Proceedings of the Twenty-Sixth International Florida Artificial Intelligence Research Society Conference

Exploiting Topical Perceptions Over Multi-Lingual Text for Hashtag Suggestion on Twitter

Amara Tariq
Div. of Computer Science
Univ. of Central Florida

Asim Karim
Dept. of Computer Science
LUMS, Pakistan

Fernando Gomez
Div. of Computer Science
Univ. of Central Florida

Hassan Foroosh
Div. of Computer Science
Univ. of Central Florida

Abstract

Microblogging websites, such as Twitter, provide seemingly endless amount of textual information on a wide variety of topics generated by a large number of users. Microblog posts, or tweets in Twitter, are often written in an informal manner using multi-lingual styles. Ignoring informal styles or multiple languages can hamper the usefulness of microblogging mining applications. In this paper, we present a statistical method for processing tweets according to users perceptions of topics and hashtags. Based on the non-classical notion of relatedness of vocabulary terms to topics in a corpus, which is quantified by discriminative term weights, our method builds a ranked list of terms related to hashtags. Subsequently, given a new tweet, our method can suggest a ranked list of hashtags. Our method allows enhanced understanding and normalization of users perceptions for improved information retrieval applications. We evaluate our method on a dataset of 14 million tweets collected over a period of 52 days. Results demonstrate that the method actually learns useful relationships between vocabulary terms and topics, and that the performance is better than a Naive Bayes suggestion system.

Introduction

Microblogging websites generate a huge amount of textual information written by their users. This information provides public opinion on topics of both local and global interest, and holds immense potential for information retrieval applications such as search, contextual advertising, and sentiment analysis (Efron 2011). Among microblogging websites today, Twitter is one of the most popular (Efron 2011). Started in March 2006, Twitter has grown rapidly with an estimated 140 million users generating 340 million tweets (microblog posts on Twitter) per day recently[1]. Twitter is also popular among researchers and developers because of the availability of Twitter Search API which is a handy tool for keyword-based search and retrieval of tweets.

Communications on Twitter can span a wide range of topics and languages, including informal and mixed-language writing styles. To help organize communications better, Twitter users can use hashtags[2] which are basically labels placed in tweets by users by prefixing a hash symbol to keywords or phrases. Consistent use of hashtags can improve information retrieval and topical access for users significantly. However, inclusion of hashtags in tweets is completely voluntary and user dependent. Moreover, the topics discussed under a specific hashtag tend to evolve over time. This is understandable when considering that some hashtags become micro-memes rather than topical labels (Huang, Thornton, and Efthimiadis 2010). Nonetheless, users develop a perception of hashtags based on their recent usage, which they then propagate in their own tweets.

Users' perception of hashtags and the contexts in which they are used is a powerful concept for hashtag suggestion and topical term identification. It has been demonstrated that readers/writers tend to associate terms with topics rather than terms with other terms (Haliday and Hassan 1976; Morris and Hirst 2004). This concept of relatedness of terms to topics is different from that of relatedness of term to terms (the classical notion of relatedness in linguistics), but is quite useful for text mining and information retrieval applications. Statistical measures, e.g. relative risk, can be used to quantify this concept of relatedness (Junejo and Karim 2008).

In this paper, we exploit users' perception of hashtags to learn a term-hashtag model for suggesting hashtags and identifying topical terms for tweets on Twitter. We compute a discriminative term weight for each term and context (identified by its hashtag) in the corpus. This is done by identifying current popular hashtags from a time-window of tweets, and measuring the relatedness of each vocabulary term with every popular hashtag based upon the discrimination information that particular vocabulary term provides for the hashtag. This process provides a list of vocabulary terms for each hashtag, sorted according to their relatedness with the hashtag. These lists of related terms are used to suggest popular hashtag for tweet based upon the terms used in it. In addition, these lists have potential applications in topic understanding and query expansion as they consist of popular aspects of communications in the context of a cer-

[1] http://blog.twitter.com/2012/03/twitter-turns-six.html

[2] Idea of hashtag was first introduced by Chris Messina in his blog http://factoryjoe.com/blog/2007/08/25/groups-for-twitter-or-a-proposal-for-twitter-tag-channels/

tain topic or hashtag. We evaluate our method on 14 million tweets in four time-window segments. The results show a significant improvement in hashtag suggestion over a baseline method. We also highlight and discuss the temporal and language characteristics of hashtag usage on Twitter.

