Curator: Enhancing Micro-blogs Ranking by Exploiting User's Context

Hicham G. Elmongui[1,2] and Riham Mansour[3]

[1] Alexandria University, Computer and Systems Engineering
Alexandria 21544, Egypt
elmongui@alexu.edu.eg
[2] Umm Al-Qura University, Science and Technology Unit
Makkah 21955, Saudi Arabia
hgmongui@uqu.edu.sa
[3] Microsoft Research Advanced Technology Lab
Cairo 11728, Egypt
rihamma@microsoft.com

Abstract. Micro-blogging services have emerged as a powerful, real-time, way to disseminate information on the web. A small fraction of the colossal volume of posts overall are relavant. We propose Curator, a micro-blogging recommendation system that ranks micro-blogs appearing on a user's timeline according to her context. Curator learns user's time variant preferences from the text of the micro-blogs the user interacts with. Furthermore, Curator infers the user's home location and the micro-blog's subject location with the help of textual features. Precisely, we analyze the user's context dynamically from the micro-blogs and rank them accordingly by using a set of machine learning and natural language processing techniques. Curators extensive performance evaluation on a publicly available dataset show that it outperforms the competitive state-of-the-art by up to 154% on NDCG@5 and 105% on NDCG@25. The results also show that location is a salient feature in Curator.

Keywords: micro-blogs recommendation, user's context

1 Introduction

Micro-blogging services, e.g., Twitter, have emerged as a powerful real-time means of disseminating information on the web. As of January 2017, there are more than 695M Twitter users; 342M of them are active users posting on the average 518M tweets every day [47]. The high volume of tweets received by the active users is continuously increasing and is reducing productivity. About 73% of companies across the United States with 100 or more employees either completely prohibited visiting social networking sites or permitted for business purposes only [8]. With 82% of the users are active on the mobile devices [48], the effect of keeping oneself "busy" skimming through the micro-blogs is becoming apparent. With many of the micro-blogs being redundant or not of interest

to the user, the need for ranking the micro-blogs is obvious so as to be able to show her the more relevant ones first on her timeline.

In this paper, we propose Curator, a micro-blogging recommendation system that ranks the micro-blogs by exploiting the user's context. Context is defined as "any information that can be used to characterize the situation of an entity. An entity is a person, place, or object that is considered relevant to the interaction between a user and an application, including the user and applications themselves" [2]. Main components of a user's context are her identity and her location. The former is directly reflected by her preferences, which we infer from the language used in her micro-blogs. The latter may represent the current location from which she reads or writes a micro-blog, the subject location about which she authors, or her home location which affects her culture and personality. In addition to other techniques, we use natural language techniques to infer the subject location and home location of the user. Time is an inherent component of a user's context. It reflects the evolving nature of the other context components.

Building micro-blogging recommendation systems is non-trivial. First, It needs to deal with a large, and consistently increasing, corpus of micro-blogs. Second, micro-blogs themselves lack context as they are short; users are limited to a maximum of 140 characters to post in any tweet on Twitter. Third, scarcity of author's location information is another challenge. A small percentage of micro-blogs are associated with location information for privacy purposes [39]. Fourth, with the dynamic property of real life, context changes over time, and needs to be maintained for each user.

The contributions of this paper can be summarized as follows:

- We propose Curator, a micro-blogging recommendation system that ranks the micro-blogs according to the progressing user's context.
- Curator continuously captures the user's preferences by looking at the micro-blog text and the user interaction (forwardings, replies, and likes).
- Curator infers the user's home location and the micro-blog's subject location through natural language processing on the text of the tweets.
- We perform an extensive performance evaluation of Curator on a publicly available dataset. Experimental results show that Curator outperforms the competitive state-of-the-art micro-blogging recommendation systems.

The rest of the paper is organized as follows. Section 2 summarizes the related work. Curator's details are described in Section 3. In Section 4, we evaluate Curator through a meticulous performance study. We conclude the paper in Section 5.

2 Related Work

The related work to Curator is two folds: micro-blogs recommendation systems and location inference techniques for micro-blogs users.

