STExNMF: Spatio-Temporally Exclusive Topic Discovery for Anomalous Event Detection

Sungbok Shin
Korea University
Seoul, South Korea

Minsuk Choi
Korea University
Seoul, South Korea

Jinho Choi
Korea University
Seoul, South Korea

Scott Langevin
Uncharted Software Inc.
Toronto, ON, Canada

Christopher Bethune
Uncharted Software Inc.
Toronto, ON, Canada

Philippe Horne
Uncharted Software Inc.
Toronto, ON, Canada

Nathan Kronenfeld
Uncharted Software Inc.
Toronto, ON, Canada

Ramakrishnan Kannan
Oak Ridge National Laboratory
Oak Ridge, TN, USA

Barry Drake
Georgia Tech. Research Institute
Atlanta, GA, USA

Haesun Park
Georgia Tech.
Atlanta, GA, USA

Jaegul Choo
Korea University
Seoul, South Korea

***Abstract*—Understanding newly emerging events or topics associated with a particular region of a given day can provide deep insight on the critical events occurring in highly evolving metropolitan cities. We propose herein a novel topic modeling approach on text documents with spatio-temporal information (e.g., when and where a document was published) such as location-based social media data to discover prevalent topics or newly emerging events with respect to an area and a time point. We consider a map view composed of regular grids or tiles with each showing topic keywords from documents of the corresponding region. To this end, we present a tile-based spatio-temporally exclusive topic modeling approach called STExNMF, based on a novel nonnegative matrix factorization (NMF) technique. STExNMF mainly works based on the two following stages: (1) first running a standard NMF of each tile to obtain general topics of the tile and (2) running a spatio-temporally exclusive NMF on a weighted residual matrix. These topics likely reveal information on newly emerging events or topics of interest within a region. We demonstrate the advantages of our approach using the geo-tagged Twitter data of New York City. We also provide quantitative comparisons in terms of the topic quality, spatio-temporal exclusiveness, topic variation, and qualitative evaluations of our method using several usage scenarios. In addition, we present a fast topic modeling technique of our model by leveraging parallel computing.**

***Index Terms*—Topic modeling; social network analysis; matrix factorization; event detection; anomaly detection**

I. Introduction

Social networking services, such as Facebook and Twitter have successfully established themselves as a new media of communication. They have deeply involved themselves into various forms of social activities in diverse areas, including businesses, health managements, and entertainments. Such affluent uses of social networking services triggered studies using social media data, one of which is that on location-based social media data. They are utilized in developing new methods of detecting anomalous events, understanding the sentiments of users, recommending point-of-interest areas for travelers, and so on.

Jaegul Choo is the corresponding author. E-mail: jchoo@korea.ac.kr.

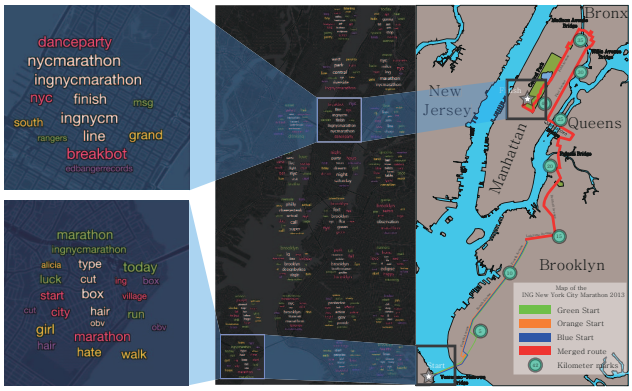


Fig. 1: Topic examples generated by our method on a tile-based map interface. The dark-colored map in the center shows topics of New York City on November 3, 2013. The map on the right shows the running course of the *2013 ING New York City Marathon*. The highlighted tiles on the left show their topics revealing the location of the start (e.g., 'start,' 'city,' and 'marathon') and the finish line (e.g., 'finish,' 'line,' and 'ingnycm') of the course.

Grid- or tile-based map systems have been broadly used in practice, especially in web-based services because of their advantage in parallel handling of large-scale data. In other words, tile-based processing splits up the entire documents into multiple small tile segments, making it more efficient when computing in a real-world environment. Various applications such as Google Maps adopt a tile-based map system as their main interface.

