Tweet Timeline Generation with Determinantal Point Processes

Jin-ge Yao[1,2], **Feifan Fan**[1], **Wayne Xin Zhao**[3], **Xiaojun Wan**[1,2], **Edward Chang**[4], **Jianguo Xiao**[1,2]
[1]Institute of Computer Science and Technology, Peking University, Beijing 100871,China
[2]Key Laboratory of Computational Linguistic (Peking University), MOE, China
[3]School of Information, Renmin University of China
[4]HTC Research, Beijing, China
{yaojinge,fanff,wanxiaojun,xiaojianguo}@pku.edu.cn, batmanfly@gmail.com, edward_chang@htc.com

Abstract

The task of tweet timeline generation (TTG) aims at selecting a small set of representative tweets to generate a meaningful timeline and providing enough coverage for a given topical query. This paper presents an approach based on determinantal point processes (DPPs) by jointly modeling the topical relevance of each selected tweet and overall selectional diversity. Aiming at better treatment for balancing relevance and diversity, we introduce two novel strategies, namely spectral rescaling and topical prior. Extensive experiments on the public TREC 2014 dataset demonstrate that our proposed DPP model along with the two strategies can achieve fairly competitive results against the state-of-the-art TTG systems.

Introduction

The microblogging service has become one of the most popular social networking platforms in recent years. When users search a query in a microblogging website such as Twitter, an archive of tweets would be retrieved related to the query topic. In many cases, users may also want to collect the recent progress of one particular event and turn to Twitter search for more information from social users. However, search results on tweets are not very informative and lack of meaningful organization. Results typically include a large amount of duplicates or near-duplicates tweets. It would be helpful if the search system produced a summary timeline about the topic.

In TREC 2014 Microblog track, the organizer introduced a novel pilot task named Tweet Timeline Generation (TTG) (Lin and Efron 2014). The main requirement of the TTG task is to produce a summary that captures relevant information for a given query Q. Highly related to topic detection and tracking (TDT) and traditional multi-document summarization, the essence of TTG task requires that a system should be able to jointly consider *topical relevance* (obtaining relevant information) and *diversity* (reducing the redundant tweets) in tweet selection.

Jointly modeling relevance and diversity has been an important and general issue in multiple fields, which has attracted much research interest. In this work we introduce a novel approach for the TTG task to jointly model individual topical relevance and overall diversity with a recently proposed probabilistic model called determinantal point processes (DPPs)(Kulesza and Taskar 2012). A determinantal point process is a probabilistic measure defined on item sets. A typical DPP model can independently characterize quality scores of each item and select items as diverse as possible. This property makes it suitable for the timeline generation task to select both representative and diverse tweets. However, tweets are short, noisy and even contain slangs or spelling errors, which makes the TTG task challenging. For example, many less relevant tweets are likely to obtain considerable relevance scores due to the fact that the query terms are unconsciously included in the tweets. Thus, a simple application of DPP model will not lead to good performance for TTG task.

There is an inevitable need to better balance relevance of tweets and selectional diversity in the context of TTG task. To tackle this issue, we proposed two novel strategies. Firstly, utilizing spectral properties of the parameter matrix of a DPP model, the magnitudes of relevance scores can be adaptively rescaled via an automatic tuning process, which is called spectral rescaling in this paper. Such a rescaling method is able to tune the effect of different relevance scores, according to the scale and distribution of the similarity values. Secondly, as the DPP model essentially defines a probability distribution, we further propose to impose a topical relevance prior to enhance the selection of relevant and topically coherent tweets, which improves the relevance measurement based on simple surface lexical similarity. Extensive experiments on the TREC 2014 dataset demonstrate the effectiveness of the proposed DPP model along with the two strategies.

The Task: Tweet Timeline Generation

The task of tweet timeline generation aims at selecting a small amount of representative tweets to generate a timeline and providing enough coverage for a given topical query.

Formally, we give a definition for the TTG task as follows:

Input: Given a topic query Q from users, we obtain a collection of tweets $\mathcal{C} = \{T_1, T_2, \cdots, T_N\}$ related to the query by traditional retrieval model, where T_i is a tweet and N is the number of retrieved tweets.