2.1 Micro-blogs Recommendation

Many systems have been propositioned as micro-blogs recommendation systems that pick which micro-blogs to show to the user. Different micro-blogs features were adopted in the recommendation; from re-tweet (i.e., forwarding) behavior as a measure of the user's interest in a tweet [15, 49] to content relevance, account authority, and tweet-specific features that were used in learning-to-rank algorithm, which ranks the tweets [11].

The challenge in the personalized recommendation of micro-blogs is to learn the preference of the user. The basic solution asked the user to specify her static topics of interests [40] or to mark her tweets with pre-defined interest labels [18]. Next, this static preference was captured without user intervention either using collaborative ranking [6] or using a graph-theoretic model [53]. Nevertheless, the user interest was represented using Latent Dirichlet Allocation (LDA) [4], which is not scalable for real-time streams of micro-blogs [38].

The user's preferences naturally changes over time. This temporal dynamic property was lately accounted for in few personalized tweet recommendation systems. In [28, 29], LDA was used for topic modeling and a binary "important" label is predicted for each tweet. A ranking classification of tweets is proposed in [13], which models the tweet topic detection also as a classification problem.

In contrast to all the previous work that use the dynamic user's preferences as the sole feature in the recommendation, Curator uses the dynamic user's preferences as one feature in addition to the other context features of the user. In fact, the home location of the user turns out to be a salient feature in the recommendation process as shows the thorough evaluation of Curator.

2.2 Micro-blogger's Location Prediction

Research efforts trying to infer the location of the micro-blogger can be categorized into graph-based, content-based, and hybrid techniques.

The graph-based techniques use the social graph, which connects each user with its followers and followees. The user's location was inferred from her friends' by looking at the social tie and the distance between the pairs [9, 37, 41], by combining weak predictors [43], or by majority voting [26]. Furthermore, the home location is inferred from landmark users who report their true locations [52] using spatial location propagation technique [14, 27].

The content-based techniques get signals solely from the text of the microblogs. Signals include point of interests [32, 42], local words [42], location indicative words [20], or latent topics [7] to infer the home location [5], or to infer the tweet source location [23]. Besides, statistical methods are used to infer the user current location as well as her home location [12, 22, 30, 35]. An extensive feature selection comparison for location inference may be found in [21].

The hybrid approaches utilize both the social graph as well as the content of the micro-blog to predict the home location and visited locations of the user [14, 17, 33, 34]. Such approaches receive added signals from both sources and therefore have improved performance over other techniques. In this work, we adopt the Injected Inferences model [14] as a building block in Curator.

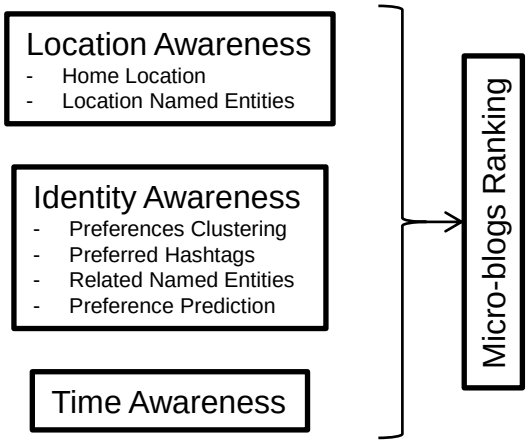

Fig. 1. Exploiting Context in Ranking

3 Curator: Micro-blogs Recommendation System

Curator is a context-aware micro-blogs recommendation system. When it ranks the micro-blogs on a timeline, it takes into account the context of its user. Therefore, it needs to be aware the identity, location, and time of the user as it appears in Figure 3. In the rest of this section, we start with the pre-processing step and the feature extraction that is done on any micro-blog prior to describing how the three context components are captured by getting signals from the micro-blogs of the user and from her interaction. Next, we show how they are incorporated in the ranking model.

3.1 Micro-blogs Textual Pre-processing and Feature Extraction

Micro-blogs are to be pre-processed in Curator. This pre-processing is needed to prepare the data for the extraction of the features used in the subsequent sections. First, the text of the micro-blog is tokenized, which removes all punctuation and other white spaces. A standard list of stop words is to be used. All URLs are also removed. Tokens containing special characters are also removed except for those starting with a hash sign, '#', which denote hashtags (e.g., #cooking). Hashtags will play a role in the classification of the user's preferences are will be described later.

