Location-aware online learning for top-k hashtag recommendation

Róbert Pálovics[1,2] Péter Szalai[1,2] Levente Kocsis[1,4]
Júlia Pap[1] Erzsébet Frigó[1,3] András A. Benczúr[1,3]
[1]Institute for Computer Science and Control, Hungarian Academy of Sciences (MTA SZTAKI)
[2]Technical University Budapest [3]Eötvös University Budapest [4]University of Szeged
{rpalovics, benczur, kocsis, pszalai, fbobee papjuli}@ilab.sztaki.hu

ABSTRACT

In this paper we investigate the problem of recommending Twitter hashtags for users with known GPS location, learning online from the stream of geo-tagged tweets. Our method learns the relevance of regions in a geographical hierarchy, combined with the local popularity of the hashtag. Unlike in typical collaborative filtering settings, trends and geolocation turns out to be more important than personalized user preferences. We evaluate in a time-aware setting, where evaluation is cumbersome by traditional measures, since we have different top recommendations at different times. We describe a time-aware framework based on individual item discounted gain.

1. INTRODUCTION

We investigate the problem of recommending Twitter hashtags for users, based on the temporal geolocation information of both the users and the hashtags. Our aim is to recommend new hashtags, i.e. hashtags that the user has not used before. The recommendations are obtained by learning online from the stream of geo-tagged tweets. In our task the novel element is that neither users nor items (hashtags) are bound to one single location. Hashtags may in fact relate to certain locations as well as be popular worldwide. Earlier results on recommendation in location-based social networks surveyed in e.g. [1, 13] combine spatial ratings for non-spatial items, nonspatial ratings for spatial items, and spatial ratings for spatial items [10]. Our new results address the problem of the fuzzy relation of users and hashtags with locations.

Since hashtag usage is highly volatile, the problem calls for an online method. Whenever a user sends a geotagged tweet with a hashtag he or she has not used earlier, we consider the event as a trigger for recommendation. We measure the accuracy of our methods in the online evaluation framework of [12] based on discounted cumulative gain (DCG) computed individually for each event and averaged over time.

LocalRec'15, September 19, 2015, Vienna, Austria.

We find that location and timing are *the* key factors with little contribution from personalized user interest. The locality of Twitter hashtag adoption in both spatial and temporal sense is observed among others by Kamath et al. [7]. They state that "hashtags are a global phenomenon [...] but distance between locations is a strong constraint on the adoption [...] and follow a spray-and-diffuse pattern".

We use a four-month collection of 400 million geotagged Twitter messages detailed in [6]. We discard the text of the tweet messages and keep only the hashtags, the timestamp and the GPS coordinates. In our experiments we focus on the user location and the new hashtags that appear in the message. As we have no information on which tweets are read by the users but we know the new hashtags they tweeted, we use the hashtag publishing information to measure user topic adoption. We consider a hashtag newly adopted if we have not observed the given user-hashtag pair before in the dataset. To guarantee at least one month for each user-hashtag pair without activity, we simply skip the first month of the stream of new hashtag usages.

Some content may have obvious connection to certain locations but others can have more widespread interest on different levels such as language, continent, or even worldwide. Dealing with this, our models rely on the hierarchy of regions from a global or continent-wide level down to a village or city district to attribute the momentary popularity of a hashtag to levels of locations. This hierarchical property of locations is surveyed also in [1]. We use the open hierarchical database of Global Administrative Areas (GADM, `http://gadm.org`). We mention that the metadata of tweets may contain not only GPS coordinates but also a *place* attribute that can contain the name and type of the place. However, we found the place attribute often ambiguous and less reliable.

We use two models to recommend hashtags at a given time and location, one based on the estimated probability of the hashtag appearance based on its recency, and another based on its temporal popularity. In both cases, our new method learns the importance of each node in the GADM tree. The final prediction arises as the weighted combination of the hashtag probabilities along the path of the GADM tree from the leaf location of the user up to the root.

As baseline method, we use online matrix factorization [12]. Surprisingly, it turns out that matrix factorization performs much weaker than the distance based methods and contributes relatively little to the final prediction. This observation justifies the importance of the temporal and geographic context of Twitter messages.