Related Work

Twitter users can generate a variety of meta-data e.g. hashtag, mention (@screen name), retweet (RT) (Efron 2011). Hashtags provide topical access to tweets and their importance and usefulness have been widely explored in the context of query expansion (Massoudi et al. 2011; Lau, Li, and Tjondronegoro 2011). Better and consistent use of hashtags by Twitter users will improve topical access to tweets and performance of any system making use of hashtags. In this paper, we have presented a hashtag suggestion system which can help Twitter users select suitable hashtags for their tweets based on vocabulary terms used in their tweets. Such suggestion systems have not been widely explored in the context of microblog posts. A. Mazzia et al. have provided a preliminary suggestion system, using a Naive Bayes approach, with much focus on pre-processing steps (Mazzia and Juett). We have quantified topic-term relationships similar to human perception of text, used much larger corpus, included text from multiple languages and explored temporal behavior of both hashtags and their relevant terms. The system develops lists of terms related to each popular topic of discussion.

Topics of discussion tend to evolve rapidly over time on microblogging websites. We have trained the system over short windows of time to account for this evolution. We have developed an evaluation system without using human judgment by considering hashtag as label/topic for tweet as hashtags are supposed to provide topical access to tweets. This idea is similar to the sentiment analysis system for Twitter presented by Go et. al., which uses emoticons as judgment on sentiment of tweet (Go, Bhayani, and Huang 2009).

Motivation

In this section, we discuss the motivation for hashtag suggestion and our proposed methodology.

Dependence on Perception of Users

In this paper, we have used hashtags present in tweets as label/topics. Humans tend to group together vocabulary terms in the context of a topic, rather than associating terms with other terms (Haliday and Hassan 1976). Our system makes use of the same idea. It quantifies relatedness of terms to topics/labels to identify topic-terms relationships. Since we use hashtags provided by users as labels/topics, these topic-term relationships are actually users' perceptions of relationships between vocabulary terms and hashtags. Our system creates lists of terms related to each hashtag e.g. 'travel' (see table 4). Terms listed as related to 'travel' may not have any lexical relationship such as antonymy, synonymy or hypernymy with 'travel'. But they are related to 'travel' in the perception of users. Moreover, we have used a similarly labeled dataset for evaluation purposes, thus providing a quantitative measure of performance of the system without using explicit human-judgment on results.

Two-way Information Retrieval from Microblog Posts

The lists of terms identified as related to a topic (see table 4) can provide two-way utility for information retrieval systems. Users may put no hashtag or use less-popular hashtags in their tweet. For example, most popular hashtag used in the context of playstation 3 was 'PS3'. Some other users used less-popular hashtag of 'playstation3' in the same context. On one hand, our system can improve consistency in the use of hashtags, in turn improving performance of IR algorithms making use these tags, by suggesting suitable hashtags for tweets.

On the other hand, microblog posts or simply microblogs corpora for opinion mining are generally collected through query-matching or keyword search from microblogging websites. The terms identified as related to a topic by our system, hints towards multiple aspects of discussion under the same topic. For example, the terms related to 'Pakistan' in table 5 hints towards flood situation ('flood' , 'floodrelief' , 'CWF'), sports personalities ('Aisam') and political parties ('MQM') discussed in the context of Pakistan. Using this list of related terms as keywords for search will help cover various aspects of the single topic in the retrieved corpus.