Micro-blogs by definition are short and lack context. Short micro-blogs make the problem worse as they do not carry enough information. Curator discards one-word-token micro-blogs.

Micro-bloggers tend to emphasize some words by repeating some letters in those words. For instance, to enthusiastically agree, one may say "yesss" instead of "yes". The #coooold shows the strong feeling of the weather being cold. For

words containing excessively repeated letters (three or more occurrences), we just keep two occurrences and drop the others. Next, we use a spell checker, (e.g., GNU Aspell [16]) to detect out-of-vocabulary tokens and replace them with the best suggested replacement according to based on lexical and phonemic distance. Some out-of-vocabulary words are in fact slang. We use a slang dictionary to get their lexical meaning and use it as a substitute [25].

Named entities are to be extracted from the micro-blog text. We use a named entity recognizer to extract them [45]. Extracted named entities include, but are not limited to, locations, which will be used in Curator's location awareness (discussed next). Other named entity types will be used in Curator's identity awareness (detailed subsequently).

The last step in the pre-processing phase is representing the micro-blog tokens in a suitable representation for the machine learning techniques of Curator. We use term frequency-inverse document frequency (TF-IDF), which is a numerical statistic that reflect how important a word is to a document in a corpus [44]. Similar to the competitor state-of-the-art [13], the weights of the hashtags and named entities are doubled since micro-blogs with hashtags get two times more engagement [24].

3.2 Location Awareness in Curator

The location context of a micro-blogger is either the current location from which she reads or writes a micro-blog, the subject location about which she authors, or her home location which affects her culture and personality. These locations may or may not be the same. For instance, a French user may be traveling to India, but is micro-blogging about Wimbledon tournament in London, UK. A Londoner may be micro-blogging about the same event from his home.

The subject location of a micro-blog is inferred from textual signals in the micro-blog. In Curator, a location named entity recognizer is used to capture such signals. Upon detection, this subject location is fed into the identity awareness component as a signal of the micro-blog to be used to detect whether this location is preferred by the user.

The current location is either reported by the user's device, upon her permission, or is detected by the micro-blogging service. Only a small fraction of the users prefer to reveal their current location. However, the proposed ranking mechanism does not dependent on the current location by itself. If the user is interested about micro-blogs related to her current location, a micro-blog's subject location would be equal to the user's current location, and this subject location is already accounted for in Curator.

The home location of a user is either reported by the user on her profile, usually as a toponym, or may be predicted from the user's micro-blogs, her behavior on the micro-blogging service, or her friends. Curator infers the home location of the user by injecting the output of the Friends classifier described in [14] as an additional feature in the state-of-the-art content-based home location identification machine learning algorithm [35]. This home location is used as a feature in the proposed ranking model as will be shown later in this section.

3.3 Identity Awareness in Curator

The identity context the user is reflected by her preferences. Curator learns the user's preferences from her engagement on the micro-blogging service. If a micro-blog is replied to, forwarded, or liked by the user, it is a signal that the subject of the micro-blog lies within her preferred topics. Curator models the problem of predicting one's preferences by clustering the micro-blogs according to the topic preferences, classifying each cluster, and then detecting which cluster is closer to the micro-blogs that the user has engagement most.

The clustering phase is important to increase the context content of the micro-blogs' text that share the same topic. We use an online incremental clustering algorithm [3] on a corpus of micro-blogs. The resultant clusters have the properties that the micro-blogs of a cluster have larger cosine similarity among themselves [36], and hence share the same topic preference.

The classification phase labels each cluster with its topic by applying a set of topic-based binary SVM classifiers, hashtags classifiers, and named entities classifiers. The SVM classifiers are trained using predefined lists of keywords that are indicative of each adopted topic. The keyword lists are retrieved from web directories that are categorized by subjects. As an example, the list of *Food* retrieved from the Open Directory Project contains *drink*, *cheese*, and *meat* [10].

