Long time, no tweets! Time-aware personalised hashtag suggestion

Morgan Harvey [1] and Fabio Crestani [2]

[1] Dept. of Maths and Info. Sciences, Northumbria University at Newcastle, UK.
[2] Faculty of Informatics, University of Lugano, Switzerland
morgan.harvey@northumbria.ac.uk, fabio.crestani@usi.ch

Abstract. Microblogging systems, such as the popular service Twitter, are an important real-time source of information however due to the amount of new information constantly appearing on such services, it is difficult for users to organise, search and re-find posts. Hashtags, short keywords prefixed by a # symbol, can assist users in performing these tasks, however despite their utility, they are quite infrequently used. This work considers the problem of hashtag recommendation where we wish to suggest appropriate tags which the user could assign to a new post. By identifying temporal patterns in the use of hashtags and employing personalisation techniques we construct novel prediction models which build on the best features of existing methods. Using a large sample of data from the Twitter API we test our novel approaches against a number of competitive baselines and are able to demonstrate significant performance improvements, particularly for hashtags that have large amounts of historical data available.

1 Introduction

Social-media update streams are fast becoming a key mode of information access on the web, with many services basing their offerings on this paradigm. One of the most popular of these is Twitter, which has become remarkably successful in recent years (10% of online Americans use the service on a typical day [7]). Twitter is a *microblogging* platform which allows users to post short messages (of up to 140 characters) to share thoughts, opinions, useful links, and insights from their personal experiences. Users are encouraged to "follow" others on the service whose posts (or *tweets*) may be of interest to them. Doing so results in all of the posts created by that user appearing on the follower's stream.

Although Twitter represents a highly valuable, user-driven and up-to-date source of information of unprecedented volume, evidence suggests that high volumes of tweets can become overwhelming for users. Nearly half of all Twitter search tasks involve re-finding previously seen tweets from the stream, a task which was reported to be amongst the most difficult [7]. A feature called *hashtags*, short keywords prefixed by a # symbol, a practise which emerged organically through use of the system, allows the topic(s) of each tweet to be specifically defined by the author. Hashtags provide users with a means to more easily search,

browse and re-find tweets, form ad-hoc communities based around a hashtag's topic and follow the evolution of discussions or breaking news stories [10].

Despite the clear utility of hashtags and their ability to promote the tweets to which they are assigned [10], only a relatively small number of tweets - as few as 8% [11] - contain them. As there is no pre-defined set of hashtags to choose from when writing a tweet, users can choose any terms they wish, leading to vocabulary mismatch problems. Given the benefits of appropriate hashtag usage and the reluctance many users have in employing them (perhaps because they find selecting the best terms difficult), the problem of hashtag recommendation is important. By recommending hashtags during the tweeting process we aim to support users in allocating terms to their posts and increase the homogeneity of hashtag usage on Twitter as a whole. Since hashtag usage has been shown to be heavily dependent on time, user interests and of course the topics of the parent tweet, we attempt to incorporate these three sources of information into our recommendation models. We test several novel approaches on real Twitter data collected over a period of one month and compare the performance of our models against competitive baselines from the literature.

2 Related Work

Twitter's popularity and the existence of a public API has led to it becoming a common topic of research interest. A large amount of early work focused on understanding how networks and communities of users on such services grow and what kind of content is posted [1] which led to studies on how and why people actually use Twitter [24]. Analysis of search behaviour showed that while users often express the desire to re-find tweets, this is usually extremely difficult [7]. Twitter content has been used for various purposes: to identify and locate events as they are occurring [4], to replace tags as information sources for URLs [8] and to predict and track natural disasters [16] or the outcome of elections [18].

Two key interactive features of Twitter have been investigated in detail: the @syntax (which allows tweets to reference a particular user) [9] and hashtags. Hashtags have been used for many applications such as tweet and topic recommendation/filtering [2], to augment existing tags on other social media sites [3] and to detect communities of users [23]. Cunah et al. [5] found that hashtag popularity follows a power-law distribution and that they are used to classify tweets, propagate ideas and to promote specific topics. Elsweiler et al. [7] state that "hashtags can be helpful ... [searching and re-finding] become noticeably more difficult for users when they are not present." Hashtags encourage convergence in query terminology, are used to promote content and to find other tweets about a given topic or other users who are interested in the same topic(s) and popular queries are much more likely to contain a hashtag than unpopular ones [17, 10].