Output: A summarized tweet timeline which consists of relevant and non-redundant, chronologically ordered tweets,

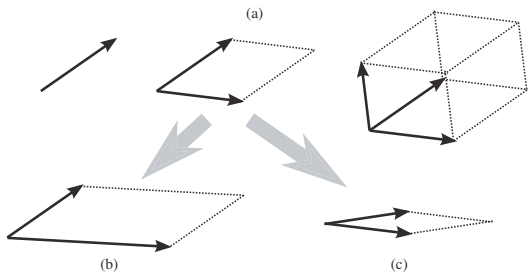

Figure 1: (a) The DPP probability of a set Y depends on the volume spanned by vectors $q_i\phi_i$ for $i \in Y$ (b) As length increases, so does volume. (c) As similarity increases, volume decreases.

i.e. $S^{(Q)} = \{T_1^{(Q)}, T_2^{(Q)}, \cdots, T_K^{(Q)}\}$, where $T_i^{(Q)}$ is a relevant tweet from $\mathcal{C}$ for query Q, and K is the number of tweets in the timeline.

Background: Determinantal Point Processes

We first review some background knowledge on the determinantal point processes (DPPs). More details can be found in the comprehensive survey (Kulesza and Taskar 2012) covering this topic.

Determinantal point processes (DPPs) are distributions over subsets that jointly prefer quality of each item and diversity of the whole subset. Formally, a DPP is a probability measure defined on all possible subsets of a group of items $\mathcal{Y} = \{1, 2, \ldots, N\}$. For every $Y \subseteq \mathcal{Y}$ we have:

$$\mathcal{P}(Y) = \frac{\det(L_Y)}{\det(L+I)}$$

where L is a positive semidefinite matrix typically called an *L-ensemble*. $L_Y \equiv [L_{ij}]_{i,j \in Y}$ denotes the restriction of L to the entries indexed by elements of Y, and $\det(L_\emptyset) = 1$. The term $\det(L+I)$ is the normalization constant which has a succinct closed-form and easy to compute. We can define the entries of L as follows:

$$L_{ij} = q_i \phi_i^\top \phi_j q_j \tag{1}$$

where we can think of $q_i \in \mathbb{R}^+$ as the *quality* of an item i and $\phi_i \in \mathbb{R}^n$ with $\|\phi_i\|_2 = 1$ denotes a normalised feature vector such that $\phi_i^\top \phi_j \in [-1, 1]$ measures *similarity* between item i and item j. This simple definition gives rise to a distribution that places most of its weight on sets that are both high quality and diverse. This is intuitive in a geometric sense since determinants are closely related to volumes; in particular, $\det(L_Y)$ is proportional to the volume spanned by the vectors $q_i\phi_i$ for $i \in Y$. Thus, item sets with both high-quality and diverse items will have the highest probability (Figure 1).

A variety of probabilistic inference operations can be performed efficiently, including sampling, marginalization, and conditioning. However, the maximum a posteriori (MAP) problem $\text{argmax}_Y \log\det(L_Y)$ that finds the item set with the largest probability is NP-hard(Gillenwater, Kulesza, and Taskar 2012b). Fortunately, efficient approximate solutions seem to be acceptable in this study. This is partly due to the submodularity [1] of the $\log\det(L_Y)$ function. Many greedy algorithms for submodular maximization have been theoretically justified to have worst-case guarantees(Vondrák, Chekuri, and Zenklusen 2011).

Our Proposed Approach

The property of joint modeling item quality and overall diversity makes DPP-based models intuitively appealing for tweet timeline generation, where it requires both good relevance from each selected tweet and diversity of all selected tweets to reduce information redundancy and improve information coverage. In this section, we first propose a general DPP-based model for tweet timeline generation. Then we introduce two improvement strategies to adapt DPP models to the TTG task.

A DPP Model for TTG

Tweet Relevance For the TTG task, we believe that the quality of each item, i.e., each candidate tweet, corresponds to its relevance to a given query. We can set q_i in Equation 1 to be the relevance score for tweet i with respect to the given query. For the relevance score used in this work, we utilize a state-of-the-art retrieval model which was top-ranked in TREC 2014 ad hoc search task. The system mainly contains three components, respectively responsible for candidate generation that includes query expansion, feature generation and learning-to-rank re-ranking model. We point interested readers to (Lv et al. 2014) for implementation details. Other relevance measurement methods can be applied equally.

Timeline Diversity As described in the previous section, diversity of an item set is indirectly characterized with a similarity measure in a DPP model. Following the same TREC 2014 participant system (Lv et al. 2014), we use a term frequency vector to represent each candidate tweet i, corresponding to ϕ_i in Equation 1. Based on such a vector representation, we simply use the ordinary cosine similarity to measure the pairwise similarity. Note that diversity itself encourages coverage, novelty and less redundancy. Therefore, we do not consider explicitly these factors in this paper, which are reflected via the diversity measurement.