1.1 Related work

Most of previous publications on geographic recommender systems work with *check-in data*, where each of the items has a predefined static location. For brevity, we do not survey these here. *Hashtag recommendations* are addressed in two recent papers: Chen et al. [3] give methods for efficiently maintaining a sliding window for time aware recommendation, and Diaz et al. [5] introduce methods to compute matrix factorization online. These results are orthogonal to our exploitation of the location information.

Spatial statistics of hashtag adoption are analyzed by Kamath et al. [7]. Cheng et al. [4] give methods to geolocalize tweets based on content. Mocanu et al. [11] use a data set similar to ours to analyze geographical properties like homogeneity and seasonal patterns of language usage at scales ranging from country-level to city neighborhoods. Similar to our use of the Global Administrative Areas, *regions-of-interests partitioning* is examined in [9] by applying k-means clustering to establish natural regions over Twitter data. None of these papers exploit the results in recommender systems.

No other results use *external data to define the hierarchy of locations* for recommendation tasks. Similar to our result, in [6], GADM is used over the same Twitter data set, but only for visualization purposes.

2. ONLINE RECOMMENDATION AND EVALUATION

We use the online recommendation framework described in [12], in which model training and evaluation happen simultaneously, iterating over the dataset only once, in chronological order. Whenever we see a new tweet, we assume that the user becomes active and reveals its location to the recommender system. Next, we recommend hashtags of potential interest for the user. The recommendation is online, hence it depends on the context at the exact time instance of the tweet. If a user u tweets with hashtag h at time t in location ℓ, our models give a score $\hat{r}(u, h', \ell, t)$ for each hashtag h' seen so far, and recommend to u the k hashtags with the largest values from those that u has not used before.

The data is implicit: the events imply only that the user is interested in a hashtag. In most of our models, we need negative instances as well for training. We use all hashtag usages as positive training instances and generate negative training instances by selecting `negRate` random hashtags uniformly at the time when a user first used a hashtag. We tested the `negRate` parameter between 1 and 300.

We use the quality metric of [12] that we adopt to hashtag recommendation. If h is the new hashtag in the message and the rank of h returned by the recommender system is $\text{rank}(h)$, then the discounted cumulative gain, DCG@k of this event is $\frac{1}{\log_2(\text{rank}(h)+1)}$ if $\text{rank}(h) \leq k$, and 0 otherwise. The overall evaluation of a model is the average cumulative DCG@k.

3. TWITTER AND GEOGRAPHICAL DATA

Dobos et al. [6] collected the dataset using the Twitter open API by requesting *geotagged tweets*. We used the data between February 1 and May 30, 2012 with February for training and observing distributions only, hence the online learning period lasts three months.

Table 1: Properties of the cleansed dataset.

number of records	6,978,478
number of unique user-hashtag pairs	2,993,183
number of users	792,860
number of hashtags	268,489
number of countries	49

Most of the hashtags in the database are quite rare, thus we use only the hashtags that appear more than 5 times. This way we exclude about 90% of the hashtags, but most of the hashtag timeline remains. We also exclude the hashtags that appear in the first month of the collection to recommend newly spreading hashtags for the users. The properties of the final cleansed dataset are summarized in Table 1.

We collected all 214,230 nodes from the GADM database, from which 190,315 are leaves. The depth of the tree is 6, and includes 5 levels from the GADM tree plus continent-country relations. The hashtag time series data covered 30,450 leaves from the tree.

4. MODELING

4.1 Recommendation by location hierarchy

In our recommendation model we use the location ℓ at time t of the user. Here ℓ notes the leaf of the tree that is closest to the current GPS location of the user. First, we get the path in the tree from the root node to location ℓ, $\text{Path}(\ell)$. Next, for a given hashtag, assume we have a recommendation method that yields scores $s(h, n, t)$ for nodes n along $\text{Path}(\ell)$. We will give two such methods in the next subsections. In order to aggregate the individual recommendation at each GADM tree node, we propose the formula

$$\hat{r}(u, h, \ell, t) = \sum_{n \in \text{Path}(\ell)} w_n \cdot s(h, n, t),$$

where w_n values are node specific weights. The weights w_n are independent of the hashtags and characterize the area n only. We learn the weights by online gradient descent by optimizing for RMSE. If we consider all positive instances and generate negative ones as described in Section 2, we will have sufficiently many implicit data to update the weights online as we read the sequence of events.