Statistical Techniques for Informal and Multi-lingual datasets

The relationship between vocabulary terms can be either based on classical lexical rules of synonymy, antonymy, hypernymy etc. (Resnik 1999) or statistical information (Haliday and Hassan 1976). Statistical measure of relatedness between terms and topics can be evaluated based upon the frequency of co-occurrence of the terms and the frequency of occurrence of the terms in the context of a given topic. Identification of lexical relationships in microblogs corpus is hard because microblogs may contain informal language or may even be written in multiple languages. Statistical techniques can readily establish topic-term relationships on such dataset. Therefore, we have focused mainly on statistical relationships between terms and topics.

Statistical Measure of Relatedness

In this section, we shall establish quantitative measures such as discriminative term-weight (DTW) and relatedness score to quantify topic-term relationships.

Discriminative Term Weight

Discriminative term weights (DTW) are typically used to quantify the evidence or hint a term provides for a certain context or topic (Tariq and Karim 2011; Junejo and Karim 2008). Measures of discrimination information have been studied extensively for classification, association rule-mining and feature selection in literature. A detailed evaluation of three of such discrimination measures (i.e. KL divergence, relative risk, log-relative risk) has been described in (Junejo and Karim 2008). According to (Tariq and Karim

2011), the discriminative term weight of term t_j for category T_k is defined as

$$dtw(t_j, T_k) = \frac{p(t_j|T_k)}{p(t_j|\bar{T}_k)} \quad (1)$$

where $p(t_j|T_k)$ denotes the probability of term t_j in documents belonging to category T_k and $\bar{T}_k$ refers to documents in all categories but T_k. If $dtw(t_j, T_k) > 1$ then term t_j provides positive discrimination information for category T_k, with larger values representing stronger discriminative power. A labeled set of documents is required to estimate these probability values. In our case, each microblog is one document with the category/label T_k provided by hashtag of the microblog. Vector x_i of size M, where M is the size of vocabulary set, represents i^{th} document/microblog. If the term weight in each microblog is a binary value (i.e. $x_{ij} \in 0, 1$ for all i and j) following the Bernoulli distribution, then the maximum likelihood estimate of $p(t_j|T_k)$ is given by

$$\bar{p}(t_j|T_k) = \frac{\sum_{i=1}^{N_k} x_{ij}}{N_k} \quad (2)$$

where N_k is the number of microblogs belonging to hashtag T_k. Division by zero is avoided by add-one smoothing. Notice that the estimation of these probabilities requires K passes over the microblog collection, where K is the total number of popular hashtags under consideration. With all of the discriminative term weights, the set of terms $\mathcal{V}_k$ providing significant positive discrimination information for hashtag T_k is defined as

$$\mathcal{V}_k = \{t_j \| dtw(t_j, T_k) > \lambda \, \forall j\} \quad (3)$$

where $\lambda \geq 1$ is a term selection threshold controlling the exclusion of insignificant terms. In general, $\mathcal{V}_k \cap \mathcal{V}_l \neq \emptyset$ for all k and l. It means that vocabulary terms can provide significant positive discrimination information for more than one hashtags. Also, depending on the value of λ, $\cup_k \mathcal{V}_k \neq \mathcal{V}$ as some terms may not provide significant discrimination information for any hashtag (Tariq and Karim 2011).

Relatedness of Terms to Topic/Context

The measure of relatedness between a term and a topic is the product of the weight of the term in the topic and the discrimination information provided by the term for the topic (Cai and van Rijsbergen 2009). We have already described the measure of discrimination information that a term provides for a hashtag in the previous section. Conditional probability of term t_j in a given hashtag/topic T_k is the weight of the term in the topic T_k. The higher this weight, stronger the hint term t_j provides for T_k. Thus, the relatedness measure can be described as the product of $dtw(t_j, T_k)$ and $p(t_j|T_k)$ (Tariq and Karim 2011).

$$rel(t_j, T_k) = p(t_j|T_k) \times dtw(t_j, T_k) \quad (4)$$

The feature extraction based upon this relatedness measure provides a readily interpretable form of features in the textual data (Tariq and Karim 2011). Most of the traditional feature extraction methods e.g. principal component analysis, linear discriminant analysis, and latent semantic indexing lack this quality.