The distribution of hashtags in Twitter changes rapidly and as such the most frequent terms in one hour may look very different from those in the next ("churn") [14]. Analysis of how hashtag popularity evolves over time shows sev-

eral types of distribution with many being "bursty" and short-lived [12]. Huang et al. [10] used the standard deviation of hashtag ages (relative to some fixed time point) to measure the spread of hashtag usage over time, asserting that many short-lived hashtags can be explained by the appearance of "micro-memes" - time-sensitive, ad hoc discussions around a topic - and breaking news stories. They showed that a hashtag's temporal spread (as determined by standard deviation) can indicate whether or not it has been triggered by a micro-meme.

Despite the utility of hashtags and the clear advantage in promoting their use, the problem of recommending them has received little attention thus far [13]. An early approach [21] used similarity metrics to compare the vectors of terms to rank tweets in terms of their closeness to the one being written. The method then took the union of hashtags from a number of top-ranked tweets as candidate hashtags to present as suggestions. Three weighting methods for the candidate hashtags were tested with one based on the score of the most similar tweet proving to be most effective. Later work [11] improved on this by using the previous hashtags chosen by the target user to introduce some personalisation, however only raw frequencies of hashtags within the top candidate tweets were used as weights. The authors found that including the user's own hashtag choices improved performance slightly, particularly in cases where the number of top tweets chosen to draw hashtags from was small.

An alternative formulation of the problem instead tried to predict which hashtags will be reused in the future [15, 20]. Yang at al. [20] considered methods for prediction of hashtag adoption and tested the hypothesis that hashtags serve as a tag of content and a symbol of community membership. They found evidence for this and built models to predict whether a user will adopt each potential hashtag within the next 10 days. However, they do not predict which tags will be assigned to a given tweet and therefore their methods are not applicable to hashtag suggestion. In this work we aim to bring together the insights from previous work together with features to exploit the strong temporal trends in the usage of hashtags by users in order to improve recommendations.

3 Recommending Hashtags

We wish to recommend hashtags to Twitter users after they have finished writing a new *target tweet* and therefore have information about the *target user* (i.e. the one who is writing the tweet), the content of the new tweet and the current time. We also have a collection of tweets which were publicly made available prior to the user beginning to write the target tweet - some of which may also have been written by the target user. The content of each tweet can be separated into two groups of terms: *hashtags* (prefixed with a # symbol) and *content terms*. To increase the likelihood of them being useful, suggested hashtags should be: (a) topically appropriate to the content of the target tweet, (b) related to the interests and vocabulary choices of the target user, and (c) temporally relevant.

We have access to a sample of tweets D with a combined vocabulary W, a combined hashtag vocabulary H, written by a set of users U. Each individual

tweet i is composed of a number of content terms from T and hashtags from H (both potentially of length 0). The counts of the wth content term and hth hashtag in the ith tweet are denoted $Cw_{i,w}$ and $Ch_{i,h}$, the author and posting time of the ith tweet are denoted $u(i)$ and $t(i)$. The summation of term counts for term w over all tweets in D is Cw_w. Each user u can also be represented by the set of all of the content terms and hashtags of their tweets (their term and hashtag profiles) using similar notation: $Cw_{u,w}$ being the count of the wth term in the uth user's profile.