To summarize, we define our DPP model for TTG using an L-ensemble representation, where each element of L is defined as:

$$L_{ij} = q_i \phi_i^\top \phi_j q_j = q_i \times \cos(i, j) \times q_j \tag{2}$$

where $\cos(i, j)$ denotes the cosine similarity between tweet i and tweet j, using TF vectors.

Once the DPP model has been set up, what we need to do is to pick up the item set that maximizes the probability, i.e., performing maximum a posteriori (MAP) inference. Since MAP inference for DPP is NP-hard, we use a greedy

[1] A set function $F : 2^U \to \mathbb{R}$ defined over subsets of a universe set U is said to be *submodular* iff it satisfies the *diminishing returns* property: $\forall S \subseteq T \subseteq U \setminus u$, we have $F(S \cup \{u\}) - F(S) \geq F(T \cup \{u\}) - F(T)$.

algorithm modified with size budget for simplicity and efficiency in this study. Adapting more advanced and more accurate MAP inference strategies to the TTG task is left for future work. The greedy algorithm is described in Algorithm 1. Our algorithm does not require the number of tweets K in the timeline generation as the input, and the score for a set of tweets S is defined as the unnormalized log probability given L: $\text{score}_L(S) = \log\det L(S)$.

Algorithm 1 A greedy algorithm for DPP model

Input:
Candidate tweet collection $\mathcal{Y} = \{1, 2, \ldots, N\}$,
L-ensemble matrix $L \in \mathbb{R}^{N \times N}$
Output:
Summarized tweet timeline S

$S \leftarrow \emptyset, i \leftarrow 0$
repeat
  $i \leftarrow i + 1$
  $s_i = \text{argmax}_{s \in \mathcal{Y}} \text{score}_L(S \cup \{s\})$
  $S \leftarrow S \cup \{s_i\}$
until $\text{score}_L(S \cup \{s_i\}) - \text{score}_L(S) < 0$
return S

Balancing Relevance and Diversity

The quality scores defined in earlier DPP-based applications, such as document summarization (Kulesza and Taskar 2011b) and item recommendation (Gillenwater et al. 2014), are typically good enough for the system to produce ideal item sets. Unfortunately this is not the case in tweet timeline generation, where the items are less informative since tweets are usually short and noisy. More care is needed to balance the quality of each tweet and overall diversity of selection.

More specifically, the magnitude of the quality score has important effect on the performance of the DPP-based system for TTG. On one hand, if the scale is too small, the system will tend to select more diverse items and result in more irrelevant tweets for the given query [2]; On the other hand, if the quality scale is too large, the quality scores will become the first priority, and the system tends to select top-ranked tweets only, which cannot achieve good diversity. Therefore we need principled ways to balance the relevance and diversity during the selection process.

Automatic Spectral Rescaling

Diversity scores are derived from inner products of normalized feature vectors, and in this work those inner products take values within $[0, 1]$ since all elements are nonnegative. In order to control the relevance scale, a scaling parameter is needed to rescale each quality score, or equivalently a scaling parameter for rescale the L-ensemble matrix in the DPP model. By re-examining the Equation 1, i.e. the definition of L-ensemble elements, we can see that scaling β on each relevance score will directly scale the matrix L by a factor of β^2. The scale of a matrix can be also reflected on its spectral properties, e.g. eigenvalues.

On the other hand, there exists an intriguing property for a DPP L-ensemble on the expected number of selected items, denoted as K:

$$K \equiv \mathbb{E}[|Y|] = \sum_i \frac{\lambda_i}{\lambda_i + 1}, \tag{3}$$

where Y is the selected item set, λ_is are eigenvalues of the L-ensemble matrix L. Thereby the problem of quality rescaling now becomes how to tune the scale of eigenvalues to identify a proper size for the selected item set. Inspired by a commonly used strategy in principal component analysis, we empirically set a 90% threshold on accumulated eigenvalues to determine K, i.e. how many components or in our context how many items, to keep. In our experiments we also find that the estimated numbers of item sets and the gold standard sizes share strong correlations.

Having known the estimated number for tweet selection, what remains is to rescale the eigenvalues to match that estimated set size. To achieve this, we propose a forward-backward scaling scheme, as described in Algorithm 2.