Topic modeling is a well-known machine learning technique that automatically extracts a set of topics from a large-scale document corpus. Each topic corresponds roughly to

2374-8486/17 $31.00 © 2017 IEEE
DOI 10.1109/ICDM.2017.53

(a) Standard NMF

(b) Our approach

Fig. 2: Topic examples extracted from geo-tagged Twitter data set from several tiles of New York City on July 13, 2013

a subset of documents, summarized by a few semantically coherent keywords in it. Topic modeling has been widely studied in analyzing large-scale data because of this effective representation capability.

Computing topic modeling of geo-tagged social media data on a tile-based interface makes it possible to recognize topics occurring in various local areas of the city, as shown in Fig. 1. However, the overflow of noise and everyday language in social media data are the main bottlenecks for extracting informative topics as the presence of trivial keywords in a topic restrains the topic from becoming semantically meaningful. Fig. 2(a) shows the topics extracted using the standard NMF from several sampled tiles of the geo-tagged Twitter data of New York City. Uninteresting keywords such as 'strong' and 'love' make topics less coherent. This disturbs users from obtaining insightful information on tile-based visual analytics systems, where each tile is usually designed to display only a few dominant topics.

In addressing these issues, our work aims to provide a novel topic modeling approach that performs a spatio-temporal analysis of geo-tagged social media data by leveraging the functionalities of a tile-based map interface. We propose STExNMF, a topic modeling algorithm that extracts geospatio-temporally prominent topics for each tile on tile-based map systems. STExNMF is based on a novel nonnegative matrix factorization (NMF) [14] technique that leverages the idea of a weighted residual matrix. Fig. 2(b) shows the topics extracted using our method.

STExNMF mainly works in two steps; first, it runs the standard NMF separately on each tile to obtain general topics that summarize the documents of the corresponding tile. Second, it runs the spatio-temporally exclusive NMF algorithm with a user-specified parameter that controls the exclusiveness of the topic. We show the effectiveness of our work using the Twitter data set of New York City.

The main contributions of this paper are as follows:

- We develop a novel NMF technique called STExNMF, which extracts geospatio-temporally exclusive topics using a weight-controlled residual matrix. We then introduce a parallel algorithm for STExNMF for a tile-based map interface.
- We present a quantitative analysis by comparing our method with other methods and by conducting sensitivity analysis to show the advantages and the characteristics of our approach.
- We perform a qualitative analysis using real-world geo-tagged Twitter data sets. Our method extracts meaningful and distinguished topics in terms of its region and time point.

The rest of this paper is organized as follows; Sec. II discusses the related work. Sec. III presents our proposed methods. Sec. IV shows the qualitative and quantitative evaluations of our method. Finally, Sec. V concludes our paper.

II. Related Work

In this section, we discuss the related works on discriminative topic modeling, topic modeling on social media data and spatio-temporal event analytics using such data.

A. Discriminative Topic Modeling

A number of studies have focused on extracting discriminative topics from a document corpus to obtain topics as diverse as possible. DiscLDA [13] has provided a discriminative learning framework based on a widely-used topic modeling method called latent Dirichlet allocation (LDA) [1]. Another LDA-driven model, LDTM has performed locally discriminative topic modeling [22]. MedLDA [25] has also been a variant of discriminative topic modeling that utilizes the max-margin principle to train supervised topic models and estimate topic representations suitable for prediction. These methods have aimed to improve the classification or regression performances of each cluster, whereas our goal is to intentionally extract discriminative topics from a dataset.

Kim et al. [10] have introduced a group-sparsity regularization method for NMF that aims to search for common and discriminative patterns among multiple groups of data items and features. However, the model does not cover the dissimilarities of unshared latent components and hence is not applicable in discriminative topic extraction. DiscNMF [8] has directly extracted common and discriminative topics of multiple joint matrices via orthogonal regularization of topic vectors. However, DiscNMF has performed redundant computations in extracting topics from each tile when applied as a topic modeling algorithm on a tile-based map interface.

B. Topic Modeling on Social Media

Numerous studies stemmed from LDA have tackled topic analysis using social media data. Hong et al. [6] have proposed an LDA-based model that encourages geographical diversity across different regions given geo-tagged tweets. Vu et al. [18] have built a system that extracts user interests from Twitter messages by ranking the linguistic patterns. Meanwhile, TM-LDA [19] has complemented the lack of a short length of microblog posts by efficiently mining text streams, such as a sequence of posts from a single author.