Identifying candidate hashtags Given a new candidate tweet j written by user $u(j)$ at time $t(j)$, we first identify similar tweets in D from which to draw candidate hashtags. This can be achieved (with some success) by using the content terms and ranking tweets by their similarity to j [22, 11, 20, 13] using the cosine similarity between vectors of TFIDF-weighted content terms. Any similarity metric could be used, but we take this approach as it reported to be the best performing [22] and calculate the similarity thus: $Sim(i,j) = \frac{\mathbf{i} \cdot \mathbf{j}}{||\mathbf{i}|| \cdot ||\mathbf{j}||}$ Where $\mathbf{i}$ (and $\mathbf{j}$) are vectors of TFIDF weights over all content terms in W such that: $\mathbf{i}_w = Cw_{i,w} \cdot IDF(w)$ and $||\mathbf{i}||$ is the magnitude (or length) of vector $\mathbf{i}$ as computed by the euclidian norm. The Inverse Document Frequency (IDF), defined as $IDF(w) = log\left(\frac{|D|}{\sum^D I\{Cw_{i,w}>0\}}\right)$, reduces the importance of terms which occur too frequently in the collection (in this case, in too many tweets) and therefore have little discriminative power.

Now that we have a similarity score of each tweet i and the target tweet j we can rank these in descending order and, after choosing the top k most similar, we can extract the union of all hashtags within these tweets.

Personalisation To personalise the suggestions we can also look for candidate hashtags which are related to the interests and vocabulary use of u. We could follow the same approach as above but instead of looking for tweets similar to the target tweet, we look for those similar to the target user. While this approach may work well in some cases, many Twitter user have only a small number of prior tweets and as such the amount of term frequency information available will be very small. Instead we can employ a collaborative filtering-like method where we take advantage of Twitter's following mechanism and make the assumption that the hashtags used by those people who u follows are likely correlated with the interests of u. Studies have found strong evidence of homophily between users and those they follow [19] meaning that they share similar topics of interest.

For each user in U we construct a vector of TFIDF values over all hashtags in H such that $\mathbf{u}_h = Ch_{u,h}$ and the new IDF is as follows: $IDF(h) = log\left(\frac{|U|}{\sum^U I\{Ch_{u,h}>0\}}\right)$. Using the same similarity measure as before we identify users who share interests with u and rank these in descending order of similarity. We again choose the top k most similar and extract the union of all hashtags within tweets posted by these users. We will refer to the set of top k tweets as $\hat{D}$, the set of top k users as $\hat{U}$ and the combined set of candidate hashtags as $\hat{H}$.

Weighting candidate hashtags We now address the problem of weighting the candidate hashtags such that their likelihood of being relevant to the new

tweet is maximised. Previous work has investigated methods for doing this [11, 21], proposing the following simple approaches:

1. *OverallPopularity* - frequency over entire collection.
2. *SamplePopularity* - frequency over the sub-set of k tweets most similar to the target.
3. *MaxSimilarity* - the greatest similarity score over the k most similar tweets.

Of these the *MaxSimilarity* method was found to be most effective [21], although some work [11] used the *SamplePopularity* method considering tweets similar to the target tweet and to the target user. Despite the Zipf-like distribution of hashtag popularity in Twitter, the *OverallPopularity* method does not seem to return particularly good rankings.

We would like a method which includes candidate hashtags from both similar tweets and similar users such that the similarity scores from the selection step are included in the score and the influence the two sets of scores have on the final candidate weighting can be varied. Our approach computes the sum of scores from the two sets and linearly combines them into an interpolated sum:

$$score(h) = \lambda \left(\sum_{i=1}^{|\hat{D}|} I\{Ch_{i,h} > 0\} Sim(i,j) \right) + (1-\lambda) \left(\sum_{i=1}^{\hat{U}} I\{Ch_{u,h} > 0\} Sim(i,u) \right)$$

where λ is a free parameter which allows us to vary the relative influence of the scores from similar tweets and similar users.

Considering temporal relevance As discussed in the related work section, analyses of hashtag usage have uncovered evidence of strong temporal patterns [12, 10]. By looking at the timestamps of tweets to which a given hashtag had been assigned Huang et al. [10] identified two categories of hashtags: those used for "organisational" means (used over long periods of time, have high variance); and "conversational" ones (short lifespan, low variance).

Figure 1 shows how two hashtags were used over the first 20 days of January 2014 with the lines representing 2-period moving averages calculated over time bins of 6 hours (4 per day). Although both hashtags are used with approximately the same frequency (266 and 280 instances respectively), they have very different temporal characteristics. The first, #happykanginday, is an example of a conversational tag and refers to the birthday of South Korean celebrity Kangin - which falls on the 17th of January - while #marketing is clearly much more general in nature. Note that the popularity of #happykanginday on the 17th is so great that it exceeds the y-axis, having a count for this bin of 246.