Algorithm 2 Forward-backward rescaling for eigenvalues

Input:
L-ensemble matrix L
Expected number of elements K
Magnitude parameter $\alpha \in (0, 1)$
Output:
Proper scaling parameter β

$\beta \leftarrow 1$, $\Lambda \leftarrow$eigenvalues(L)
repeat
  $\Lambda \leftarrow \beta\Lambda$
  est_size$= \sum_i \frac{\lambda_i}{\lambda_i + 1}$
  if est_size $< K$ **then**
    $\beta \leftarrow (1 + \alpha)\beta$
  else if est_size $> K$ **then**
    $\beta \leftarrow \alpha\beta$
  end if
until est_size$= K$
return β

The procedure is similar to binary search to some extent, and a note is that scalings from two directions are not symmetric [3]. We prefer item sets with smaller estimated sizes, since we adopt the greedy approximate inference, which empirically tends to select more items than expected. Once β is obtained, we use it to rescale the L-ensemble, i.e. set $L \leftarrow \beta L$. Greedy inference (Algorithm 1) remains the same. Note that the βs are different for each query, since the spectral properties of corresponding L-ensembles vary.

A Topical Prior for Enhancing Global Coherence

As mentioned in the introduction section, one of the major challenges of TTG is that tweets are usually too short to be

[2] Consider the geometric intuition in Figure 1, if the magnitude of all vectors shrinkages, then the degree of angle will become more decisive for maximizing the volume.

[3] For any constant T and scaling factor $\alpha < 1$, we always have $\alpha T \times (1 + \alpha)$ less than the original T.

sufficiently informative. The relevance scores based on lexical similarity with query terms are not reliable to accurately capturing query relevance. Many less relevant tweets containing partial literal overlaps with query terms are likely to be ranked higher than those topically coherent tweets with less overlaps. For example, given the query *Obama healthcare law unconstitutional*, our implementation of the retrieval system in (Lv et al. 2014) assigns high ranks to less relevant tweets discussing *Obama honors science Olympiad gold medal winners*, as they contain the important entity term *Obama*.

This phenomenon can incorrectly guide DPP-based models (or any other methods that prefer diversity, as later we will see in the experiments): irrelevant or topically incoherent tweets with substantial relevance scores are selected to encourage diversity. Therefore, if we want to model the TTG task using DPPs, we need to take special care on topical coherence between candidate tweets and the query.

To be more specific, in the previous settings we assume that DPPs are selecting items from candidates that are mostly coherent to the query topic. The aforementioned DPP-based models are in fact only modeling $\mathcal{P}(Y|C_{Y,t} = 1)$ instead of the joint probability $\mathcal{P}(Y, C_{Y,t} = 1)$, where $C_{Y,t}$ is a binary variable that takes value 1 if and only if tweet i is highly relevant and coherent with the given query topic t for $\forall i \in Y$. Another view is that previously we indeed maximize the joint probability $\mathcal{P}(Y, C_{Y,t} = 1) \propto \mathcal{P}(Y|C_{Y,t} = 1)P(C_{Y,t} = 1)$, but with $P(C_{Y,t} = 1)$ naively set to be a uniform prior. Assuming that topical coherence among different tweets are independent, we have

$$P(C_{Y,t} = 1) = \prod_{i \in Y} P(C_{i,t} = 1). \quad (4)$$

This indicates that different candidate is will have the same prior probability to become coherent to the query topic t, regardless of the specific content of each tweet.

Therefore, a natural improvement over the original DPP model is to introduce a nonuniform prior $P(C_{i,t} = 1)$ for each tweet i. Tweets with coherent content to the query topic should be assigned with higher prior probability mass.

In this work we adopt a simple yet effective way to set such prior. Here we change the notation to explicitly emphasize that we are inferring topical coherence based on the content:

$$P(C_{i,t} = 1) = P(C_t = 1|i) = \frac{P(C_t = 1, i)}{P(C_t = 1, i) + P(C_t = 0, i)}$$

For both $P(C_t = 1, i)$ and $P(C_t = 0, i)$ we impose a naive Bayes assumption on words contained in each specific tweet i:

$$P(C_t = k) = \frac{P(C_t = k) \prod_{w \in i} P(w|C_t = k)}{\sum_{k' \in \{0,1\}} P(C_t = k') \prod_{w \in i} P(w|C_t = k')}$$

To estimate the parameters appeared in the RHS, theoretically we need to have tweets that are labeled with $C_t = 1$ or $C_t = 0$. Since we do not have such information, we assume that tweets with the most overlap (typically containing all words in the query topic t) are topically coherent, i.e. $C_t = 1$, and tweets without any overlap with those coherent tweets are set to be $C_t = 0$. This is a very strong assumption and we leave to future work about how to relax it. Parameters are estimated on these tweets using empirical counts with Laplace smoothing [4].