DTW-based Hashtag Suggestion System

Algorithm 1 DTW-based hashtag suggestion system

Input: $\{d_s\}_{s=1}^{N_a}$ (Tweets generated during time period [t1,t2]), $\{d_f\}_{f=1}^{N_b}$ (Tweets generated during time period [t2,t3]), θ_t (threshold of frequency for selection of vocabulary term), θ_K (threshold of frequency for selection of hashtag), θ_S (threshold of similarity score for suggestion of hashtag)
Output: $\{y_k\}_{k=1}^{K}$ (Vectors for each popular hashtag), $\{\mathcal{T}_f\}_{f=1}^{N_b}$ (lists of suggested hashtags for all d_f)
// Training of model
Extract all hashtags $\{T_l\}_{l=1}^{L}$ from $\{d_s\}_{s=1}^{N_a}$
Count number of occurrences Tc_l of each T_l
$\mathcal{T}_K \leftarrow \{T_k\}_{k=1}^{K}$ (set of popular hashtags) consisting of each T_l with $Tc_l > \theta_K$, $K < L$
Set of tweets $\{d_i\}_{i=1}^{N}$, consisting of every d_s which have at least one hashtag from set $\mathcal{T}_K$, $N < N_a$
tokenize every d_i to get corresponding size M binary vector representation x_i and ordered set of vocabulary terms $\mathcal{V} \leftarrow \{t_j\}_{j=1}^{M}$, frequency of t_j is at least θ_t
// $dtw(t_j, T_k)$ and $rel(t_j, T_k)$
for $k = 1 \rightarrow K$ **do**
  for $j = 1 \rightarrow M$ **do**
    $dtw(t_j, T_k) \leftarrow$ discriminative term weight of term t_j for hashtag T_k (Eq. 1)
    $rel(t_j, T_k) \leftarrow$ relatedness of term t_j to hashtag T_k (Eq. 4)
  end for
  $\mathcal{V}_k \leftarrow$ significant terms for hashtag T_k (Eq. 3)
  $y_k \leftarrow$ vector of size M corresponding to T_k, $rel(t_j, T_k)$ on all indices corresponding to t_j in $\mathcal{V}_k$, zero on all other indices
end for
//Hashtag suggestion
tokenize all d_f and get corresponding binary term-weight vector x_f over vocabulary set $\mathcal{V}$
// for each d_f
for $k = 1 \rightarrow K$ **do**
  $S_k \leftarrow x_f \cdot y_k$
  if $S_k < \theta_S$ **then**
    $S_k \leftarrow 0$
  end if
end for
list $\mathcal{T}_f$ for tweet d_f, consisting of all T_k corresponding to non-zero S_k, sorted in decreasing order of S_k

The relatedness between terms and context, described by equation 4, has been employed in dimensionality reduction algorithm for textual data, named as feature extraction based on discrimination information pooling (FEDIP) (Tariq and Karim 2011). We have used the relatedness score, calculated based on discriminative term weights (DTW), to identify topic-term relationships between vocabulary terms used in microblog posts and their topics. Our system uses these relationships to suggest hashtags for microblog posts.

Overall structure of the DTW-based hashtag suggestion

system has been presented in algorithm 1. Training algorithm generates a vector corresponding to each popular hashtag. System outputs a list of suitable hashtags for a tweet in testing data, ranked on the basis of cosine similarity score between tweet and hashtag vectors. In reality, these hashtags may be vocabulary terms already used in tweet, prompting user to put hash sign before those terms or they may be term not included in tweet, prompting user to put these terms along with hash sign in tweet. If all hashtags get similarity scores weaker than threshold θ_S, it indicates that tweet does not mention any popular topic.

