimate and evaluate.

5. DISCUSSION

The small-scale experiments in this paper show that predicting popular activities for a later moment based on people's tweets is feasible, but challenging. Our naive, mainly heuristics-based approach manages to extract likely activities for people to participate in later that evening, but we also stumble upon various issues in exploring this task. Below, we discuss seven issues and their potential solutions, leading to future research directions related to this task.

[2]A Dutch soap opera, "Goede Tijden Slechte Tijden"

Activity classification. The main issue we need to address is the classification of tweets into those mentioning an activity and those that do not. To this end we could, for example, explore the usage of activity indicators (words) like "to watch," "to," or other verbs.

Summarization. Summarization is very important in presenting activity predictions to users. Although previously proposed methods work reasonably well, they need to be able to deal with a limited number of tweets. On top of that, the summaries need verbs to make it clear they present activities.

Term extraction. Currently, the method we use to extract unexpected activities depends on log-likelihood type methods. However, this way we only extract unexpected activities and we ignore very frequent, recurring activities. Future work should also look at other ways of extracting terms (and activities).

Time indication. People use various ways to refer to the same time of day. For now, we focused on filtering tweets that contained the words tonight ("vanavond") or 2nite ("vnvnd"), however, there are other ways to refer to tonight which might occur frequently. In the future we want to identify terms that refer to tonight, preferably in an automated way. Llorens et al. [10] have looked into using semantic knowledge to extract temporal phrases and events, work which could be very relevant for activity prediction.

Tweet segmentation. Tweets are short, but it is rare for the whole tweet to refer to the future. People often use templates like "first have to go to work, tonight party," or "this morning school, but tonight going to the movies." Simple methods identify "work" or "school" as activities related to tonight, whereas they clearly are not referring to future activities. Determining the scope of future references is necessary to further improve activity prediction.

Evaluation. Evaluation of the activity prediction task is hard. It needs to be done at three levels: (i) is the extracted activity really an activity, (ii) is the activity "suitable" for the given future timeframe, and (iii) is it a popular activity? Second, besides difficulties with measuring precision, it is hard to measure recall, i.e., did a system extract *all* popular activities for the given timeframe? Finally, even with just two annotators, there is a fair amount of disagreement. We need to look into ways of doing (semi) automatic evaluation, e.g., (i) by looking into other, more numeric data streams (e.g., viewing or visitor statistics, news coverage) or (ii) by using after-the-facts tweets to extract activities that were popular during the evening.

Combining multiple sources. The work in this paper only uses Twitter as a data source, but future work should focus on combining signals from multiple sources into one prediction. Other relevant sources are, for example, (a) Facebook event pages ("attending"); (b) visitor statistics and "likes" for event URLs; (c) shared calendars (e.g., Google Calendar) and meet-up websites (e.g., meetup.com).

6. CONCLUSIONS

In this paper we introduced a particular instance of life mining: activity prediction. The task is, given a set of tweets and a future timeframe, to extract a set of activities that will be popular during that timeframe. We showed, using a small-scale experiment, that predicting popular activities for the coming evening based on Dutch tweets is feasible, but challenging. Based on our observations we identified a set of seven directions for future work, which will further our understanding of activity prediction as a particular case of life mining.

Acknowledgments. This research was supported by the European Union's ICT Policy Support Programme as part of the Competitiveness and Innovation Framework Programme, CIP ICT-PSP under grant agreement nr 250430, the European Community's Seventh Framework Programme (FP7/2007-2013) under grant agreements nr 258191 (PROMISE) and 288024 (LiMoSINe), the Netherlands Organisation for Scientific Research (NWO) under project nrs 612.061.814, 612.061.815, 640.004.802, 380-70-011, 727.011.-005, 612.001.116, the Center for Creation, Content and Technology (CCCT), the Hyperlocal Service Platform project funded by the Service Innovation & ICT program, the WAHSP and BILAND projects funded by the CLARIN-nl program, the Dutch national program COMMIT, and by the ESF Research Network Program ELIAS.

References

[1] S. Asur and B. Huberman. Predicting the Future with Social Media. In *WI-IAT 2010*, pages 492–499, 2010.

[2] B. Balasubramanyan, B. Routledge, and N. Smith. From Tweets to Polls: Linking Text Sentiment to Public Opinion Time Series. In *ICWSM 2010*, pages 122–129, 2010.

[3] J. Bollen, H. Mao, and X. Zeng. Twitter mood predicts the stock market. *J. of Computational Science*, 2(1):1–8, 2011.

[4] S. Carter, W. Weerkamp, and E. Tsagkias. Microblog language identification: Overcoming the limitations of short, unedited and idiomatic text. *Language Resources and Evaluation Journal*, 2012.

[5] M. De Choudhury, H. Sundaram, A. John, and D. D. Seligmann. Can Blog Communication Dynamics be Correlated with Stock Market Activity? In *HT 2008*, pages 55–60, 2008.

[6] D. Gayo-Avello, P. Metaxas, and E. Mustafaraj. Limits of Electoral Predictions using Twitter. In *ICWSM 2011*, pages 165–171, 2011.

[7] S. A. Golder and M. W. Macy. Diurnal and seasonal mood vary with work, sleep, and daylength across diverse cultures. *Science*, 333(6051):1878–1881, 2011.

[8] D. Gruhl, R. Guha, R. Kumar, J. Novak, and A. Tomkins. The Predictive Power of Online Chatter. In *SIGKDD 2005*, pages 78–87, 2005.

[9] M. Joshi, D. Das, K. Gimpel, and N. A. Smith. Movie Reviews and Revenues: An Experiment in Text Regression. In *HLT-NAACL 2010*, pages 293–296, 2010.

[10] H. Llorens, E. Saquete, and B. Navarro-Colorado. Applying semantic knowledge to the automatic processing of temporal expressions and events in natural language. *Information Processing & Management*, Available online, 2012.

[11] A. Oghina, M. Breuss, E. Tsagkias, and M. de Rijke. Predicting IMDB Movie Ratings Using Social Media. In *ECIR 2012*, pages 503–507, 2012.

[12] A. Ritter, Mausam, O. Etzioni, and S. Clark. Open domain event extraction from twitter. In *KDD 2012*, 2012.

[13] B. Sharifi, M.-A. Hutton, and J. Kalita. Summarizing microblogs automatically. In *HLT-NAACL 2010*, pages 685–688, 2010.

[14] G. Szabo and B. A. Huberman. Predicting the Popularity of Online Content. *Comm. of the ACM*, 53(8):80–88, 2010.

[15] M. Tsagkias, M. de Rijke, and W. Weerkamp. Predicting the Volume of Comments on Online News Stories. In *CIKM 2009*, pages 1765–1768, 2009.

[16] M. Tsagkias, W. Weerkamp, and M. de Rijke. News Comments: Exploring, Modeling, and Online Prediction. In *ECIR 2010*, pages 191–203, 2010.

[17] A. Tumasjan, T. Sprenger, P. Sandner, and I. Welpe. Predicting Elections with Twitter: What 140 Characters Reveal about Political Sentiment. In *ICWSM 2010*, pages 178–185, 2010.

[18] S. Yu and S. Kak. A Survey of Prediction Using Social Media, 2012. http://arxiv.org/abs/1203.1647.