Imagine that we want to re-weight candidate hashtags based on this temporal information. If the target tweet is being written on the 17th and one of the candidate tags happens to be #happykanginday then an increase in the weight of this tag would be sensible. If instead the tweet was being written on another day then it is much less likely to be relevant and therefore should be assigned a negative temporal weight. However, for the #marketing tag the likelihood of

relevance is uniform over time and therefore we would not want to assign it such an extreme temporal weight (neither negative nor positive).

We need a way to measure, in a single point statistic, how spread out the distribution of the ages of previous tweets is. An obvious candidate is the standard deviation, which was used by Huang et al. [10] and is easy to calculate. Another is the entropy of the relative frequencies over evenly-spaced time windows, likely a better measure as it uses more information about the distribution and does not assume that is symmetrical [6]. If we know the frequencies of occurrence of the hashtag over a continuous set of time windows index by i, $Ch(i)$, we can calculate the normalised entropy as follows:

$$\mathbb{H}(X) = -\frac{\sum^X P(x_i) log_b(P(x_i))}{log_b(|X|)}, \;\; where\ P(x_i) = \frac{Ch(i) + 0.01}{\sum^X Ch(i) + 0.01|X|}$$

Note that the probability calculations are smoothed to ensure that the entropy is always finite. In our Twitter data (described later) high-entropy examples are general topical terms or long-running entertainment phenomenon (such as the TV series *The Walking Dead* and the Chicago Bears) are appear with uniform frequency over time. The low-entropy one are instead more specific and usually related to mercurial Internet memes or short-term news events.

To understand how to model the temporal patterns in the hashtags we analysed how the probability of a hashtag being relevant at a given time is related to its age. We split a data set of tweets obtained in January 2014 into two parts with an 80:20 ratio. For each tweet in the 20% part we try to predict which hashtags were actually assigned to it using the method described earlier. For each one we output the top 10 candidate hashtags and the following statistics: entropy, standard deviation, minimum age, maximum age, mean age and median age as well as whether or not each candidate was relevant (i.e. was actually assigned to the target tweet).

The hashtags are separated into two categories - those with entropy less than 0.5 and those with entropy equal to or greater than 0.5 - and then divided into 100 equal-sized bins. For each bin we calculate the probability of relevance as the number of relevant hashtags divided by the total number of hashtags within that bin. Logistic regression models predicting the relevance of a hashtag using each of the measures of location determined that the minimum age has the greatest predictive power. To understand why this is so, and to see how age affects relevance differently for the high- and low-entropy tags, in figure 2 we plot the probability of relevance over the range of minimum ages.

The figure shows that for both sets there is a clear trend of decay in the probability of relevance as minimum age increases, however the rate is much steeper for the low-entropy queries. This confirms the intuition that low-entropy tags relate to short-lived topics or are merely conversational in nature. If the hashtag has a low entropy and the minimum age (i.e. the time since it was last used) is high then it is unlikely that it will be used again and its score in the ranking function should be heavily penalised. However, if it has a high entropy then although there will still be some decay of interest over time, we should not

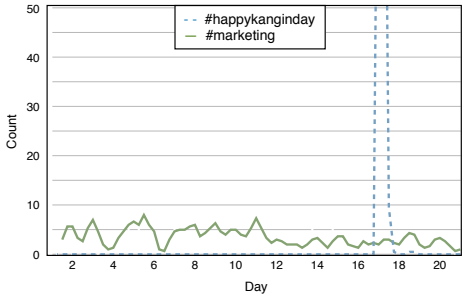

Fig. 1: Trend lies for temporal activity of two hashtags #happykanginday and #marketing.