The above estimation method can be explained with query expansion. Words that co-occurs frequently with all query words are considered as coherent with the query topic, called *extended words*. In this way, tweets with extended words but not containing all query terms can still be assigned with large coherence prior. On the other hand, words appears frequently with a subset of the query words but without extended words will no longer be assigned with high probability. The aforementioned example "*Obama honors medal winners* " will be considered to be topically incoherent since all words except *Obama* do not co-occur with all the other query words.

The incorporation of the topical prior given the topical query t changes the definition of score in the greedy inference procedure (Algorithm 1) to be $\text{score}_L(S) = \log\det L(S) + \log P(C_{S,t} = 1)$. With the independence assumption (4), we can factorize probabilistic scores over each item. In each iteration, a score is still assigned to each item but with the modification of the topical-prior weighted increase, resulting in a tiny change of Line 4 in Algorithm 1 into

$$s_i = \operatorname*{argmax}_s \log\det L(S \cup \{s\}) + \log P(C_{s,t} = 1).$$

Experiments

Data Preparation

We evaluate the proposed TTG systems over 55 official topics in the TREC 2014 Microblog track (Lin and Efron 2014). Topics in TREC 2011-2012 are used as the development set for parameter tuning.

For each topic query, we use the aforementioned retrieval system to obtain candidate tweets. We use a threshold for relevance scores to filter out most irrelevant tweets on TREC 2011-2012 data.

Evaluation Metrics

Following the TREC 2014 task on tweet timeline generation, our evaluation metrics mainly focus on the clustering performance as timeline quality. There exists groups of tweets called semantic clusters that are annotated by human assessors, representing an equivalence class of tweets that contain the same information (i.e., retrieving more than one cluster member is redundant). The ideal TTG run would retrieve one and only one tweet from each cluster of each topic. TTG results will be evaluated by two different versions of the F_1 metric, i.e., an unweighted version and a weighted version, which are used in TREC 2014 Microblog Track (Lin and Efron 2014). F_1 metric is combined by cluster precision and

[4] Stopwords and tokens start with '@' or '#' are not considered. We also tried to utilize all tweets by using an EM estimation on those unlabeled tweets but we did not observe substantial difference in experimental results.

cluster recall. We first introduce the unweighted version as follows.

- **Cluster precision (unweighted).** Of tweets returned by the system, how much proportion of distinct semantic clusters are represented.
- **Cluster recall (unweighted).** Of the semantic clusters discovered by the assessor, how much proportion has been represented in the system's output.

For unweighted version, the system does not get "credit" for retrieving multiple tweets from the same semantic cluster. Different from unweighted F_1, the weighted F_1 (denoted as F_1^w) attempts to account for the fact that some semantic clusters are intuitively more important than others. Each cluster will be weighted by relevance grade: minimally-relevant tweets get a weight of one and highly-relevant tweets get a weight of two. These weights are then factored into the precision and recall computations. The F_1^w score is the main evaluation metric for TTG task in TREC 2014.

Methods to Compare

We consider the following methods as comparisons in our experiments.

- **TTGPKUICST2**: Hierarchical clustering algorithm based on adaptive relevance estimation and Euclidean distance, as used in (Lv et al. 2014), which achieved the best performance in TREC 2014 Microblog Track.
- **EM50**: kNN clustering approach applied in (Walid, Wei, and Tarek 2014), using a modified Jaccard coefficient (i.e. EM) and used top K retrieved results as candidates for clustering, which won the second place in TREC 2014 Microblog Track.
- **hltcoeTTG1**: The novel detection approach proposed in (Xu, McNamee, and Oard 2014). Unlike clustering methods, they framed the problem of tweet timeline generation as a sequential binary decision task. Therefore, they proposed a binary classifier to determine whether a coming tweet is novel and then compose the novel tweets as the summarized tweet timeline, which ranked third in TREC 2014 Microblog Track.
- **DivRank**: We implement the DivRank algorithm (Mei, Guo, and Radev 2010) that is also well known for jointly modeling relevance and diversity in a much simpler way – a linear combination of two terms.
- **StarClustering**: In star clustering (Wang and Zhai 2007), each cluster is star-shaped, and the center document is treated as the most representative tweet for each cluster. The final timeline is the collection of such representative tweets. This can also be regarded as a model that balances individual relevance and diversity.
- **DPP**: This is the vanilla version of our proposed DPP-based model utilizing relevance scores and cosine similarity.
- **DPP+SR**: This is our DPP-based model with the spectral rescaling strategy applied.
- **DPP+TP**: The DPP model integrated with topical prior.
- **DPP+SR+TP**: The DPP model involving both spectral rescaling and topical prior.