Tweet are treated as a Unicode encoded strings. Tokenization is based upon occurrence of any non-alphanumeric character. This method is blind to the language(s) used in a tweet and provides crude tokenization. Only the term providing significant positive discriminative information for a hashtags have non-zero value in trained model (line 16 of algorithm 1). This step effectively filters out a lot of gibberish terms produced because of naive tokenization system. Moreover, tweets in test set are stripped of hashtag information at this step before they are fed to system for suggestion of suitable hashtags.

Data-set and Empirical Evaluation

We have evaluated our system over an extensive dataset of 14 million tweets collected over a period of 52 days during September and October of 2010. About 14.5% of all tweets have hashtags. The dataset has been divided into 4 parts, based on time of entry of tweet, to account for temporal evolution of topics over microblog-space. For each part, the system is trained over tweets generated in 12 consecutive days and evaluated over tweets containing at least one popular hashtag, generated in next 6 days. The set of popular hashtags (i.e. $\mathcal{T}_K$) for each part includes hashtags occurring with more than a certain frequency (θ_K) in that time period. We have experimented with different values of θ_K, resulting in five different sizes of $\mathcal{T}_K$ and the training and test datasets for each of four temporally divided portions of data. Table 1 provides the statistics for each part of the dataset for different $\mathcal{T}_K$ used for evaluation. For evaluation purposes, the system suggests as many top-ranked hashtags as there are actual hashtags attached with tweet by user. Table 2 provides mean precision, mean recall, mean F-score per hashtag and number of hashtags with non-zero recall. These performance measures are popular for automatic image annotation systems. Since our system provides keyword annotations for tweets, we have used similar evaluation criterion.

Similar pre-processing steps, the training and test datasets, the set of popular hashtags and the evaluation criterion have been used for all systems. We have generated baseline results using a simple Naive Bayes scoring system in which suitable hashtags for each tweet are selected on the basis of their Naive Bayes score in context to that tweet. The DTW-based system performs much better than this baseline system. We have also proposed a variation of the DTW-based system, named as FEDIP-Naive Bayes. In FEDIP-Naive bayes, dimensionality of tweets dataset is reduced using FEDIP (Tariq and Karim 2011) and then suitable hashtags for each tweet are suggested on the basis of their Naive Bayes scores. The performance of this system is comparable to the DTW-based system and much better than baseline system. These results indicate that discriminative term weight is an effective indicator of topic-term relationship. Since our proposed hashtags suggestion system relies on correct identification of these relationships, it benefits greatly if discriminative term weights are incorporated in the overall system.

Table 1: Statistics of Training and Test datasets

Part of Dataset	No. of popular hashtags	No. of tweets in Training set	No. of tweets in Test set
Part1	3621	313937	117550
	1729	275238	102950
	843	236046	89439
	293	180514	69197
	133	146247	56958
Part2	3836	343351	142383
	1887	305055	127326
	903	264169	110932
	318	207388	87363
	143	169834	72047
Part3	4193	372822	143981
	2093	332053	128383
	1015	288375	111529
	354	228372	88790
	166	186882	71653
Part4	4045	353366	122305
	2007	313163	109145
	945	270593	95002
	330	212167	74513
	150	170691	61110

The performance of our system improves non-linearly as θ_K is increased, resulting in reduced size of set $\mathcal{T}_K$. As θ_K is increased, model is trained better because of increased number of tweets for each hashtag. Moreover, the system has to chose from a smaller pool of possible hashtags as K decreases. We have conducted another set of experiments where the system suggests all hashtags for a certain tweet whose relatedness scores are greater than the chosen threshold, to observe effects of hashtag selection threshold (i.e. θ_S in algorithm 1). As this threshold increases, fewer but higher scoring hashtags are used as annotations resulting in increase in precision but decrease in recall. Figure 2 provides curves depicting change in overall precision and recall of the DTW-based and baseline systems, generated by varying this threshold θ_S, for each value of θ_K in all four parts of dataset. The DTW-based system's performance is far superior to that of baseline system.