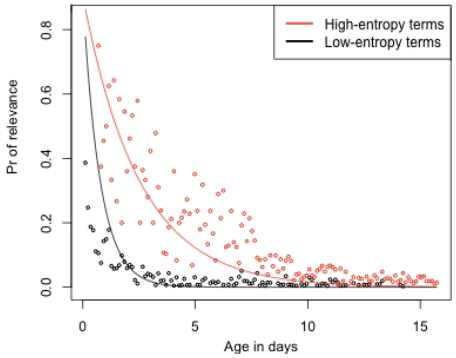

Fig. 2: Pr. of relevance for low- and high-entropy hashtags by minimum age.

penalise it so aggressively. Note that in the case of hashtags which have general relevance (such as the #marketing example) the entropy will be high and the min age will be low meaning that it should receive a positive temporal weighting as the likelihood of being used is always quite high.

Figure 2 also shows lines fitted to each of the two sets calculated using an exponential decay function: $N(t) = e^{-\eta t}$. $N(t)$ is the expected value at time t and η is the rate of decay, which can be learned from the data as in the example in the figure. The output of this function is between 0 and 1 however since the weight should have a positive effect where $N(t)$ is high and a negative effect when it is low we add a constant of 0.5. Multiplying the original similarity-based scores with the output of this function gives an increased weight where $N(t) > 0.5$ (as the output will be between 1 and 1.5) and a decrease when $N(t) < 0.5$. To model the two categories of hashtag we use two different values of η in the function: one for the low- and one for the high-entropy hashtags (η_l and η_h). When weighting the candidate hashtags we calculate their entropies over previous tweets in the data set and if the entropy is < 0.5 we use η_l, otherwise we use η_h.

We have devised sensible functions for identifying candidate hashtags and then ranking those tags based on their similarity to the target tweet and the target author's expanded interest profile weighted by their temporal relevance. We now detail how we collected a suitable data set for testing our methods and describe the results achieved by them. We conclude by discussing the results and commenting on potential avenues for future work.

4 Experiments

Data set A sample of 5,016 Twitter users was collected from the Twitter API [3] by first downloading tweets from the Twitter streaming API - which we assume

[3] Twitter REST API version 1.1:
https://dev.twitter.com/docs/api/1.1

to be random - and then listing all users who posted any sampled tweets. The account details of these users were obtained and the list filtered by removing: verified users (usually celebrities or news organisations), those with unusually high numbers of friends (spammers), those who had joined within the past week and had more than 1000 tweets (spammers) and users with no followers (potential spammers), resulting in a list of 2,576 users. From this list we randomly sampled 300 users and collected all tweets written by users they follow - 379,919 - between the 1st and 20th of January 2014, yielding 3,303,016 tweets, appearing a total of 3,528,564 times (a single tweet can appear on more than one user's timeline).

Since this data will be used to suggest hashtags we restricted our dataset to those tweets that have at least 1 hashtag and, in keeping with literature, we do not use retweets in our data set as our similarity search would return an identical retweet, clearly distorting the results. The final data set consisted of 333,784 tweets (10.1% of the original tweets) from 23,476 unique authors with a hashtag vocabulary of 51,899 unique tags.

Models and baselines Here the models used for hashtag suggestion are briefly described. An ***** indicates that the method was newly developed for this work.

1. *TweetMax* - hashtags drawn from similar tweets only, weights each candidate by max. similarity score. Best-performing method of Zangerle et al. [21].
2. *UserMean** - uses only similar users to draw hashtags from, weights each candidate by the mean similarity score over all similar users.
3. *CombCount* - uses union of hashtags from similar users and tweets, weighted by total count of hashtag over all similar tweets and users. Slightly more sophisticated version of best-performing method used by Kywe et al. [11].
4. *CombInt** - uses union of hashtags from similar users and tweets, candidates weighted by linearly interpolated scores from similar tweets and users.
5. *TemporalTweetMax** - hashtags drawn from similar tweets only, weighted by the maximum similarity score multiplied by temporal relevance score.
6. *TemporalCombInt** - uses union of hashtags from similar users and tweets, candidates weighted by linearly-interpolated scores from similar tweets and users multiplied by temporal relevance score.