Note that both in re-implemented systems and the proposed system, we obtain the top-ranked 300 tweets from the ranked list achieved by the retrieval models as the candidates for TTG process. For spectral rescaling, we simply set $\alpha = 0.5$.

Results and Discussion

Table 1 shows the TTG performance of different methods.

We can observe that previous methods modeling both relevance and diversity have achieved large recall along with unsatisfactory precision, e.g., *DivRank* and *StarClustering*. Such methods have a strong tendency to select more tweets to enhance diversity, while ignores the fact that relevance scores themselves from the retrieval systems are not always reliable. They are quite sensitive to the noisiness from relevant tweet retrieval.

With the proposed strategy for balancing relevance and diversity via the proposed spectral rescaling method for the DPP model, **DPP+SP** significantly improves the precision score over **DPP**. By comparing **DPP** with **DPP+TP**, we can observe that the incorporation of the topical relevance prior also gives better precision performance. The two strategies together achieve the optimal performance for DPP based models (**DPP+SR+TP**), in terms of the weighted F_1 metric.

We present an illustrative example in the topic of "MB225" to show the effect of the topical prior. Given the topical query *Barbara Walters, chicken pox*, we observed that several tweets about *Barbara Walters's clash with Elisabeth Hasselbeck* have also achieved high relevance scores even if the two events are not related. The topical prior derived from retrieved tweets can set low probability to the words *Elisabeth Hasselbeck* for coherent tweets, which makes tweets with these words unlikely to be selected by the DPP system in the final timeline.

Related Work

Topic detection and tracking TDT task mainly conveys the recognition and evolution of the topics contained in text streams. Many previous works detect topic through discovering topic bursts from a document stream. (Lappas et al. 2009) presented a approach to model the burstiness of a term, using discrepancy theory concepts. They could identify the time intervals of maximum burstiness for a given term, which is an effective mechanism to address topic detection in the context of text stream. (Lin et al. 2012) proposed burst period detection based on the phenomenon that at some time point, the topic-related terms should appear more frequently than usual and should be continuously frequent around the time point. (Agarwal, Ramamritham, and Bhide 2012) discovered events as dense clusters in highly dynamic graphs. Following the same idea, (Lee, Lakshmanan, and Milios 2014) applied DBSCAN clustering to recognize the evolution pattern of a topic. Though these methods have been successfully adopted in TDT task, they

Table 1: Performance comparisons of the proposed methods and baselines. †, ‡ and $*$ indicate statistically significant differences($p < 0.05$) with **DPP+SR**, **DPP+TP** and **DPP+SR+TP**, respectively.

Method	Recall	Recall^w	Precision	F_1	$F_1{}^w$
TTGPKUICST2	0.3698	0.5840	0.4571	0.3540	0.4575
EM50	0.2867	0.4779	0.4150	0.2546	0.3815
hltcoeTTG1	0.4029	0.5915	0.3407	0.2760	0.3702
DivRank	0.5059	0.6796	0.3383†‡*	0.3514	0.4029†‡*
StarClustering	**0.5221**	**0.7016**	0.2682†‡*	0.2691†‡*	0.3276†‡*
DPP	0.4979	0.6808	0.3447†‡*	**0.3547**	0.4099†‡*
DPP+SR	0.3105	0.5276	0.4581*	0.3230‡*	0.4365*
DPP+TP	0.3424	0.5562	0.4655*	0.3437	0.4585
DPP+SR+TP	0.3234	0.5422	**0.4747**	0.3381	**0.4600**

are not applicable to the TTG problem. The timeline generation problem represents a natural extension of traditional retrieval (Lin and Efron 2014), which means the generation process is based on the documents returned by the search engines. Therefore, major techniques used in TDT such as burst period detection and dense-based clustering cannot be well applied in generating timeline since many subtopics or aspects in the timeline just contain exactly one document.