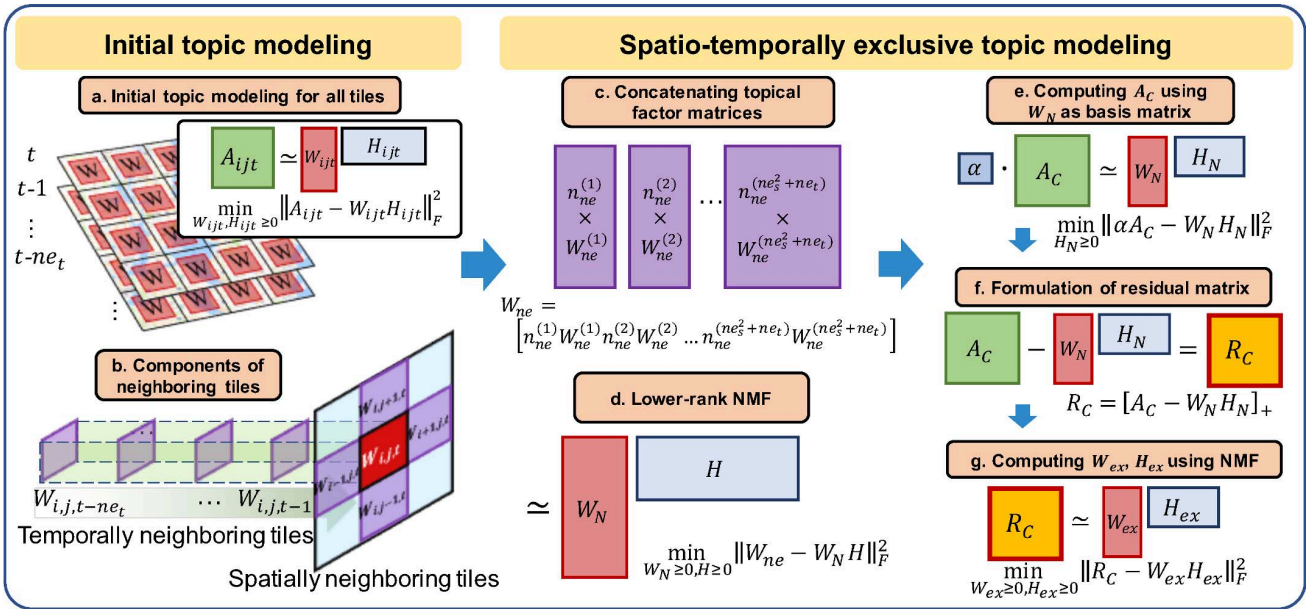


Fig. 3: Overview of STExNMF, composed of initial topic modeling followed by spatio-temporally exclusive topic modeling. (a) We compute the initial topic modeling of each tile using standard NMF. (b) After collecting the topic matrices of the spatio-temporally neighboring tiles obtained from the initial topic modeling, (c) we aggregate them to form a matrix W_{ne} (Eq. (2)) containing the topic information of the spatio-temporally neighboring tiles. (d) We further perform a lower-rank NMF on W_{ne} to remove any possible redundancies in it, resulting in W_N (Eq. (3)). (e) We then compute the explainable part of αA_C using the basis matrix W_N (Eq. (4)) and (f) subtract $W_N H_N$ from A_C to create the residual matrix R_C (Eq. (5)). (g) Finally, we compute NMF on R_C to obtain the resulting exclusive topic matrix W_{ex} (Eq. (6)).

There have also been numerous studies that conducted social media topic analyses using other methods such as NMF and sketching technique [5]. L-EnsNMF [17] has extracted interesting local topics from ensemble learning of residual matrices. Matsutani et al. [16] have addressed the problem of predicting trusklinks among users in social media sites. Moreover, a count-min sketching technique which efficiently computes the summary of the streaming data has been combined with topic modeling to detect bursty topics from Twitter data in a real-time basis [23].

C. Spatio-Temporal Event Analytics for Social Media

Developing temporal event analytic systems using social media data has been an active area of research. Lin et al. [15] have introduced a model that tracks the evolution of a topic and reveals the latent diffusion paths of that topic in a social community. LPTA [24] has tackled the problem of latent periodic topic analysis from time-stamped documents by exploiting the periodicity and the co-occurrences of topics. In addition, various temporal event analytic models have been developed with the help of visualization. TIARA [20] has combined a ThemeRiver style of visualization and topic modeling for time-evolving topics. TargetVue [2] has displayed a novel glyph visualization of the anomalous users' temporal usage patterns in social media.