In our experiments we also investigate models where we set all w_n values constant, i.e. we do not learn the weights, but simply sum all $s(h, n, t)$ values along $\text{Path}(\ell)$.

4.2 Temporal popularity

Given a predefined time discretization that we test between a minute and a day, for each location in the tree, we compute the number of occurrences of the hashtag in the time interval at the given location. As it follows power-law distribution, we use the logarithm of the temporal popularity values as node scores: $s(h, n, t) = \log(\text{pop}(h, n, t))$, where $\text{pop}(h, n, t)$ denotes the number of occurrences of hashtag h in node n in the time interval ending at time t.

4.3 Hashtag recency

Our next method estimates the chance of the appearance of a hashtag by considering its most recent usage. The advantage of this method is that it is more sensitive to changes

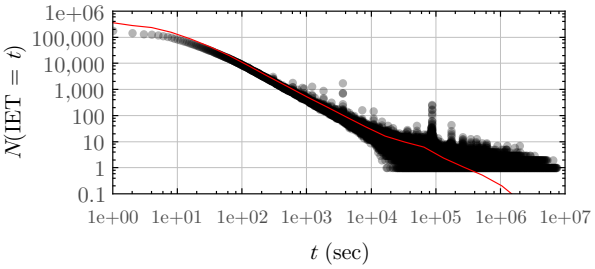

Figure 1: Inter-event distribution.

in trends. While it may more aggressively overfit to single events, overall it performs similar to and combines very well with the popularity based method. In Fig. 1, we investigate the distribution of the time elapsed between the same hashtag appearing in tweets. This inter-event time distribution follows power law, in accordance with several earlier observations [2, 14, 8], $P(\tau = t) = (\alpha - 1) \cdot t^{-\alpha}$, whence we easily get

$$P(t < \tau \leq t + \Delta t \mid \tau > t) = 1 - \left(1 + \frac{\Delta t}{t}\right)^{(1-\alpha)}. \quad (1)$$

For location sensitive prediction we maintain the last appearance of each hashtag for every node in the geolocation tree. We compute the estimate of (1) in each node by using the global measured value $\alpha = 1.2$.

4.4 Online matrix factorization

We apply stochastic gradient descent factorization for the user-hashtag matrix as in [12]. Batch stochastic gradient descent iterates several times over the training set until convergence. Online recommenders seem to be more restricted than those that may iterate over the data set several times. However, online matrix factorization proved to be superior to batch in [12], since it gives much more emphasis on recent events.

5. EXPERIMENTS

In our graphs we show the average cumulative DCG for the first three weeks, by when all of our methods reach stable performance. Here the cumulative average corresponds to cumulative time average. We set $k = 100$ to compare our methods in detail. All methods show a slight performance degradation, which is due to the fact that the number of possible hashtags to recommend increase in time.

5.1 Online matrix factorization

We used the online version of stochastic gradient descent (SGD) matrix factorization algorithm of [12]. We applied the mean square error with user and item regularization terms of weight `regRate` as our objective function. Since our data is implicit and contains only positive interactions, we generated negative samples. Every time a user *first* posts about a hashtag, we generate for her `negRate` hashtags that she have not used in her past as described in Section 2. In all cases, we set `negRate`=99, learning rate `lRate`=0.4, and `regRate`=0.01.

As we show next, our tree based methods and their baseline variants to exploit geographical information resulted in better performance in our experiments.

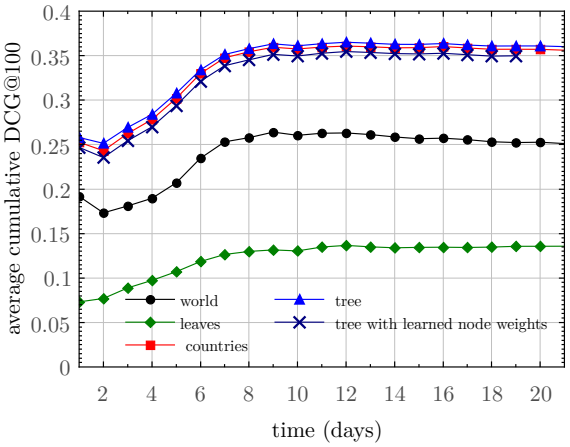

Figure 2: Average cumulative DCG@100 for the tree based popularity model and its baseline variants.