Splitting the data set and optimising parameters The data was sorted by time in ascending order and split into two sections in the ratio 80:20. Although it is normal to use split-fold testing with multiple splits, this is not possible as we are interested in the specific temporal aspects of the data and therefore cannot test on data generated before the training data. The last 20% of the largest split was used to optimise any model parameter values: the η values for the low- and high-entropy exponential decay functions (η_l and η_h respectively) and the λ parameter controlling the linear interpolation of hashtag ranking scores from similar tweets and users. All parameters were optimised via an exhaustive search resulting in the following optimised values: $\eta_l = 1.2$, $\eta_h = 0.6$ and $\lambda = 0.4$.

The smaller split of the data (66,757 tweets) was used to test the models. For each model we wish to predict which hashtags were actually chosen by the

author. To do so all data which existed prior to each tweet in time was used to train the similarity models and learn the entropies and minimum ages of the hashtags. The content terms of each test tweet as well as the user ID of the author were then input into each model which returned a ranked list of 5 candidate hashtags. These suggestions were then compared with the hashtags actually assigned to the tweet (which we take to be relevant, with all other hashtags being non-relevant). The standard IR metrics of precision and recall were calculated for ranks 1 through 5, where precision is the number of relevant returned over the number returned and recall is the number of relevant returned over the total number relevant. Note that often there is only one relevant tag.

4.1 Results

Method	**P@1**	**P@5**	**R@5**
TweetMax	0.256	0.086	0.311
CombinedCount	0.153	0.069	0.267
UserMean	0.238 (-8.2%)	0.089 (3.5%)	0.348* (11.9%)
CombInt	0.292* (14.1%)	0.106* (23.3%)	0.416* (33.8%)
TemporalTweetMax	0.310* (21.9%)	0.102* (18.6%)	0.359* (15.4%)
TemporalCombInt	0.314* (22.7%)	0.109* (26.7%)	0.429* (37.9%)

Table 1: Results table for all methods compared. * indicates a statistically significant improvement over *TweetMax*, 2-sample t at 95% confidence.

Table 1 summarises the performance of the 6 hashtag suggestion models. P@1 indicates a model's ability to return a relevant tag at position one in the ranking and P@5 indicates the ability to return at least one relevant tag within the top 5 candidates. R@5 describes, on average, what ratio of all relevant tag the model is able to suggest. We can see that the additions made to the basic models in this work served to increase both the accuracy and coverage of the suggested hashtags. Figures in parentheses indicate the percentage difference of each model relative to the most competitive baseline (*TweetMax*).

The worst-performing model is *CombinedCount*, probably because of its lack of sophisticated weighting, relying as it does on combined frequencies of each candidate hashtag over the similar tweets and similar users. In terms of P@1, *TweetMax* is able to achieve better performance than *UserMean*, which is expected as it relies on information about the tweet itself and not just about the user, however surprisingly *UserMean* is able to out-perform it over the next 4 rank positions. Linearly interpolating candidate hashtags and scores (*CombInt*) performs significantly better than either of the single components on their own.

The addition of the temporal weighting seems to have a very positive impact on suggestion performance as the two models which include this weighting returned better performance than the equivalent models without it. The most sophisticated method (*TemporalCombInt*) yields the best performance figures for all metrics and does particularly well in terms of recall, being able to predict 42.9% of all hashtags correctly within the top 5 rank positions.

Changes in rank position To examine the performance improvements resulting from including temporal information in the ranking we look in more detail

at the relative performance between *TweetMax* and *TemporalTweetMax* (which are otherwise identical). The variation in performance can be better understood by considering the difference in the ranks of the relevant hashtags. Figure 3 shows the distribution of the difference in the ranking of relevant hashtag for single-hashtag tweets. Red bars show the number of tweets where the ranking was improved by using temporal information, while the green ones indicate a deterioration and "other" refers to all rank changes greater than 5. The chart shows - as one would expect from table 1 - that the temporal information results in a better ranking far more often than a worse one (74% of cases are better). However it also shows that in the majority of negative cases, the ranking is only deteriorated by a couple of rank positions - 48.1% of deteriorated rankings are only by one or two positions. On the other hand, in 37.8% of cases where the temporal information has a positive effect the improvement is dramatic (i.e. an improvement of more than 5 rank positions).