Timeline Generation There are also several works studying the timeline generation recently. A greedy algorithm based on approximation of Minimum-Weight Dominating Set Problem (MWDS) is exploited in (Wang, Li, and Ogihara 2012; Lin et al. 2012; Zhou et al. 2014). Among these works, (Wang, Li, and Ogihara 2012) proposed an approach that combines image and text analysis to generate a timeline containing textual, pictorial and structural information. They first constructed a multi-view graph, in which each node contains textual and pictorial information, and then selected the representative nodes by finding a minimum dominant set on the graph. Based on the same idea, (Lin et al. 2012) adopted the method to tweet timeline generation. (Xu, McNamee, and Oard 2014) proposed a novel detection approach, framing the problem of redundant tweet removal as a sequential binary decision task. There are also studies that try to characterize different aspects of timeline generation, e.g. from the users' perspective (Li and Cardie 2014) or social attention (Zhao et al. 2013).

Timeline generation is also related to document summarization tasks. Among many variants of traditional summarization, temporal summarization (Yan et al. 2011; Aslam et al. 2013; Guo, Diaz, and Yom-Tov 2013) is in particular related to the TTG task. The difference is that temporal summarization is mainly operating on news documents, without much noise from microblog short text data.

Determinantal Point Processes One of the key features of determinantal point processes is their ability to model the notion of diversity while respecting quality, a concern that underlies the broader task of subset selection where balancing quality with diversity is a well-known issue. Such tasks include document summarization (Kulesza and Taskar 2011b), video summarization (Gong et al. 2014), recommender systems (Gillenwater et al. 2014), and even in some topics in the scope of neural science (Snoek, Zemel, and Adams 2013). Perhaps the most relevant study is the previous work to select diverse structured threads with k-structured DPPs (Gillenwater, Kulesza, and Taskar 2012a).

A related model is fixed-size DPP, also known as k-DPP (Kulesza and Taskar 2011a). The difference is that our spectral rescaling strategy is not designed for fixing the number of selected items. We use the estimated set size to rescale L-ensembles and help balancing relevance and diversity.

Conclusion and Future Work

In this work we propose a novel approach based on determinantal point process for the task of tweet timeline generation. To balance the effects between query relevance and selectional diversity, we design a forward-backward rescaling algorithm to automatically control the overall scaling of relevance scores. Since many irrelevant candidate tweets may be still assigned with high relevance scores from the underlying retrieval system, DPPs that favors diversity tend to select such tweets. We tackle this issue by setting a topical prior. Experimental results suggest our novel DPP-based method along with the proposed strategies gives competitive performance against several state-of-the-art systems which may involve more hand-tuned parameters, features or rules.

With the flexible nature of probabilistic modeling, many possibilities for further extensions exist. The definition of relevance score can be substituted with a mixture model that integrates results from multiple relevance scoring systems. We can also consider using supervised methods to improve the current approach.

The topical prior in our model is designed to capture global topical coherence of the retrieved timeline. Evidences from other related tasks suggest that local temporal coherence can sometimes still play a role. We are going to explore how to capture local coherence well and further improve the performance.

On the other hand, there exists some recent progress on MAP inference for DPPs. Meanwhile, drawing samples from a DPP as a probabilistic model has been shown to be quite efficient. We may perform minimum Bayes risk decoding to get better solutions, according to certain task-specific loss functions on those random samples.

Acknowledgments

We thank the anonymous reviewers for valuable suggestions on an earlier draft of this paper. This work was jointly supported by National Hi-Tech Research and Development Program (863 Program) of China (2015AA015403, 2014AA015102) and National Natural Science Foundation of China (61502502, 61170166, 61331011).

References

Agarwal, M. K.; Ramamritham, K.; and Bhide, M. 2012. Real time discovery of dense clusters in highly dynamic graphs: identifying real world events in highly dynamic environments. *Proceedings of the VLDB Endowment* 5(10):980–991.

Aslam, J.; Ekstrand-Abueg, M.; Pavlu, V.; Diaz, F.; and Sakai, T. 2013. Trec 2013 temporal summarization. In *TREC' 13*.

Gillenwater, J. A.; Kulesza, A.; Fox, E.; and Taskar, B. 2014. Expectation-maximization for learning determinantal point processes. In *Advances in Neural Information Processing Systems*, 3149–3157.