5.2 Popularity and recency based methods

The temporal popularity based method using the GADM tree of Section 4.2 achieved best results by setting the time frame around 2 hours. In the recency based model of Section 4.3, the parameter Δt had relatively little effect, we set $\Delta t = 12h$. We compared the popularity and recency based methods separately in Figs. 2 and 3, resp., by using different levels of the GADM tree and turning recency and node weight learning on and off.

In the figures, "world" denotes the methods that use global values only and do not take user geolocation into account, while "leaves" and "countries" use the corresponding level of the tree. Note that country level popularity and recency performed very well while leaves worked only with recency not popularity. Note that using countries but no temporal information at all performs the poorest.

Best performance is obtained when using the whole tree for recommendation by adding all recency values along the path corresponding to the current user location in the tree, marginally improving the country based results. However, by applying the gradient method to learn node specific weights as in Section 4.1, we could achieve significantly better results for recency but not for popularity. By using the recency based tree learning algorithm, we were able to focus on the active and representative part of the tree. We achieved our best results with `lRate`=0.0001 and `negRate`=4.

5.3 Online combination

In our final experiments we compared and combined our strongest methods. In Figure 4 we plotted the average cumulative DCG@100 as the function of time for our best models. Surprisingly, the tree based methods strongly outperform online matrix factorization, while the best popularity based model overtook the best recency based method a little bit in the long run. Next, we considered the strongest one, the tree based popularity without node weight learning, to improve it by combining it with the best factor model and recency recommender. We used the SGD based double layer combination method introduced in [12] with mean squared error as objective function. In Figure 5 we show the results of the combination. The popularity model can be improved by using the best recency method that uses the tree with

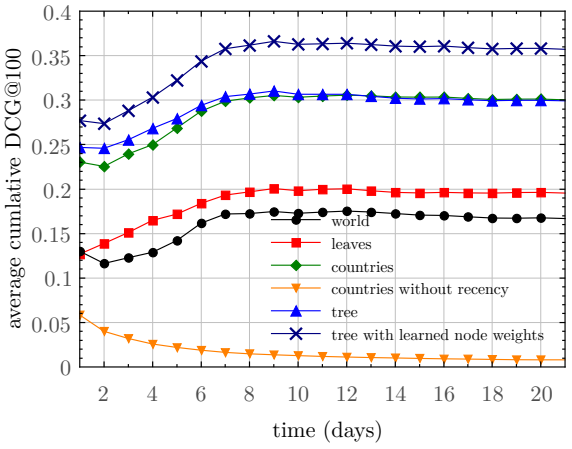

Figure 3: Average cumulative DCG@100 curves of the recency based methods.

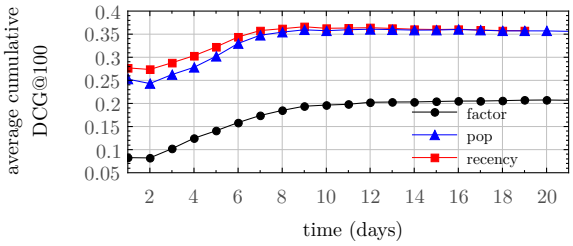

Figure 4: Best performances achieved with the factor, the popularity, and the recency based recommenders.

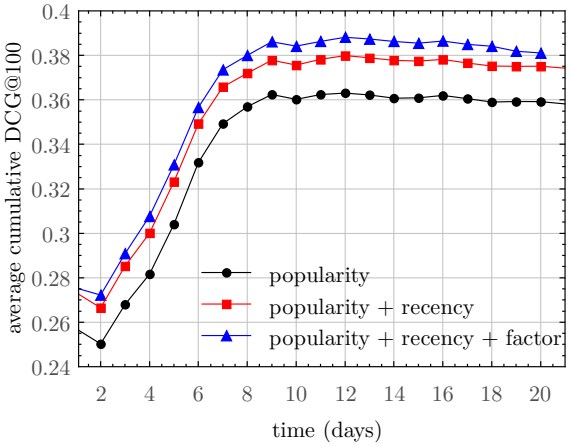

Figure 5: Combination of the best three different models.

learned node weights. We could further improve our results by including the factor model in our recommendation. In Table 2 we collected the overall performance of our best methods and their combinations.