Event analytics systems that consider both spatial and temporal events have been developed. Pairfac [21] has been an event analytic model that considers the location, time, and venue of each activity using tensor decomposition. Hu et al. [7] have developed a location recommendation model by capturing the spatio-temporal aspects of user check-ins based on topic modeling. Chen et al. [3] have developed another spatio-temporal visual analytics system which facilitates the understanding of people's movements using geo-tagged social media data.

Our approach aims to detect anomalous events by deploying a novel topic modeling approach using social media data specialized for a tile-based map interface. It extracts discriminative topics of a region with respect to multiple neighboring tiles using NMF on a weighted residual matrix. It reveals latent topics by extracting spatio-temporally exclusive topics that can be utilized for analyzing anomalous events. Furthermore, our approach is capable of parallelizing the process, enabling efficient computations when employed in tile-based map interface.

III. STExNMF

This section presents the geospatio-temporally exclusive nonnegative matrix factorization, or STExNMF. First, we describe the initial step of computing tile-wise topics via the standard NMF. Next, we formulate our novel topic modeling approach called STExNMF, which extracts geospatio-

TABLE I: Notations used in the paper

Notation	Description
m	Total number of keywords
k	Number of topics created during initial topic modeling
k_{ne}	Number of topics in W_N
k_{ex}	Number of exclusive topics
α	Exclusiveness parameter
$n_{ne}^{(i)}$	Number of documents contained in the neighboring tile i
$W_{ijt} \in \mathbb{R}_+^{m \times k}$	Topic matrix of tile (i, j, t)
$W_{ne}^{(i)} \in \mathbb{R}_+^{m \times k}$	Topic matrix of the neighboring tile i
$W_{ne} \in \mathbb{R}_+^{m \times k\left(ne_s^2+ne_t\right)}$	Column-concatenated topic matrix of $W_{ne}^{(i)}$'s
$W_N \in \mathbb{R}_+^{m \times k_{ne}}$	Lower-rank topical factor matrix of W_{ne}
$W_{ex} \in \mathbb{R}_+^{m \times k_{ex}}$	Topic matrix of R_C
$A_{i,j,t} \in \mathbb{R}_+^{m \times n_{ijt}}$	Term-document matrix corresponding to tile (i, j, t)
$A_C \in \mathbb{R}_+^{m \times n_C}$	Term-document matrix of a tile of interest
$R_C \in \mathbb{R}_+^{m \times n_C}$	Residual matrix

temporally exclusive topics of a particular region on a tile-based map visualization. Fig. 3 shows the method overview.

A. Initial Topic Modeling on Spatio-Temporal Tiles

Problem Setting. Given the entire document corpus where each document has a timestamp and a geo-location information (e.g. geo-tagged Twitter data), let us consider its term-document matrix representation $A \in \mathbb{R}_+^{m \times n}$. $\mathbb{R}_+$ denotes the set of nonnegative real numbers. m is the vocabulary size, and n is the total number of documents. The matrix element indicates the number of occurrences of a particular term in a particular document.

We split the documents into equally sized spatio-temporal grids called *tiles*, with respect to different time points and locations. Let us denote $A_{i,j,t} \in \mathbb{R}_+^{m \times n_{ijt}}$ (or A_{ijt} in short) as a term-document (sub-)matrix containing the subset of columns of A, whose corresponding documents belong to the tile with the latitude and the longitude indices i and j, respectively, and the time index t, setting one day as the basic time unit.

