User Interest Modeling in Twitter with Named Entity Recognition

Deniz Karatay
METU Computer Engineering Dept.
06800 Ankara, Turkey
deniz.karatay@ceng.metu.edu.tr

Pinar Karagoz
METU Computer Engineering Dept.
06800 Ankara, Turkey
karagoz@ceng.metu.edu.tr

ABSTRACT

Considering wide use of Twitter as the source of information, reaching an interesting tweet for a user among a bunch of tweets is challenging. In this work we propose a Named Entity Recognition (NER) based user profile modeling for Twitter users and employ this model to generate personalized tweet recommendations. Effectiveness of the proposed method is shown through a set of experiments.

Categories and Subject Descriptors

H.4 [**Information Systems Applications**]: Miscellaneous

General Terms

Theory

Keywords

Named Entity Recognition, Tweet Segmentation, Tweet Classification, Tweet Ranking, Tweet Recommendation

1. INTRODUCTION

As a service that embodies both social networking and microblogging, Twitter has become one of the most important communication channels with its ability of providing the most up-to-date and newsworthy information [6]. In this study, we present a technique for constructing user interest model, in which user interests are defined by means of relationship between the user and his friends as well as named entities extracted from tweets. We demonstrate the use of this model for tweet recommendation.

To extract information from this large volume of tweets generated by Twitter's millions of users, Named Entity Recognition (NER), which is the focus of this work, is already being used by researchers. NER can be basically defined as identifying and categorizing certain type of data (i.e. person, location, organization names, date-time and numeric expressions) in a certain type of text. On the other hand, tweets are characteristically short and noisy. Considering the fact that tweets generally include grammar mistakes, misspellings, and informal capitalization, performance of the traditional methods is incompetent on tweets and new approaches have to be generated to deal with this type of data. Recently, tweet representation based on segments in order to extract named entities has proven its validity in NER field [4, 3].

In this work, it is aimed to reduce the Twitter user's effort to access to the tweet carrying the information of interest. To this aim, a tweet recommendation method under a user interest model generated via named entities is presented. To achieve our goal, a graph based user interest model is generated via named entities extracted from user's followees' and user's own posts. In the user interest model, each included followee is ranked based on their interactions with the user via retweets and mentions, and named entities are scored via ranking of the user posting them.

2. PROPOSED METHOD

The general overview of the system architecture can also be seen in Figure 1. The method used in this study segments the tweets and generates named entity candidates. These candidates have to be validated so that they can be used as an indicator of the user's interest. In this step, Wikipedia is chosen as a reference for a segment to be a named entity, or not. Since our Tweet collection is in Turkish, Turkish Wikipedia dump published by Wikipedia is obtained.

For named entities to be extracted successfully, the informal writing style in tweets has to be handled. Generally named entities are assumed as words written in uppercase or mixed case phrases where uppercased letters are at the beginning and ending, and almost all of the studies bases on this assumption. However, capitalization is not a strong indicator in tweet-like informal texts, sometimes even misleading. To extract named entities in tweets, the effect of the informality of the tweets has to be minimized as possible. The preprocessing tasks applied can be divided into two logical group:. Pre-segmenting, and Correcting. Removal of links, hashtags, mentions, conjunctives, stop words, vocatives, slang words and elimination of punctuation are considered as presegmentation. It is assumed that parts in the texts before and after a redundant word, or a punctuation mark cannot form a named entity together, therefore every removal of a word is considered as it segments the tweet as well as punctuation does it naturally. Removal of repeating characters that are used to express a feeling such as exaggerating,

Published as part of the #Microposts2015 Workshop proceedings, available online as CEUR Vol-1395 (http://ceur-ws.org/Vol-1395)

#Microposts2015, May 18th, 2015, Florence, Italy.