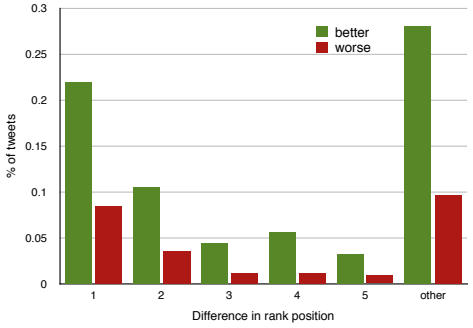

Fig. 3: Δ in rank position of relevant hashtag between the rankings from *TweetMax* and *TemporalTweetMax*.

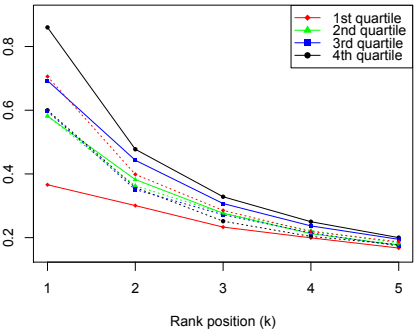

Fig. 4: Average precision by amount of historical data. Solid lines = *TemporalTweetMax*, dotted = *TweetMax*.

Do we have enough data? Given that the temporal part of our models is based on historical information about each hashtag and our data set represents only a small sample of all tweets posted on Twitter between the crawling dates, we now investigate how the quantity of information available about a hashtag affects performance. We again compare the performance of *TweetMax* and *TemporalTweetMax* and only consider instances where the target tweet has a single hashtag. However, here we sample to ensure that both models were able to return the single relevant hashtag somewhere within the first 20 rank positions.

Figure 4 shows how the performance of the two models changes (over the first 5 rank positions) as we vary the amount of historical data available in the training set about the relevant hashtag (in 4 equal quartiles). The *TemporalTweetMax* (solid lines) returns poorer performance when we have less information about the relevant hashtag but much better performance when we have more information.

This pattern is, however, not evident for the *TweetMax* model (dotted lines) which returns similar performance regardless of the amount of data available about the relevant hashtag. This indicates that the performance improvements given by the inclusion of temporal information could be even greater if we had more training data to base our entropy and minimum age statistics on.

5 Conclusions and future work

In this paper we proposed new methods for hashtag suggestion which could lead to more frequent, accurate and useful assignment of hashtags to tweets. We began by identifying the most effective measures for basic hashtag recommendation in the literature and proceeded to investigate ways to improve performance by including more information in the model and using existing information in a more intelligent fashion. We analysed temporal patterns in hashtags from the perspective of relevance and identified trends which we hypothesised could be exploited to make suggestions more temporally relevant. By analysing the ages of tweets containing candidate hashtags, relative to when a new tweet was posted, we developed a method to re-weight candidate scores by their temporal relevance.

Using a sample of real-world Twitter data from January 2014, we tested the performance of our novel methods against two competitive baselines from the literature, demonstrating significant performance improvements, although these were restricted by the amount of training data available and therefore have the potential to be better still. We showed that these improvements came from both the temporal information and the more sophisticated use of user interest data, augmented by a collaborative filtering approach. Further analysis showed that the improvements in rank position of relevant hashtags brought by including temporal information in rankings are often quite large. Perhaps more importantly, in the few instances where the temporal weighting is not successful, it rarely results in a large detrimental change to the ranking.

In future work we would like to first investigate more nuanced ways of identifying similar tweets and similar users, perhaps using some form of dimensionality reduction to mitigate the issue of vocabulary mismatch. Similar approaches to addressing this problem could also consider term expansion of the initial list of candidate hashtags. We also intend to investigate how the temporal information could be more subtly utilised in the models. Instead of grouping hashtags into two categories (low- and high-entropy) with tuned η values, it may be possible to learn a smooth mapping between a hashtag's entropy and the appropriate value of η in the temporal weighting function.

References

1. Danah Boyd, Scott Golder, and Gilad Lotan. Tweet, tweet, retweet: Conversational aspects of retweeting on twitter. In *HICSS*, pages 1–10, 2010.
2. Kailong Chen, Tianqi Chen, Guoqing Zheng, Ou Jin, Enpeng Yao, and Yong Yu. Collaborative personalized tweet recommendation. In *SIGIR*, pages 661–670, 2012.