Gillenwater, J.; Kulesza, A.; and Taskar, B. 2012a. Discovering diverse and salient threads in document collections. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, 710–720. Association for Computational Linguistics.

Gillenwater, J.; Kulesza, A.; and Taskar, B. 2012b. Near-optimal map inference for determinantal point processes. In *Advances in Neural Information Processing Systems*, 2735–2743.

Gong, B.; Chao, W.-L.; Grauman, K.; and Sha, F. 2014. Diverse sequential subset selection for supervised video summarization. In *Advances in Neural Information Processing Systems*, 2069–2077.

Guo, Q.; Diaz, F.; and Yom-Tov, E. 2013. Updating users about time critical events. In *Advances in Information Retrieval*. Springer. 483–494.

Kulesza, A., and Taskar, B. 2011a. k-dpps: Fixed-size determinantal point processes. In *Proceedings of the 28th International Conference on Machine Learning (ICML-11)*, 1193–1200.

Kulesza, A., and Taskar, B. 2011b. Learning determinantal point processes. In *UAI*.

Kulesza, A., and Taskar, B. 2012. Determinantal point processes for machine learning. *arXiv preprint arXiv:1207.6083*.

Lappas, T.; Arai, B.; Platakis, M.; Kotsakos, D.; and Gunopulos, D. 2009. On burstiness-aware search for document sequences. In *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, 477–486. ACM.

Lee, P.; Lakshmanan, L. V.; and Milios, E. E. 2014. Incremental cluster evolution tracking from highly dynamic network data. In *Data Engineering (ICDE), 2014 IEEE 30th International Conference on*, 3–14. IEEE.

Li, J., and Cardie, C. 2014. Timeline generation: Tracking individuals on twitter. In *Proceedings of the 23rd international conference on World wide web*, 643–652. ACM.

Lin, J., and Efron, M. 2014. Overview of the TREC-2014 Microblog Track. In *TREC'14*.

Lin, C.; Lin, C.; Li, J.; Wang, D.; Chen, Y.; and Li, T. 2012. Generating event storylines from microblogs. In *Proceedings of the 21st ACM international conference on Information and knowledge management*, 175–184. ACM.

Lv, C.; Fan, F.; Qiang, R.; Fei, Y.; and Yang, J. 2014. PKUICST at TREC 2014 Microblog Track: Feature Extraction for Effective Microblog Search and Adaptive Clustering Algorithms for TTG.

Mei, Q.; Guo, J.; and Radev, D. 2010. Divrank: the interplay of prestige and diversity in information networks. In *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining*, 1009–1018. ACM.

Snoek, J.; Zemel, R.; and Adams, R. P. 2013. A determinantal point process latent variable model for inhibition in neural spiking data. In *Advances in Neural Information Processing Systems*, 1932–1940.

Vondrák, J.; Chekuri, C.; and Zenklusen, R. 2011. Submodular function maximization via the multilinear relaxation and contention resolution schemes. In *Proceedings of the forty-third annual ACM symposium on Theory of computing*, 783–792. ACM.

Walid, M.; Wei, G.; and Tarek, E. 2014. QCRI at TREC 2014:Applying the KISS principle for TTG task in the Microblog Track.

Wang, X., and Zhai, C. 2007. Learn from web search logs to organize search results. In *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*, 87–94. ACM.

Wang, D.; Li, T.; and Ogihara, M. 2012. Generating pictorial storylines via minimum-weight connected dominating set approximation in multi-view graphs. In *AAAI*.

Xu, T.; McNamee, P.; and Oard, D. W. 2014. Hltcoe at trec 2014: Microblog and clinical decision support.

Yan, R.; Kong, L.; Huang, C.; Wan, X.; Li, X.; and Zhang, Y. 2011. Timeline generation through evolutionary trans-temporal summarization. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, 433–443. Association for Computational Linguistics.

Zhao, X. W.; Guo, Y.; Yan, R.; He, Y.; and Li, X. 2013. Timeline generation with social attention. In *Proceedings of the 36th international ACM SIGIR conference on Research and development in information retrieval*, 1061–1064. ACM.

Zhou, W.; Shen, C.; Li, T.; Chen, S.; Xie, N.; and Wei, J. 2014. Generating textual storyline to improve situation awareness in disaster management. In *In Proceedings of the 15th IEEE International Conference on Information Reuse and Integration (IRI 2014)*.