Standard Topic Modeling. For our initial topic modeling, we apply the standard NMF on each A_{ijt} as the input matrix corresponding to a particular region (i, j) and a time point t. Given A_{ijt} and the number of topics $k \ll \min(m, n_{ijt})$, NMF solves a lower-rank approximation as follows:

$$\min_{W_{ijt},\, H_{ijt} \geqq 0,} \parallel A_{ijt} - W_{ijt} H_{ijt} \parallel_F^2, \tag{1}$$

where $W_{ijt} \in \mathbb{R}_+^{m \times k}$ and $H_{ijt} \in \mathbb{R}_+^{k \times n_{ijt}}$ are the two factor matrices. W_{ijt} represents A_{ijt} as k nonnegative column vectors of W_{ijt} corresponding to the bag-of-words representation of k topics. The l-th nonnegative column vector $w_l \in \mathbb{R}_+^{m \times 1}$ represents the l-th topic as a weighted combination of m keywords. A larger element in a particular column vector indicates the keyword more relevant to the topic. Without loss of generality, we assume that each column of W_{ijt} is normalized to have a unit L_2-norm. The i-th column vector $h_i \in \mathbb{R}_+^{k \times 1}$ of H_{ijt} represents a set of n documents, each of which is described as a weighted combination of k topics.

B. Spatio-Temporally Exclusive Topic Modeling

STExNMF aims to extract the spatio-temporally exclusive topics of each tile that convey the newly emerging or anomalous event information corresponding to the tile. STExNMF accomplishes the task by leveraging the information of the spatio-temporally neighboring tiles (i.e., W_{ijt}'s computed from such neighboring tiles with respect to the chosen tile), as shown in Fig. 3(a).

Let us denote the spatial and the temporal indices of a chosen tile $C = \{i_c, j_c, t_c\}$, as illustrated in Fig. 3(b). Let $A_C \in \mathbb{R}_+^{m \times n_C}$ be the term-document matrix of the chosen tile. Consider the topic matrices, W_{ijt}s of the spatially neighboring tiles, where the tile indices (i, j, t) of spatially neighboring tiles are defined as

$$(i_c \pm 1, j_c \pm 1, t_c), (i_c \pm 2, j_c \pm 2, t_c), \\ \cdots, (i_c \pm ne_s, j_c \pm ne_s, t_c).$$

ne_s is the window size of the spatial neighbors. Next, consider those of the temporally neighboring tiles, where the tile indices (i, j, t) of the temporally neighboring tiles are defined as

$$(i_c, j_c, t_c - 1), (i_c, j_c, t_c - 2), \cdots, (i_c, j_c, t_c - ne_t),$$

where ne_t is the window size of the temporal neighbors.

For simplicity, we enumerate all the spatially or temporally neighboring topic matrices W_{ijt}'s and their number of documents n_{ijt}'s as $W_{ne}^{(i)}$ and $n_{ne}^{(i)}$, respectively, for $i = 1, 2, \cdots, ne_s^2 + ne_t$.

Each $W_{ne}^{(i)}$ can be viewed as k virtual documents summarizing the documents belonging to a particular neighboring tile. Collecting them together, we form a set of $k\left(ne_s^2 + ne_t\right)$ virtual documents and further apply the standard NMF to its column-concatenated (virtual) term-document matrix representation to remove any redundancies and improve the quality of these summary topics.

Different tiles have varying numbers of documents, $n_{ne}^{(i)}$, but each of them is summarized by the same number k of virtual documents (or topics). Considering this, we assign different weights to $W_{ne}^{(i)}$'s, which results in the following (virtual) term-document matrix as

$$\begin{aligned} W_{ne} &= \left[\, n_{ne}^{(1)} W_{ne}^{(1)} \quad n_{ne}^{(2)} W_{ne}^{(2)} \right. \\ &= \left. \cdots \quad n_{ne}^{\left(ne_s^2+ne_t\right)} W_{ne}^{\left(ne_s^2+ne_t\right)} \,\right] \in \mathbb{R}_+^{m \times k\left(ne_s^2+ne_t\right)}. \end{aligned} \tag{2}$$

The following procedure is described in Fig. 3(c). We then apply NMF on this matrix, that is,

$$\min_{W_N \geq 0,\, H \geq 0} \|W_{ne} - W_N H\|_F^2, \tag{3}$$

where $W_N \in \mathbb{R}_+^{m \times k_{ne}}$, $H \in \mathbb{R}_+^{k_{ne} \times k\left(ne_s^2+ne_t\right)}$, and k_{ne} is the reduced rank of this factorization, as shown Fig. 3(c).

The purpose of the procudure is to summarize W_{ne} so that redundant topics are discarded. Consequently, W_N acts as a matrix representing the topics from all the spatio-temporally neighboring tiles.