3. Denzil Correa and Ashish Sureka. Mining tweets for tag recommendation on social media. In *SMUC*, SMUC '11, pages 69–76, New York, NY, USA, 2011. ACM.
4. Anqi Cui, Min Zhang, Yiqun Liu, Shaoping Ma, and Kuo Zhang. Discover breaking events with popular hashtags in twitter. In *CIKM*, pages 1794–1798, 2012.
5. Evandro Cunha, Gabriel Magno, et al. Analyzing the dynamic evolution of hashtags on twitter: A language-based approach. In *LSM*, pages 58–65, 2011.
6. Nader Ebrahimi, Esfandiar Maasoumi, and Ehsan S Soofi. Ordering univariate distributions by entropy and variance. *J. of Econometrics*, 90(2):317–336, 1999.
7. D. Elsweiler and M. Harvey. Engaging and maintaining a sense of being informed: Understanding the tasks motivating twitter search. *JASIST*, 2014.
8. M. Harvey, M. Carman, and D. Elsweiler. Comparing tweets and tags for urls. In *ECIR*, pages 73–84, 2012.
9. Courtenay Honeycutt and Susan C. Herring. Beyond microblogging: Conversation and collaboration via twitter. In *HICSS*, pages 1–10, 2009.
10. Jeff Huang, Katherine M. Thornton, and Efthimis N. Efthimiadis. Conversational tagging in twitter. In *HT*, pages 173–178, 2010.
11. Su Mon Kywe, Tuan-Anh Hoang, Ee-Peng Lim, and Feida Zhu. On recommending hashtags in twitter networks. In *SocInfo*, pages 337–350, 2012.
12. Janette Lehmann, Bruno Gonçalves, José J. Ramasco, and Ciro Cattuto. Dynamical classes of collective attention in twitter. In *WWW*, pages 251–260, 2012.
13. T. Li and Y. Z. Yu Wu. Twitter hashtag prediction algorithm. In *WORLDCOMP*, 2011.
14. Jimmy Lin and Gilad Mishne. A study of "churn" in tweets and real-time search queries. In *ICWSM*, 2012.
15. Zongyang Ma, Aixin Sun, and Gao Cong. Will this #hashtag be popular tomorrow? In *SIGIR*, pages 1173–1174, 2012.
16. Takeshi Sakaki, Makoto Okazaki, and Yutaka Matsuo. Earthquake shakes twitter users: Real-time event detection by social sensors. In *WWW*, pages 851–860, 2010.
17. Jaime Teevan, Daniel Ramage, and Merredith Ringel Morris. #twittersearch: A comparison of microblog search and web search. In *WSDM*, pages 35–44, 2011.
18. A. Tumasjan, T. O. Sprenger, P. G. Sandner, and I. M. Welpe. Predicting elections with twitter: What 140 characters reveal about political sentiment. In *ICWSM*, pages 178–185, 2010.
19. Jianshu Weng, Ee-Peng Lim, Jing Jiang, and Qi He. Twitterrank: Finding topic-sensitive influential twitterers. In *WSDM*, pages 261–270, 2010.
20. Lei Yang, Tao Sun, Ming Zhang, and Qiaozhu Mei. We know what @you #tag: Does the dual role affect hashtag adoption? In *WWW*, pages 261–270, 2012.
21. E. Zangerle and W. Gassler. Recommending #-tags in twitter. In *CEUR Workshop*, 2011.
22. Eva Zangerle, Wolfgang Gassler, and Günther Specht. On the impact of text similarity functions on hashtag recommendations in microblogging environments. *Social Network Analysis and Mining*, 3(4):889–898, 2013.
23. Yang Zhang, Yao Wu, and Qing Yang. Community discovery in twitter based on user interests. *Journal of Computational Information*, 3:991–1000, 2012.
24. Dejin Zhao and Mary Beth Rosson. How and why people twitter: The role microblogging plays in informal comms at work. In *GROUP*, pages 243–252, 2009.