During the classification, a micro-blog may not fall in any of the existing clusters, and therefore cannot be labeled using the aforementioned SVM classifiers. For such micro-blogs, the hashtag classifiers are used to predict the topic of the micro-blog. If the micro-blog does not contain any indicative hashtags, the named entity classifiers are used for the topic prediction.

The hashtag classifier is built from the corpus used to create the clusters. Each of these hashtags are assigned a score that reflects how confident we are that the hashtag is related to the topic assigned to that cluster. Let $\text{conf}(m)$ denote the SVM confidence score of the topic predicted for a micro-blog m. Let $\text{tpcs}(h)$ denote the set of topics assigned of the clusters in which a hashtag h appears. Therefore, for each topic, t, each hashtag gets a score, $S(h|t)$.

$$S(h|t) = \frac{\sum\limits_{\substack{m \in t \\ h \in m}} \text{conf}(m)}{|\text{tpcs}(h)| + \sum\limits_{h \in m} \text{conf}(m)} \tag{1}$$

where $m \in t$ denote that micro-blog m is assigned to a cluster that is labeled with topic t. From the above equation, a hashtag gets a high value when a big fraction of its micro-blogs belong to a certain topic. The number of topics in which a hashtag appears, $|\text{tpcs}(h)|$, distinguishes between the heavily-used and lightly-used hashtags when such hashtags appear in a single topic as it prevents $S(h|t)$ from being 1. We would like to note that Equation 1 looks similar but not exact to Equation 1 in [13].

The topic with the highest score is assigned to that hashtag as shown in Equation 2. A micro-blog is assigned to the topic of a contained hashtag if that

hashtag receives a topic score above a certain threshold, $\mathbb{S} = 0.7$. We call this hashtag an indicative hashtag.

$$T(h) = \arg\max_{t} S(h|t) \tag{2}$$

The named entities classifiers are used when a micro-blog does not fall in any cluster and does not contain any indicative hashtag. A named entities classifier predicts the topic of a micro-blog if it contains a named entity. The different resources, i.e., canonical named entities, of Wikipedia [50] are retrieved along with their types from DBpedia [31]. An example resource type is *Musical Artist*. We project the types of the resources on the micro-blogs clusters and assign each resource type the same topic of preference of the corresponding cluster. Transitively, names entities of a certain resource type are assigned its assigned topic of preference. Also, synonyms to named entities are assigned their topic of preferences. Synonyms of canonical named entities are retrieved using WikiSynonyms service [51]. Examples of Synonyms of Elizabeth II are Queen Elizabeth II, Elizabeth II of England, and Her Majesty Queen Elizabeth II.

3.4 Time Awareness in Curator

Curator is aware of the current clock. Rankings of micro-blogs change over the time as the context itself changes over the time. The subject location changes with time as users move and talk about different places. This location variation is already accounted for as this subject location is detected separately for each arriving micro-blog in real time.

The user preferences also may change with time as situations progress. A user may be interested in micro-blogs about sports when a major tournament takes place, and then she gets interested in travel when she is arranging for an annual vacation. This is why Curator accounts for an adaptive preference detection.

The preference of a user is computed from the micro-blogs with which she engages. These contain the micro-blogs she liked, forwarded, or replied to. We denote such micro-blogs for a certain day, d, as M_d. The computation uses a $\text{conf}(m)$ function, which gives Curator's confidence in its prediction of the topic t of a micro-blog m. For micro-blogs that fall in any cluster and hence take its topic, this function returns the SVM confidence of the classifier corresponding to the assigned topic. The function returns 1 if the predicted topic was using the hashtag or named entities classifiers. Otherwise, $\text{conf}(m) = 0$.

Equations 3- 5 give the computation for a certain user. A daily topic preference, $\text{Pref}_d(t)$, is computed from that topic's micro-blogs with which that user has engaged on her timeline. A moving average on this daily topic preference is computed with a weekly window to produce the recent topic preference, $\text{Pref}(t)$. The user's preference in a micro-blog is computed by multiplying the confidence in predicting its topic with that topic's recent preference as shown in Equations 5.

