Identifying Interesting Postings on Social Media Sites

Swathi Seethakkagari
School for Electronics & Computer Systems
University of Cincinnati
Cincinnati, OH 45221-0030
seethasi@mail.uc.edu

Anca Ralescu
Machine Learning & Computational Intelligence Lab
School of Computing Sciences & Informatics
University of Cincinnati
Cincinnati, Oh 45221-0030
anca.ralescu@uc.edu

Abstract

This paper considers the classification of messages posted on social networking sites as a step towards identifying interesting/non-interesting messages. As a first approximation, a message is represented by two attributes – the *message length* (number of words), *posting frequency* (time difference between consecutive messages) for the same sender. A classifier, trained according to a user's perception of whether a message is interesting or not, is used to label each message. *Facebook* is considered for illustration purposes.

Keywords

Social networks, classification, k-nearest neighbors.

Introduction

Social networking has long been an activity within social communities. Whether through relatives, friends, or acquintances people are routinely using their social connections to further their careers, and improve and enjoy their lives. With the advent of online social networks and sites this type of activity has increased, making possible networking on a large scale between people at great physical distances. Friendship and contacts can now be maintained over longer period of time, idea can be exchanged between massive groups of people. Social networks have become an excellent communication source.

Analysis of the networking sites has led to many interesting research issues, in a field that is rapidly growing of social computing and cultural modeling. A natural, and often used way to represent a network is through graphs, in which a vertex corresponds to an entity in the network, usually an individual, and an edge connecting two vertices represents some form of relationship between the corresponding individuals (Al Hasan et al. 2006). *"Social network analysis provides a significant perspective on a range of social computing applications. The structure of networks arising in such applications offers insights into patterns of interactions, and reveals global phenomena at scales that may be hard to identify when looking at a finer-grained resolution"* (Leskovec, Huttenlocher, and Kleinberg 2010).

Predicting the network evolution in time is central to such studies. In particular, detecting communities in a network, predicting the links between nodes in the network, have become much studied subjects in social computing, and other domains based on network representations. *"Prediction can be used to recommend new relationships such as friends in a social network or to uncover previously unknown links such as regulatory interactions among genes"* (Tan, Chen, and Esfahanian 2008).

By contrast with studies to reveal "global phenomena" one can consider the local, self-centric social network to which an individual belongs. The ability of anytime anywhere communication that online social networks provides to users has lead to an explosion of user generated data. Therefore, extracting patterns, global or local, from social networks, is necessary if we are to make sense of what a social network conveys about its users, and society at large. In this paper we consider the setting of a social network (such as Facebook, for example) where each user is free to post various messages (in Facebook this is done via the user *status* which the user updates. Some users are inclined to post frequent and often relatively uninteresting updates, others post them more rarely and their contents are more interesting. Then again, the evaluation "interesting/uninteresting" is subjective and varies from user to user. Therefore, to support a user whose community (friends) is large, filtering or classification of postings which can take into account this user's preference is necessary.

In the remainder of this paper we explore the idea of classifying postings/updates in a user's centered community based on the user's perception of their contents as interesting or not.

Analysis of messages on the social network

In the setting of a Facebook-like network, let I denote a generic individual, and $\mathcal{F}(I)$ the collection of friends (direct or indirect) of I. A snapshot of such Facebook community is illustrated in Figure 1.

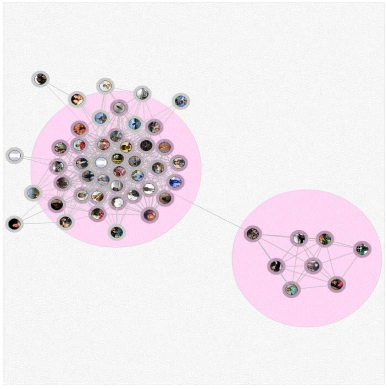

Figure 1: Snapshot from a Facebook friends network showing a group of clusters.

For each friend $f \in \mathcal{F}(I)$, $u_I(f)$ is an update of posted by f. The extent to which an update is interesting is, of course, a matter of its contents. This means that in principle, a text analysis of its contents should be done. However, while this would not doubt provide a deeper understanding of the actual content, other characteristics, such as message length, and frequency of messages from a particular f might give a close enough idea of how interesting a message is. For the purpose of this paper then, $u_I(f; n, t)$ denotes the update of length n (words), posted by the friend f at time interval t. The attributes, n and t are used to classify the update $u_I(f; n, t)$ as interesting or not.

k-Nearest Neighbors Classification of Postings

The well known k-nearest neighbor classifier (Hart 1967) is used to classify a newly posted message. The classification rule used by the k-nearest algorithm is very simple: using a set of labeled examples, a new example is classified according to its k-nearest neighbors, where k is a parameter of the algorithm. The k nearest neighbors "vote" each for the class with it has been labeled. Variations of the algorithm make possible to weigh a vote by the actual distance from each of these neighbors to the new example(Al Hasan et al. 2006): the vote of a closer neighbor counts more than that of a neighbor farther away. In the small experiment described below the simpler version of the algorithm is used. Algorithm 1 describes the steps for this classification.