Table 3 provides best hashtags suggested for a few tweets written without hashtags. Inclusion of such tweets in the test set will require human judgment for evaluation.

Discussion

Popular topics of discussion over microblog-space tends to evolve over time, with different topics evolving with different patterns. Since hashtags are meant to provide topical access to microblog posts, evolution patterns of topics are reflected in the evolution of corresponding hashtags. Figure 1 depicts evolution patterns for 4 different hashtags. Some hashtags e.g. 'VMA' are only popular for a short period of

Table 2: Performance comparison between three different systems

Part of Dataset	DTW-based system				Naive Bayes (baseline)				FEDIP-Naive Bayes			
	Mean Precision	Mean Recall	Mean F-score	No. of hashtags with recall>0	Mean Precision	Mean Recall	Mean F-score	No. of hashtags with recall>0	Mean Precision	Mean Recall	Mean F-score	No. of hashtags with recall>0
Part1	61.07	80.70	64.42	3216	37.67	18.74	20.80	1486	62.15	71.68	63.43	3131
	67.62	84.39	71.18	1605	54.50	27.42	30.00	1054	68.75	76.12	69.50	1589
	72.93	86.43	75.84	800	68.05	36.43	39.28	657	73.73	78.90	74.13	797
	79.46	87.72	80.73	282	75.41	49.83	52.15	260	79.78	83.22	79.90	282
	82.72	87.62	82.27	129	75.45	61.92	63.37	121	84.09	87.18	84.37	129
Part2	59.84	80.42	63.27	3416	31.57	14.03	15.80	1300	60.40	69.83	61.83	3308
	65.95	83.57	69.18	1767	45.72	20.14	22.50	955	66.70	74.68	67.73	1745
	70.71	86.29	74.06	870	62.44	27.87	30.84	634	72.32	78.73	72.93	862
	77.41	87.95	79.36	309	75.18	41.77	44.63	275	78.35	82.73	78.25	308
	84.23	90.17	84.19	141	78.61	55.87	58.78	131	85.20	88.54	85.03	141
Part3	59.07	79.84	62.45	3743	31.23	13.32	15.24	1405	59.99	68.87	61.00	3635
	64.21	82.11	67.55	1940	45.49	18.95	21.54	1053	64.20	71.58	64.84	1911
	70.06	85.77	73.44	970	59.41	25.97	28.87	684	70.26	77.16	71.25	962
	77.16	87.62	79.19	342	72.90	38.79	41.72	303	78.18	82.70	78.65	340
	80.01	87.81	80.63	161	75.32	51.52	53.69	149	80.81	85.09	80.98	160
Part4	59.18	79.47	62.53	3579	30.01	14.42	15.97	1314	59.84	69.00	61.00	3447
	64.91	83.57	68.42	1880	44.29	20.48	23.04	991	65.60	73.64	66.49	1834
	70.40	86.05	73.90	909	61.61	29.33	32.54	659	71.65	78.03	72.47	897
	76.82	88.39	79.47	320	73.84	42.59	45.64	287	78.82	83.12	78.81	319
	81.12	88.31	81.73	146	76.56	55.61	58.17	139	82.81	86.23	82.79	144

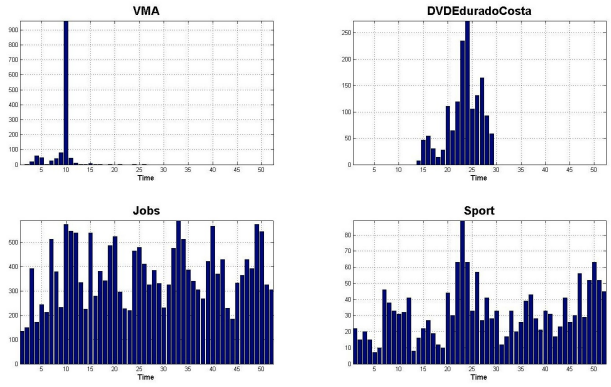

Figure 1: Number of tweets per day containing four different hashtags; each bar represents one day