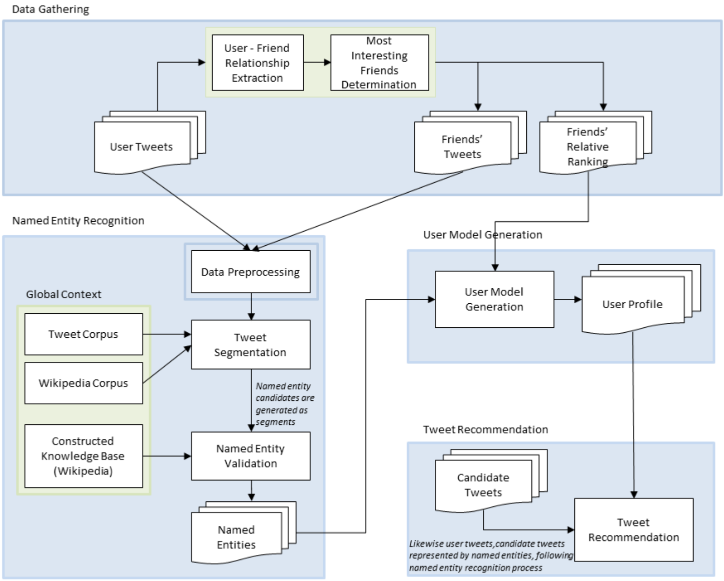
Data Gathering
User - Friend Relationship Extraction
Most Interesting Friends Determination
User Tweets
Friends' Tweets
Friends' Relative Ranking
Named Entity Recognition
Data Preprocessing
Global Context
Tweet Corpus
Wikipedia Corpus
Constructed Knowledge Base (Wikipedia)
Tweet Segmentation
Named entity candidates are generated as segments
Named Entity Validation
Named Entities
User Model Generation
User Model Generation
User Profile
Tweet Recommendation
Candidate Tweets
Tweet Recommendation
Likewise user tweets,candidate tweets represented by named entities, following named entity recognition process

Figure 1: System Architecture

or yelling, handling mistyping and asciification related problems are considered as correcting and can be thought of conversion of tweets from informal to formal. In the following subsections, we describe the NER and user profile modeling and recommendation steps in more detail.

2.1 Finding Named Entities

In this study, the idea of segmenting a tweet text into a set of phrases, each of which appears more than random occurence [1, 4] is adopted. Therefore, a corpus serving this purpose in Turkish is needed. To this aim, *TS Corpus*, which indexes Wikipedia articles and also Tweets [5], is used. In the proposed solution, *TS Corpus* is used for gathering statistical information for various segmentation combinations by means of a dynamic programming algorithm. While collecting statistical information for segment combinations, tweet collection of *TS Corpus* is also used while computing probability of a segment to be a valid named entity, which is different from the previous studies. The knowledge base that is constructed using Turkish *Wikipedia* dump is used to validate the candidate named entities.

Segmentation constitutes the core part of named entity recognition method. The aim here is to split a tweet into consecutive segments. Each segment contains at least one word. For the optimal segmentation, the following objective function is used, where F is the *stickiness* function, t is an individual tweet, and s represents a segment.

$$\arg\max_{s_1 \ldots s_n} F(t) = \sum_{i=1}^{n} F(s_i) \tag{1}$$

Although the term *stickiness* is generally used for expressing tendency of a user to stay longer on a web page by a user, Li et. al defined it as the metric of a word group to be seen together in documents frequently, or not [4] and it is used in the same way in this study. The *stickiness* function basically measures the *stickiness* of a segment or a tweet represented based on word collocations. A low *stickiness* value of a segment means that words are not used commonly together and can be further split to obtain a more suitable word collocation. On the other hand, a high *stickiness* value of a segment indicates that words in the segment are used together often and represent a word collocation, therefore cannot be further split. In order to determine the correct segmentation, the objective function above is used, where a tweet representation with the maximum *stickiness* is chosen to be the correct segmentation. Instead of generating all possible segmentations and compute their stickiness, dynamic programming algorithm described in [4] is adapted to this study to compute stickiness values efficiently. The algorithm basically segments the longer segment, which can be tweet itself, into two segments and evaluates the *stickiness* of the resultant segments recursively. More formally, given any segment $s = w_1 w_2 ... w_n$, adjacent binary segmentations $s_1 = w_1 ... w_j$ and $s_2 = w_j + 1 ... w_n$ is obtained by satisfying the following equation.

$$\arg\max_{s_1, s_2} F(s) = F(s_1) + F(s_2) \tag{2}$$

Thus far, tweets are segmented making use of the stickiness function. In the result of this phase, tweet segments, which are candidate named entities, are obtained. These candidate named entities have to be validated whether they are real named entities or not, so that they can be used as an indicator of the user's interest. For this purpose, as explained before, Wikipedia is chosen as a reference for a segment to be a named entity, and a graph-based knowledge-base based on Wikipedia is constructed. If the segment, which is actually a candidate named entity, matches exactly with a Wikipedia title in the constructed knowledge base, then it is accepted to be a named entity. In case of inexact match, we use the

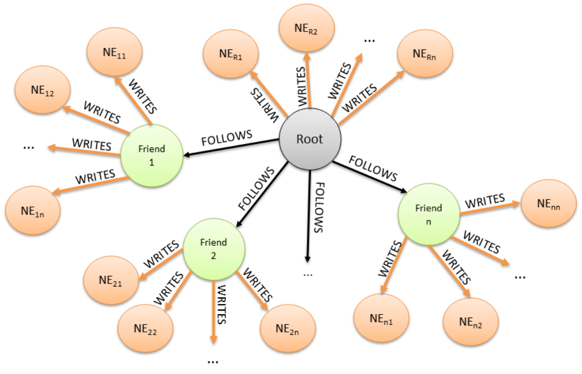

Figure 2: Structure of the User Interest Model Graph

Levenshtein distance [2] to measure the similarity of a segment to a Wikipedia title.