The next step is to compute the part of A_C explainable by using W_N as a basis matrix, which will then be removed from A_C, as portrayed in Fig. 3(d). To control how aggressively to remove the explainable part from A_C, we introduce a parameter α, such that $0 \leq \alpha \leq 1$. α is used to solve the nonnegativity-constrained least squares (NCLS) problem as

$$\min_{H_N \geq 0,} \|\alpha A_C - W_N H_N\|_F^2 , \tag{4}$$

as shown in Fig. 3(e). W_N is a constant matrix previously computed from Eq. (3). α controls the amount of A_C to be explained by W_N up to αA_C. The lower the α is, the less portion of A_C is approximated by W_N. Hence, in the topic modeling context, the influence of the neighboring matrices to the explained part of A_C decreases as α decreases.

We then define the residual matrix R_C as the nonnegative projection of the difference of $W_N H_N$ from A_C, as illustrated in Fig. 3(e).

$$\begin{aligned} R_C &= [A_C - W_N H_N]_+ \\ &= [(1-\alpha) A_C + (\alpha A_C - W_N H_N)]_+ . \end{aligned} \tag{5}$$

The residual matrix R_C is composed of (1) the remaining portion $(1-\alpha) A_C$ of the original A_C and (2) the unexplained part of αA_C even after using W_N as a basis matrix. Finally, as in Fig. 3(f), STExNMF performs the last stage of NMF on this residual matrix R_C as

$$\min_{W_{ex} \geq 0, H_{ex} \geq 0} \|R_C - W_{ex} H_{ex}\|_F^2 , \tag{6}$$

where $W_{ex} \in \mathbb{R}_+^{m \times k_{ex}}$ and $H_{ex} \in \mathbb{R}_+^{k_{ex} \times n_C}$ are the two factor matrices. The residual matrix R_C mainly contains the information unexplained by the topics from the neighboring tiles. Hence, the resulting topic matrix W_{ex} can reveal the exclusive topics of the chosen tile against its spatio-temporal neighbors. Algorithm 1 summarizes the entire STExNMF procedure.

Algorithm 1: Spatio-temporally Exclusive NMF (**STExNMF**)

Input: Term-document matrix of a chosen tile $A_C \in \mathbb{R}_+^{m \times n_C}$. Topic matrix of the neighboring tiles $W_{ne}^{(i)} \in \mathbb{R}_+^{m \times k}$. Number of topics of W_N, k_{ne}, and $\alpha \in \mathbb{R}$ $(0 \leq \alpha \leq 1)$.
Output: $W_{ex} \in \mathbb{R}_+^{m \times k_{ex}}$ and $H_{ex} \in \mathbb{R}^{k_{ex} \times n_C}$.
Compute the initial topic modeling on all tiles using Eq. (1).
Compute W_N using Eq. (3).
Compute H_N using Eq. (4).
Compute R_C using Eq. (5).
Compute W_{ex} and H_{ex} using Eq. (6).

C. Efficient Algorithm for STExNMF

A main aspect of STExNMF is to compute the topic modeling tile-by-tile. Tiles subdivide the number of the entire documents. This leads a reasonable number of documents per tile. Hence, we keep the reduced ranks in the NMF computation, such as k, k_{ne}, and k_{ex} in Eqs. (1), (3), and (6), respectively, as a relatively small value (e.g., 2 or 4). To this end, we extend a highly efficient NMF algorithm based on a successive rank-2 matrix factorization in a hierarchical manner (HierNMF2) [12], which significantly increases the efficiency for small rank values.

STExNMF basically performs each NMF in a two-block coordinate descent framework, which iteratively computes W while fixing H and vice-versa. Each sub-problem can be formulated in detail as the NCLS problem, that is,

$$\min_{V \geq 0} \|UV - X\|_2^2 = \min_{\mathbf{v}_i \geq 0} \sum_i \|U\mathbf{v}_i - \mathbf{x}_i\|_2^2 \tag{7}$$

where U, V, and X correspond to H^T, W^T, and A_{ijt}^T, respectively, when solving for W in Eq. (1), and $\mathbf{x}_i$ and $\mathbf{v}_i$ are the i-th columns of X and V, respectively. This problem can be independently solved by minimizing each term inside the summation as

$$\min_{\mathbf{v}_i \geq 0} \|U\mathbf{v}_i - \mathbf{x}_i\|_2^2 \tag{8}$$

for all i's. Each element of $\mathbf{v}_i$ will either be zero or positive because of the nonnegativity constraint. Eq. (8) can then be equivalently reduced to an unconstrained least squares problem [11] if we consider the set $\mathbb{P}$ of dimension indices of strictly positive values,

$$\|U(:, \mathbb{P})\, \mathbf{v}_i(\mathbb{P}) - \mathbf{x}_i(\mathbb{P})\|_2^2 , \tag{9}$$

which can be easily solved.