The moving average definition of the topic preference enables its computation incrementally. Each day, it is updated by including a new day and removing the oldest day in the window. It is computed once a day for each topic for each user.

$$\mathrm{Pref}_d(t) = \sum_{\substack{m \in M_d \\ m \in t}} \mathrm{conf}(m) \tag{3}$$

$$\mathrm{Pref}(t) = \mathrm{MovingAverage}\big(\mathrm{Pref}_d(t)\big) \tag{4}$$

$$\mathrm{Pref}(m) = \mathrm{Pref}(t) * \mathrm{conf}(m) \quad \text{, where } m \text{ is of topic } t \tag{5}$$

3.5 Curator's Context Aware Micro-blogs Ranking

Curator uses a variation of the learning-to-rank model of RankSVM to rank the micro-blogs [11]. For a micro-blog m written by author a and appearing on the timeline of user u, Curator uses the following features:

- The home location of user u predicted as shown in Section 3.2.
- The micro-blog subject location as shown in Section 3.2.
- The user's adaptive topic preferences computed as described in Sections 3.4.
- The number of forwardings and likes of that micro-blog.
- The number of the author's followers, followees, and micro-blogs.
- The number of hashtags in a micro-blog.
- Was u mentioned in the micro-blog.
- Does the micro-blog contain a hashtag that u used last week.
- The number of times u mentioned, liked, or replied to a's micro-blogs.
- The number of common users both of a and u follow.
- The number of days since the last time a and u interacted together.

RankSVM, and consequetly Curator, learns the ranking function as well as the weights of the used features. The micro-blogs are shown on the user's timeline according to the learned ranking score.

4 Experimental Evaluation

We performed extensive performance evaluation of Curator against the state of the art. The machine learning algorithms were run through the WEKA suite [19]. We used a public Twitter dataset, which was used in [13, 14, 34] and is publicly available at [1]. This dataset contains 50M tweets for 3M users who have 284M following relationships. To reproduce the results of the competitor algorithm, TRUPI, we used the same sampling algorithm as in [13], which produced 10M tweets for 20K users who have 9.1 million following relationships. We also downloaded the user engagements from Twitter using its REST API [46].

As evaluation metrics, we use the micro-averaged F-measure (F1) and the normalized discounted cumulative gain (NDCG@k) and Mean Average Precision (MAP) for the ranked micro-blogs [36].

4.1 Evaluation of Binary Micro-blog Filtering

The binary filtering of micro-blogs refers to predicting whether or not the micro-blog is important to the user and will receive engagement from her through a reply, a like, or a forwarding [28].

The features used for this binary filtering are the same used in Section 3.5. The competitive baselines are the state-of-the-art binary recommendation systems that adopt a dynamic preference of the user, namely DynLDALOI and TRUPI. The major difference in both baselines is that the former uses LDA to detect the topic of interest of the user. For fairness, We compared against the J48 classifier of DynLDALOI, which gives better performance for it as shown in [28].

Table 1 shows the 10-fold cross validation for the binary micro-blog filtering. Being context-aware, Curator outperforms DynLDALOI with a relative gain of 11.3% in the micro-averaged F measure (F1). It also outperforms TRUPI with a relative gain of 6.8% on the same metric.

Table 1. 10-fold Cross Validation for Binary Micro-blog Filtering

Technique	Precision	Recall	F1
DynLDALOI	74.2%	88.6%	80.7%
TRUPI	85.7%	82.7%	84.1%
Curator	93.7%	86.4%	89.9%

4.2 Evaluation of Curator Context-Aware Ranking

We performed extensive experimentation to evaluate Curator and to compare it against the state of the art recommendation systems that rank micro-blogs. We compared Curator against the 5 baselines: 1) RetweetRanker [15], whose metric of measuring user's interest is her re-tweet behavior; 2) RankSVM [11], which produces a ranking score by learning the ranking function and the weights of the input features; 3) DecisionTreeClassifier [49], which uses the tweet re-tweeting behavior to build a decision tree classifier that is used in its ranking model; 4) GraphCoRanking [53], which represents the preferences using LDA; and 5) TRUPI [13], which does not account for the home location of the author or the subject location of the micro-blog.