2.2 Generating User Interest Model based on Named Entities

At this step, named entities with their frequency counts in a tweet obtained from followees' posts, and followees' relative ranking obtained in data gathering phase is processed as shown in Figure 1. Using these data, a user interest model is generated. It is basically a graph based relationship model. Let $G = (V, E)$ be a weighted labelled graph with the node set V and edge set E. Node set V is labelled with the label set L_1 where $L_1 \in \{Root, Followee, NamedEntity\}$ and Edge set E is labelled with the label set L_2 where $L_2 \in \{Follows, Writes\}$. In other words, a user interest model graph has three types of nodes; *Root, Friend, Named Entity*, along with two types of weighted edges; *Writes, and Follows*. Weight of *Writes* edge represents the appearance count of a named entity for a followere's posts where weight of the *Follows* edge represents relative ranking of a followed. Therefore, a twitter profile is represented as *Root* node *Follows* one or many *Followee*s, and a *Followee* node *Writes* one or many *Named Entities*. The structure of the graph is shown in Figure 2.

2.3 Tweet Recommendation

Determining whether a tweet is interesting or not is achieved by comparing NE representation of the tweet with the generated user interest model. This comparison results in a ranking of candidate tweets. As the first step, candidate tweets are processed to obtain their NE representations. NE representation of a tweet simply includes the NEs, and their frequency counts. In order to compare with the candidate tweet, user interest model has to be interpreted by including the ranking score factor of the friends. Every followee's named entities and their appearance counts are first multiplied with the friend's ranking, and then summed. Therefore, a set of named entities with their scores based on the user interest model is obtained. The mathematical interpretation to calculate the score of a single named entity is given in Equation 3, where SC_{NE} represents the overall score of a named entity, C represents the frequency count of a named entity for a user, n represents the count of friends included in the user interest model, RR represents the relative ranking score of a followed, and U represents the user himself. With the same approach, the final score of all of the named entities appearing in the user interest model is calculated.

$$SC_{NE} = \sum_{i=1}^{n} RR_i \cdot C_i + RR_U \cdot C_U \qquad (3)$$

After overall score is calculated for all of the named entities in the user interest model, final scores for candidate tweets are calculated in the following approach: Overall score of named entities in NE representation of a candidate tweet are multiplied with the frequency count in the NE representation of itself. This operation is done for every named entity in the tweet representation, and then by summing these values, final score of a candidate tweet is obtained. If a named entity in a candidate tweet's NE representation, does not appear in the user interest model, its overall score is accepted as 0 and not taken into consideration assuming the user is not interested in the subject that particular named entity represents. Once final scores for all candidate tweets are calculated, candidate tweets are sorted in descending order, and hence, they are ranked.

$$SC_T = \sum_{i=1}^{m} SC_{NE_i} \cdot C_{NE_i} \qquad (4)$$

3. EXPERIMENTAL RESULTS

To evaluate the system from recommendation point of view, two types of datasets as candidate tweets for recommendation and two types of user groups to recommend tweets are formed. The first dataset of candidate tweets, *GNRL*, is a general dataset containing 100 tweets crawled from newspapers' Twitter accounts. The second dataset, *PSNL* is a personal dataset containing 100 tweets that are crawled from the followees of followees of the selected users. There are 10 users volunteered for this experiment where half of them are active Twitter users, whereas the other half are inactive Twitter users. *Active Users* are the users that use Twitter frequently, have retweeting and mentioning habits, and update followed list when necessary where *Inactive Users* do not post, retweet, or mention often, and do not update followee list frequently. Volunteered users are categorized on the basis of the information they provided about their Twitter usage habits.