6. CONCLUSION AND FUTURE WORK

We gave online, location based learning methods for Twitter hashtag recommendation. Since hashtags are not directly bound to a location, may be geographically spread, and vary in popularity at different times, we designed methods that exploit the time and location context. Surprisingly, user personalization has little contribution to recommendation quality, hence our best methods apply in the user cold start setting as well.

Table 2: Best performance methods and their combination, with relative improvement.

	DCG@100	DCG@10
factorization	0.206	0.180
recency w/ tree learning	0.355	0.323
popularity w/ tree learning	0.359	0.335
popularity + recency	0.374 (*4.1%*)	0.342 (*2%*)
popularity + recency + factor	0.381 (*6.1%*)	0.35 (*4.2%*)

7. REFERENCES

[1] J. Bao, Y. Zheng, D. Wilkie, and M. F. Mokbel. A survey on recommendations in location-based social networks. *ACM Trans. Intelligent Systems and Technology*, 2013.

[2] A.-L. Barabási. The origin of bursts and heavy tails in human dynamics. *Nature*, 435(7039):207–211, 2005.

[3] C. Chen, H. Yin, J. Yao, and B. Cui. Terec: A temporal recommender system over tweet stream. *Proceedings of the VLDB Endowment*, 6(12):1254–1257, 2013.

[4] Z. Cheng, J. Caverlee, and K. Lee. You are where you tweet: a content-based approach to geo-locating twitter users. In *Proc. CIKM*, pp. 759–768. ACM, 2010.

[5] E. Diaz-Aviles, L. Drumond, L. Schmidt-Thieme, and W. Nejdl. Real-time top-n recommendation in social streams. In *Proc. RecSys*, pp. 59–66. ACM, 2012.

[6] L. Dobos, J. Szüle, T. Bodnár, T. Hanyecz, T. Sebők, D. Kondor, Z. Kallus, J. Stéger, I. Csabai, and G. Vattay. A multi-terabyte relational database for geo-tagged social network data. In *Cognitive Infocommunications (CogInfoCom), 2013 IEEE 4th International Conference on*, pp. 289–294. IEEE, 2013.

[7] K. Y. Kamath, J. Caverlee, K. Lee, and Z. Cheng. Spatio-temporal dynamics of online memes: a study of geo-tagged tweets. In *Proc. WWW*, pp. 667–678, 2013.

[8] M. Kivelä and M. A. Porter. Estimating inter-event time distributions from finite observation periods in communication networks. *arXiv preprint arXiv:1412.8388*, 2014.

[9] R. Lee and K. Sumiya. Measuring geographical regularities of crowd behaviors for twitter-based geo-social event detection. In *Proceedings of the 2nd ACM SIGSPATIAL international workshop on location based social networks*, pp. 1–10. ACM, 2010.

[10] J. J. Levandoski, M. Sarwat, A. Eldawy, and M. F. Mokbel. Lars: A location-aware recommender system. In *Proc. ICDE*, pp. 450–461. IEEE, 2012.

[11] D. Mocanu, A. Baronchelli, N. Perra, B. Gonçalves, Q. Zhang, and A. Vespignani. The twitter of babel: Mapping world languages through microblogging platforms. *PloS One*, 8(4):e61981, 2013.

[12] R. Pálovics and A. A. Benczúr. Temporal influence over the last.fm social network. *Social Netw. Analys. Mining*, 5(1):4, 2015.

[13] P. Symeonidis, D. Ntempos, and Y. Manolopoulos. Location-based social networks. In *Recommender Systems for Location-based Social Networks*, pp. 35–48. Springer, 2014.

[14] A. Vazquez, B. Rácz, A. Lukács, and A.-L. Barabási. Impact of non-poissonian activity patterns on spreading processes. *Physical review letters*, 98(15):158702, 2007.