Algorithm 1 Pseudo code for labeling a message using the k-NN Classifier

Require: 2-class training data set of of size n
Require: Test data point and k (odd values to avoid ties)
Require: classification technique: k Nearest Neighbors Classifier
Ensure: The test data point is labeled with its class based on classifier output. Labels are set to -1 or +1.
  Compute the classifier based on the distance of a test datum to the k nearest neighbors.
  for $i = 1, \ldots, n$ **do**
    Calculate the distance, $dist(i)$ with the ith data point in the training set
    Sort the distances
    Extract the k nearest neighbors
    if the sum of the top k labels is positive **then**
      label of test data point is set to 1;
    else
      label of test data point is set to -1;
    end if
  end for

A small real example

Table 1 shows a small set of updates posted on one of authors (S. Seethakkagari) Facebook page. The labels, $\pm$, are assigned according to her subjective evaluation of each posting.

Table 1: A small set of postings extracted from S. Seethakkagari's wall on facebook. N is the message length, T, the time from the last message of the same sender. The Label is assigned according her subjective evaluation of the posting content.

ID	1	2	3	4	5	6	7	8	9
N	23	16	26	20	30	22	32	16	12
T	272	81	149	287	10	26	4	1	36
Label	1	-1	1	1	1	1	-1	1	1

ID	10	11	12	13	14	15	16	17	18
N	6	4	7	1	32	4	17	15	3
T	0	64	39	558	199	72	52	32	216
Label	-1	-1	-1	1	-1	1	1	1	1

ID	19	20	21	22	23	24	25
N	61	2	38	13	2	6	23
T	27	594	63	0	80	39	57
Label	-1	1	1	-1	-1	1	1

Experimental Results

The data of labeled postings, shown in Table 1 are plotted in the $length \times time$ space as shown in Figure 2.

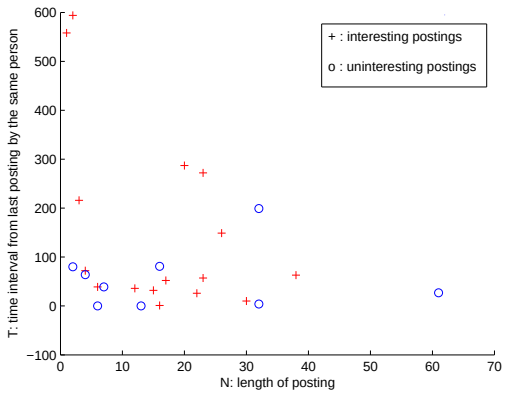

Figure 2: Plot of the 25 postings from S. Seethakkagari's site.

The following experiment was carried out: all the possible test data sets of four postings were generated for a total of 12650 sets. The corresponding training sets were obtained by eliminating the test data sets from the set of postings. The number of neighbors was selected to be $k = 3$. Table 2 and Figure 3 show the classification results of all test postings.

Table 2: Results of classification for all postings of four messages with respect to 21 postings used as training data.

accuracy(%)	0	25	50	75	100
frequency	252	1865	4544	4455	1534
average	60.1858				
mode	50				
median	50				

Conclusion and future work

We explored the use of classification of postings on a social media site into two classes: interesting versus non-interesting. Each message was encoded using two attributes,length, expressed as the N the number of words in the message and the frequency with which a its sender posts messages. Experiments were run for the set shown in Table 1, a small, but *real* data set, using a k-nearest neighbor classifier, with $k = 3$. We consider the results encouraging, as the probability of classification accuracy greater than or equal to 50% is over 83%. As a future study, a larger attribute set may be used. For example, the comments (their length and/or contents) received for previous message, the ID of a message sender, can be considered. However, the tradeoff between classification accuracy and computational efficiency. For example, as the number of attributes, or the number of neighbors k increase, the complexity in calculation increases. Feedback from the user may be used to adapt

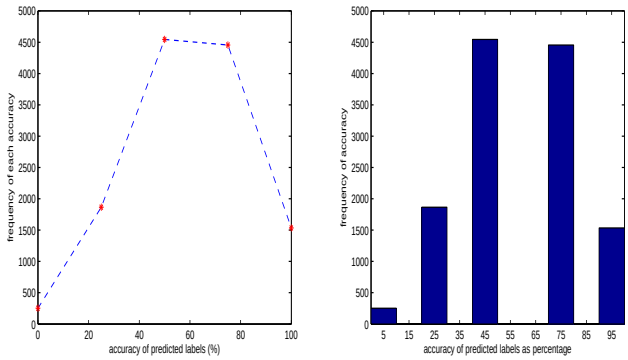

Figure 3: Accuracy of prediction when 21 messages are used to predict labels of a subset of four messages.

the classifier so as to achieve a better tradeoff between speed and accuracy.

References

Al Hasan, M.; Chaoji, V.; Salem, S.; and Zaki, M. 2006. Link prediction using supervised learning. In *SDM06: Workshop on Link Analysis, Counter-terrorism and Security*.

Hart, P. 1967. Nearest neighbor pattern classification. *IEEE Transactions on Information Theory* 13(1):21–27.

Leskovec, J.; Huttenlocher, D.; and Kleinberg, J. 2010. Signed networks in social media. In *Proceedings of the 28th international conference on Human factors in computing systems*, 1361–1370. ACM.

Tan, P.; Chen, F.; and Esfahanian, A. 2008. A Matrix Alignment Approach for Link Prediction. In *Proceedings of ICPR 2008*.