For each user, user interest model is constructed under SCP measure on Wikipedia Corpus along with length normalization for stickiness function, which gives the best results according to the validation experiments. In addition, the best N_T and N_F values are experimentally obtained, therefore 20 followees and 10 tweets of each followed are included in the model. Candidate tweets are scored by comparing with user's model as explained in Section 2.3 and then ranked. Meanwhile, each user is asked to classify and score tweets in *GNRL* and *PSNL* datasets. Volunteered users made a two-step evaluation on each tweet for each dataset. They are asked to mark the tweet as interesting or uninteresting, and then if the tweet is interesting, they are asked to score the tweet in the range of $[1-3]$ where 1 is the least score, and 3 is the highest score for interestingness. In the

		Classification Acc. (%)		Ranking Acc. (nDCG)	
		$GNRL$	$PSNL$	$GNRL$	$PSNL$
Inactive Users	$User_1$	47	49	0.520	0.612
	$User_2$	42	39	0.573	0.654
	$User_3$	36	37	0.433	0.478
	$User_4$	43	36	0.322	0.301
	$User_5$	49	47	0.567	0.514
Average (IU)		**43.40**	**41.60**	**0.483**	**0.512**
Active Users	$User_6$	68	64	0.777	0.909
	$User_7$	66	61	0.699	0.768
	$User_8$	62	56	0.760	0.782
	$User_9$	71	72	0.720	0.815
	$User_{10}$	72	65	0.601	0.677
Average (AU)		**67.80**	**63.60**	**0.711**	**0.790**
Average (Overall)		**54.10**		**0.624**	

Table 1: Tweet Recommendation Experiment Results with respect to the Baseline Method

		Classification Acc. (%)		Ranking Acc. (nDCG)	
		$GNRL$	$PSNL$	$GNRL$	$PSNL$
Inactive Users	$User_1$	69	66	0.723	0.773
	$User_2$	62	58	0.684	0.796
	$User_3$	52	55	0.656	0.616
	$User_4$	67	52	0.590	0.623
	$User_5$	72	69	0.734	0.691
Average (IU)		**64.40**	**60.00**	**0.677**	**0.700**
Active Users	$User_6$	88	86	0.809	0.958
	$User_7$	79	74	0.795	0.888
	$User_8$	74	68	0.812	0.826
	$User_9$	88	85	0.815	0.904
	$User_{10}$	80	77	0.773	0.872
Average (AU)		**81.80**	**78**	**0.801**	**0.890**
Average (Overall)		**71.05**		**0.767**	

Table 2: Tweet Recommendation Experiment Results with Respect to the Proposed Method

baseline method, followee rankings are neglected and hence every named entity has equal weight. Generated recommendations are compared against the user preferences in terms of classification, and ranking.

The results in Table 1 show that the baseline method is able to decide whether a tweet is interesting for a user or not with the accuracy of 54,10% on average with classification and 0,624 *nDCG* value on average with ranking, which are lower than the results of our system. The performance of the baseline method in some cases decreases down to 36% correct prediction at classification, and 0,322 *nDCG* value at ranking quality. On the other hand, the results shown in Table 2 shows that the proposed system is able to decide whether a tweet is interesting for a user or not with the accuracy of 71,05% on average for classification and 0,767 *nDCG* value on average for ranking. Given the suitable user habits, performance of the system increases up to the 88% correct prediction for classification, and 0,958 *nDCG* value at ranking quality. The comparison of two tables show that the proposed user interest modeling approach increases the performance.

4. CONCLUSIONS

This paper proposes a new approach to Twitter user modeling and tweet recommendation by making use of named entities extracted from tweets. A powerful aspect of NER approach adopted in this study, tweet segmentation, is that it does not require an annotated large volume of training data to extract named entities, therefore a huge overload of annotation is avoided. In addition, this approach is not dependent on the morphology of the language. Experimental results show that the proposed method is capable of deciding on tweets to be recommended according to the user's interest. Experimental results show the applicability of the approach for recommending tweets.

5. REFERENCES

[1] D. Downey, M. Broadhead, and O. Etzioni. Locating complex named entities in web text. In *Proceedings of the 20th International Joint Conference on Artificial Intelligence*, IJCAI'07, pages 2733–2739, San Francisco, CA, USA, 2007. Morgan Kaufmann Publishers Inc.

[2] V. Levenshtein. Binary Codes Capable of Correcting Deletions, Insertions and Reversals. *Soviet Physics Doklady*, 10:707, 1966.

[3] C. Li, A. Sun, J. Weng, and Q. He. Exploiting hybrid contexts for tweet segmentation. In *Proceedings of the 36th international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '13, pages 523–532, New York, NY, USA, 2013. ACM.

[4] C. Li, J. Weng, Q. He, Y. Yao, A. Datta, A. Sun, and B.-S. Lee. Twiner: named entity recognition in targeted twitter stream. In *Proceedings of the 35th international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '12, pages 721–730, New York, NY, USA, 2012. ACM.

[5] T. Sezer. TS Corpus, The Turkish Corpus, 2014. [Online; accessed 14-December-2014].

[6] Twitter. About twitter, inc., 2014. [Online; accessed 14-December-2014].