While comparing these techniques, the used ground truth was whether the micro-blog got any engagement from the user; i.e., whether it was replied to, forwarded, or liked by the user. Table 2 gives the evaluation of Curator and its competitor baselines using NDCG@k metric for the values of $k = 5$, 10, 25, and 50, whereas Figure 2 gives the evaluation between the same techniques using the MAP metric. On NDCG@k, Curator consistently outperforms all other competitive baselines for all the used values of k. Specifically, Curator outperforms RetweetRanker by 154%, 117%, 105%, and 107% on NDCG@5, NDCG@10,

Table 2. Personalized Ranking - NDCG@k Metric

Technique	k=5	k=10	k=25	k=50
RetweetRanker	0.217	0.274	0.303	0.342
RankSVM	0.222	0.290	0.326	0.372
DecisionTreeClassifier	0.342	0.401	0.429	0.487
GraphCoRanking	0.411	0.455	0.462	0.538
TRUPI	0.508	0.543	0.577	0.615
Curator	0.551	0.595	0.622	0.706

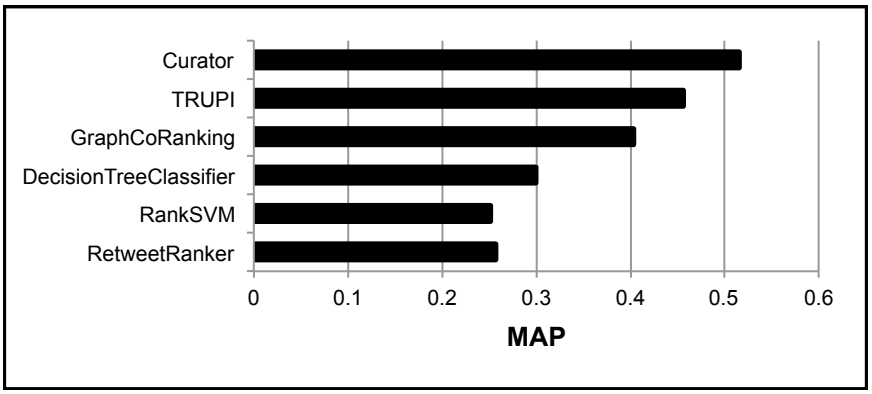

Fig. 2. Personalized Ranking - MAP Metric

NDCG@25, and NDCG@50 respectively. Curator outperforms the closest competitor, TRUPI, by 8%, 10%, 8%, and 15% on the same metrics. On MAP, Curator outperforms TRUPI by 13%.

4.3 Curator's Context Awareness Effect

Curator is aware of three context components, namely, time, identity, and location. From Section 4.2, the closest competitor was TRUPI. TRUPI already accounts for the dynamic level of interest of a user in the topic of the tweets. In this experiment, we compose a version of Curator that is not aware of the location by discarding the first two location-related features that are used in the ranking model in Section 3.5. We compare this version against the proposed Curator.

Table 3 and Figure 3 give the evaluation of Curator with and without the location context using both the NDCG@k and MAP metrics. Including the location context in Curator indeed improved its performance by 12%, 12%, 10%, 18%, and 16% on NDCG@5, NDCG@10, NDCG@25, NDCG@50, and MAP, respectively. This is why we believe that the location context is a salient feature in Curator.

Table 3. Curator Context Awareness Effect - NDCG@k Metric

Context	k=5	k=10	k=25	k=50
Identity + Time	0.493	0.531	0.563	0.600
Location + Identity + Time	0.551	0.595	0.622	0.706

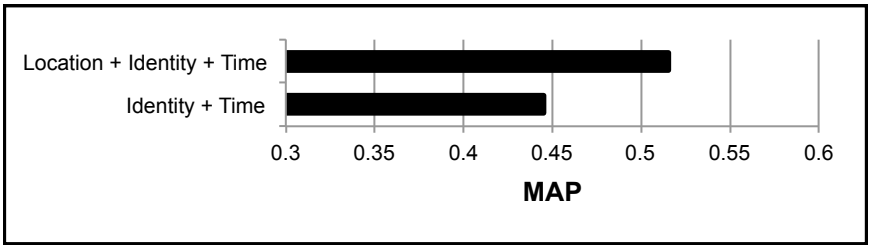

Fig. 3. Curator Context Awareness Effect - MAP Metric

5 Conclusion

In this paper, we proposed Curator, a context-aware micro-blogging recommendation system that is used to rank the micro-blogs according to the user's identity, time, and location contexts. Curator learns the user's time variant preferences from the text of the micro-blogs she engages with. Moreover, Curator infers the user's home location and the micro-blog's subject location with the help of textual features from the micro-blog. We performed an extensive performance evaluation on a publicly available dataset. Curator outperforms the competitive state-of-the-art by up to 154% on NDCG@5 and 105% on NDCG@25. The results also show that location is a salient feature in Curator.