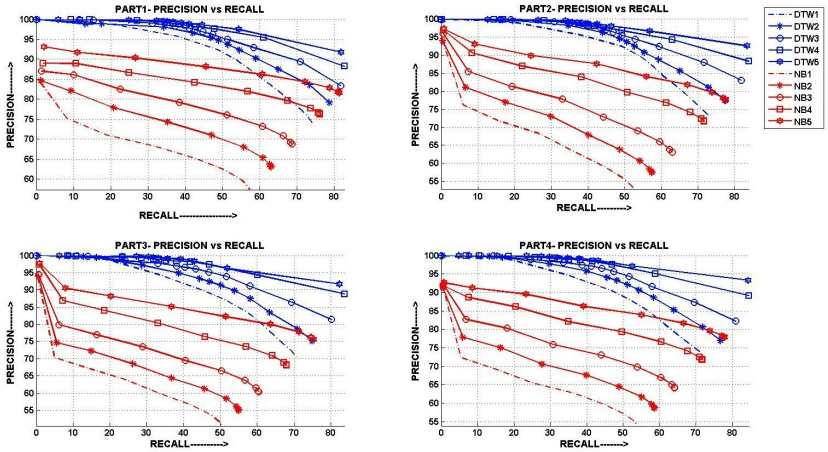

Figure 2: Precision-recall curves for baseline (NB) system and DTW-based system (DTW). Style of curve indicates specific size of training and test sets; color indicates system.

Table 3: Sample hashtags suggested for tweets with no popular hashtag

news	40 Days for Life launching London campaign :: Catholic News Agency (CNA): http://bit.ly/aINNQt via @addthis
freelance	1000s of freelance jobs available, from programming to writing! Earn cash while building your portfolio! http://tinyurl.com/37pvcfh
travel	American Airlines Labor Day Sale from $136 Round Trip (travel September 4,5,8) http://weurge.com/vacations. Denver to Los Angeles $136

time while other hashtags e.g.'travel' remain popular for an extended period. Table 4 provide lists of terms strongly related to a few topics. These lists evolve over time to keep up with evolution pattern within a topic. Table 5 provides lists of strongly related terms to hashtag 'Pakistan' over two windows of time. Our system has identified popular aspects of discussion in the context of 'Pakistan' which can be verified against the time-line of actual events e.g. 'Aisam' appears only in one window, corresponding to the time when Pakistani player named 'Aisam' won US open doubles semi-final.

Table 4: Sample of lists of vocabulary terms related to topics

tech	'tech', 'gadgets', 'intmkt', 'nolabel', 'ED' ,'mobile' 'Gree', 'cell', 'OpenLierox','LCDs'
travel	'travel', 'Augustine', 'vacations', 'tourism', 'lp' 'traveldudes', 'florida', 'tours', 'pr','VisaGuru'
sports	'sports', 'martialarts', 'college', 'baseball', 'sportsnews' 'superbowl', 'Betting', 'hockey', 'nba','tennis'
health	'health', 'wellness', 'diet', 'HealthHabits', 'weightloss' 'fitness', 'obesity', 'nutrition', 'yoga','allergies'

Table 5: Sample of terms related to hashtag 'Pakistan' from two different parts of dataset

Pakistan; Part1 (04 to 15 September,2010)
'floods', 'Pakistan', 'floodrelief', 'HelpPakistan, 'Aisam' 'victims', 'relief', 'CWF', 'savethechildren', 'PKrelief'
Pakistan; Part2 (16 to 25 September,2010)
'Pakistan', 'PKfloods', 'CWF', 'karachi', 'HelpPakistan' 'StateDept' , 'Aafia', 'MQM', 'WFP', 'Holbrooke'

Many tweets have an iso-language code, assigned to the user who created the tweet. Users who are assigned one language code, may write tweets in other languages. Still, iso-language code provides a rough estimate for different languages being used for a particular hashtag. We have observed 27 different iso-language codes in our dataset. Both English('en') and Persian('fa') are dominant languages in the context of hashtag 'iranElection'. Some hashtags describing generalized topics e.g. 'sports' also tend to have multiple dominant languages. Table 6 provides a few hashtags with multiple or non-English dominant languages.