Acknowledgement

This material is based on work supported in part by (1) Research Sponsorship from Microsoft Research, (2) the KACST National Science and Technology and Innovation Plan under grant 14-INF2461-10.

References

1. https://wiki.cites.illinois.edu/wiki/display/forward/Dataset-UDI-TwitterCrawl-Aug2012 (2012)
2. Abowd, G.D., Dey, A.K., Brown, P.J., Davies, N., Smith, M., Steggles, P.: Towards a better understanding of context and context-awareness. In: HUC (1999)
3. Becker, H., Naaman, M., Gravano, L.: Beyond trending topics: Real-world event identification on twitter. In: ICWSM (2011)
4. Blei, D.M., Ng, A.Y., Jordan, M.I.: Latent dirichlet allocation. Journal of Machine Learning Research 3 (2003)

5. wen Chang, H., Lee, D., Eltaher, M., Lee, J.: @phillies tweeting from philly? predicting twitter user locations with spatial word usage. In: ASONAM (2012)
6. Chen, K., Chen, T., Zheng, G., Jin, O., Yao, E., Yu, Y.: Collaborative personalized tweet recommendation. In: SIGIR (2012)
7. Cheng, Z., Caverlee, J., Lee, K.: You are where you tweet: A content-based approach to geo-locating twitter users. In: CIKM (2010)
8. CIO Survey: http://rht.mediaroom.com/index.php?s=131&item=790 (2009)
9. Compton, R., Jurgens, D., Allen, D.: Geotagging one hundred million twitter accounts with total variation minimization. In: BigData (2014)
10. DMOZ - the Open Directory Project: http://www.dmoz.org/ (2014)
11. Duan, Y., Jiang, L., Qin, T., Zhou, M., Shum, H.Y.: An empirical study on learning to rank of tweets. In: COLING (2010)
12. Eisenstein, J., O'Connor, B., Smith, N.A., Xing, E.P.: A latent variable model for geographic lexical variation. In: EMNLP (2010)
13. Elmongui, H.G., Mansour, R., Morsy, H., Khater, S., El-Sharkasy, A., Ibrahim, R.: TRUPI: Twitter Recommendation Based on Users' Personal Interests. In: CICLing (2015)
14. Elmongui, H.G., Morsy, H., Mansour, R.: Inference models for twitter user's home location prediction. In: AICCSA (2015)
15. Feng, W., Wang, J.: Retweet or not?: Personalized tweet re-ranking. In: WSDM (2013)
16. GNU Aspell: http://aspell.net/ (2011)
17. Gu, H., Hang, H., Lv, Q., Grunwald, D.: Fusing text and frienships for location inference in online social networks. In: WI-IAT'12 (2012)
18. Guo, Y., Kang, L., Shi, T.: Personalized tweet ranking based on ahp: A case study of micro-blogging message ranking in t.sina. In: WI-IAT (2012)
19. Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., Witten, I.H.: The WEKA data mining software: an update. SIGKDD Explorations 11(1) (2009)
20. Han, B., Cook, P., Baldwin, T.: Geo-location prediction in social media data by finding location indicative words. In: COLING (2012)
21. Han, B., Cook, P., Baldwin, T.: Text-based twitter user geolocation prediction. Journal of Artificial Intelligence Research 49(1) (2014)
22. Hecht, B., Hong, L., Suh, B., Chi, E.H.: Tweets from justin bieber's heart: The dynamics of the location field in user profiles. In: CHI (2011)
23. Hong, L., Ahmed, A., Gurumurthy, S., Smola, A.J., Tsioutsiouliklis, K.: Discovering geographical topics in the twitter stream. In: WWW (2012)
24. Huffington Post's Twitter Statistics: http://www.huffingtonpost.com/belle-beth-cooper/10-surprising-new-twitter_b_4387476.html (2013)
25. Internet Slang Dictionary & Translator: http://www.noslang.com/ (2014)
26. Jr., C.A.D., Pappa, G.L., de Oliveira, D.R.R., de Lima Arcanjo, F.: Inferring the location of twitter messages based on user relationships. Transactions in GIS 15(6) (2011)
27. Jurgens, D.: That's what friends are for: Inferring location in online social media platforms based on social relationships. In: ICWSM (2013)
28. Khater, S., Elmongui, H.G., Gracanin, D.: Personalized microblogs corpus recommendation based on dynamic users interests. In: SocialCom (2013)
29. Khater, S., Elmongui, H.G., Gracanin, D.: Tweets You Like: Personalized Tweets Recommendation based on Dynamic Users Interests. In: SocialInformatics (2014)
30. Kinsella, S., Murdock, V., O'Hare, N.: "i'm eating a sandwich in glasgow": Modeling locations with tweets. In: SMUC (2011)

31. Lehmann, J., et al.: DBpedia - a large-scale, multilingual knowledge base extracted from wikipedia. Semantic Web Journal (2014)
32. Li, C., Sun, A.: Fine-grained location extraction from tweets with temporal awareness. In: SIGIR (2014)
33. Li, R., Wang, S., Chang, K.C.C.: Multiple location profiling for users and relationships from social network and content. PVLDB 5(11) (2012)
34. Li, R., Wang, S., Deng, H., Wang, R., Chang, K.C.C.: Towards social user profiling: Unified and discriminative influence model for inferring home locations. In: KDD (2012)
35. Mahmud, J., Nichols, J., Drews, C.: Home location identification of twitter users. ACM TIST 5(3) (2014)
36. Manning, C.D., Raghavan, P., Schütze, H.: Introduction to Information Retrieval. Cambridge University Press (2008)
37. McGee, J., Caverlee, J., Cheng, Z.: Location prediction in social media based on tie strength. In: CIKM (2013)
38. Mikolov, T., Chen, K., Corrado, G., Dean, J.: Efficient estimation of word representations in vector space. In: ICLR Workshops (2013)
39. Morstatter, F., Pfeffer, J., Liu, H., Carley, K.M.: Is the sample good enough? comparing data from twitter's streaming api with twitter's firehose. In: ICWSM (2013)
40. Pennacchiotti, M., Silvestri, F., Vahabi, H., Venturini, R.: Making your interests follow you on twitter. In: CIKM (2012)
41. Rout, D., Bontcheva, K., Preoţiuc-Pietro, D., Cohn, T.: Where's @wally?: A classification approach to geolocating users based on their social ties. In: HT (2013)
42. Ryoo, K., Moon, S.: Inferring twitter user locations with 10 km accuracy. In: WWW Companion (2014)
43. Sadilek, A., Kautz, H., Bigham, J.P.: Finding your friends and following them to where you are. In: WSDM) (2012)
44. Salton, G., Buckley, C.: Term-weighting approaches in automatic text retrieval. Information Processing & Management 24(5) (1988)
45. Twitter NLP Tools: https://github.com/aritter/twitter_nlp (2011)
46. Twitter REST API: https://dev.twitter.com/docs (2014)
47. Twitter Statistics: http://www.statisticbrain.com/twitter-statistics/ (2017)
48. Twitter Usage: http://about.twitter.com/company (2014)
49. Uysal, I., Croft, W.B.: User oriented tweet ranking: a filtering approach to microblogs. In: CIKM (2011)
50. Wikipedia: http://www.wikipedia.org/ (2001)
51. WikiSynonyms: http://wikisynonyms.ipeirotis.com/ (2012)
52. Yamaguchi, Y., Amagasa, T., Kitagawa, H.: Landmark-based user location inference in social media. In: COSN (2013)
53. Yan, R., Lapata, M., Li, X.: Tweet recommendation with graph co-ranking. In: ACL (2012)