Table 6: Language related observations for hashtags

Hashtags with multiple dominant languages
'sport' , 'freelance' , 'profile' , 'iranElection'
Hashtags with non-English dominant language
'Cuba(es)','sorteio (pt)','kaskus(id)' ,'novosti(ru)'

Conclusion

Textual data posted by users on online social networks has huge potential for information retrieval (IR) application because of its very high growth rate and coverage of vast variety of topics. We have focused on a large dataset collected from Twitter and used discriminative-term-weights to establish topic-term relationships, without using dictionary or grammar of any language. This system learns users' perception of topic-term relationships and then suggest suitable hashtags to users. Consistent use of hashtags will improve performance of all IR systems for Twitter which depend upon topical access provided by hashtags. The topic-term relationships learned through this system evolve over time. Thus, this system implicitly provides a temporally evolving summary or a pool of topics for opinion mining. In future, a hybrid system making use of both statistical and rule-based relationships between topic and terms can be developed.

References

Cai, D., and van Rijsbergen, C. J. 2009. Learning semantic relatedness from term discrimination information. *Expert Systems with Applications* 36:1860–1875.

Efron, M. 2011. Information search and retrieval in microblogs. In *Journal of the American Society for Information Science and Technology (JASIST)*, volume 62, 996–1008.

Go, A.; Bhayani, R.; and Huang, L. 2009. Twitter sentiment classification using distant supervision. http://www.stanford.edu/~alecmgo/papers/TwitterDistantSupervision09.pdf.

Haliday, M. A. K., and Hassan, R. 1976. *Cohesion in English*. London, UK: Longman.

Huang, J.; Thornton, K. M.; and Efthimiadis, E. N. 2010. Conversational tagging in twitter. In *Proceedings of the 21st ACM conference on Hypertext and hypermedia*, HT '10, 173–178. New York, NY, USA: ACM.

Junejo, K., and Karim, A. 2008. A robust discriminative term weighting based linear discriminant method for text classification. In *Data Mining, 2008. ICDM '08. Eighth IEEE International Conference on*, 323 –332.

Lau, C. H.; Li, Y.; and Tjondronegoro, D. 2011. Microblog retrieval using topical features and query expansionl. In *Text REtrieval Conference*.

Massoudi, K.; Tsagkias, M.; de Rijke, M.; and Weerkamp, W. 2011. Incorporating query expansion and quality indicators in searching microblog posts. In *Proceedings of the 33rd European Conference on Advances in Information Retrieval*, ECIR'11, 362–367. Berlin, Heidelberg: Springer-Verlag.

Mazzia, A., and Juett, J. Suggesting hashtags on twitter. http://www-personal.umich.edu/~amazzia/pubs/545-final.pdf.

Morris, J., and Hirst, G. 2004. Non-classical lexical semantic relations. In *Proceedings of the HLT-NAACL Workshop on Computational Lexical Semantics*, CLS '04, 46–51. Stroudsburg, PA, USA: Association for Computational Linguistics.

Resnik, P. 1999. Semantic similarity in a taxonomy: An information-based measure and its application to problems of ambiguity in natural language. *Journal Of Artificial Intelligence Research* 11:95–130.

Tariq, A., and Karim, A. 2011. Fast supervised feature extraction by term discrimination information pooling. In *Proceedings of the 20th ACM International Conference on Information and Knowledge Management*, CIKM '11, 2233–2236. New York, NY, USA: ACM.