$\mathbb{P}$ is unknown a priori. Therefore, the algorithm iteratively refines $\mathbb{P}$ and solves Eq. (9) until the optimal $\mathbb{P}$ is reached. However, this refinement process generally requires numerous iterations because of the exponentially growing number of possibilities in terms of dimensions. HierNMF2 expedites this process by exhaustively solving Eq. (9) for all the different $\mathbb{P}$'s when the rank is set as two. It then chooses the best one with the smallest loss function value. HierNMF2 builds a binary tree of rank-2 NMF by recursively splitting data items into child nodes, until the number of leaf nodes reach the desired rank for a rank larger than two.

Computational Complexity. STExNMF computation is composed of three different stages: (1) computing NMF for dimension reduction (Eq. (3)), (2) solving the NCLS problem (Eq. (4)), and (3) performing NMF on the residual matrix R_C (Eq. (6)).

Determining the computational complexity of NMF or NCLS problems is generally difficult because the required number of iterations until convergence varies depending on the random initialization. The dominant computation for the NCLS problem in Eq. (7), which is also the NMF subproblem, takes $O(mnk)$, where the size of an input matrix X is

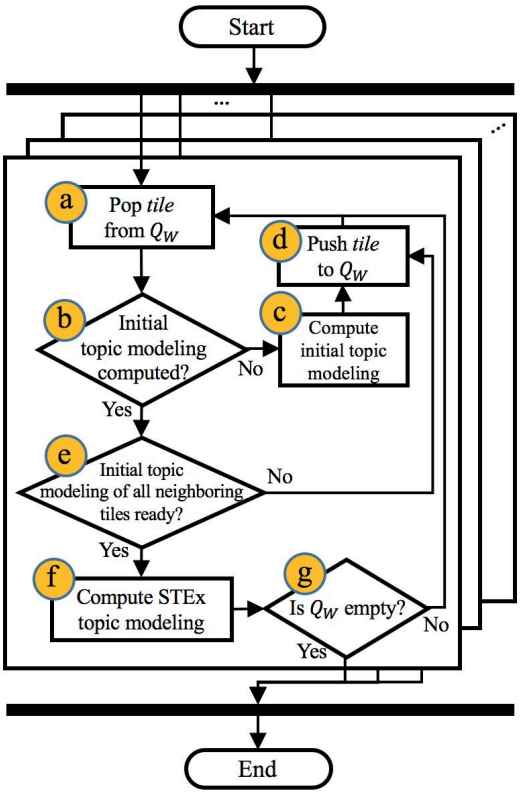


Fig. 4: Parallellized algorithm of STExNMF

$m \times n$, and the rank is k. Considering the maximum size of the NCLS problem involved in the above-mentioned three stages, its computational complexity in STExNMF is derived as $O\left(m \times \max_{ijt}\left(n_{ijt}\right) \times \max\left(k,\, k_{ne}, k_{ex}\right)\right)$.

D. STExNMF Parallelization

This section presents the proposed parallelization strategy of the entire STExNMF process on a tile-map based interface so that it can efficiently compute the topic modeling results on a daily basis, given ever-increasing document data.

We first define a data structure for tile $\mathcal{T}_{ijt}$ containing its sparse term-document matrix A_{ijt}, its standard topic matrix W_{ijt} computed from Eq. (1), and the exclusive topic matrix W_{ex} computed from Eq. (6). We mainly divide the task into two stages. As mentioned in Sec. III-A, given a tile $\mathcal{T}_{ijt}$, we first compute its initial topic modeling to obtain W_{ijt} via the standard NMF. We then compute spatio-temporally exclusive topic modeling to obtain W_{ex} as described in Sec. III-B.

As summarized in Fig. 4, our parallelization algorithm works as follows: in a multi-threaded environment, a worker thread pops a tile $\mathcal{T}_C$ from a global task queue Q_W (Fig. 4(a)), where Q_W can be simultaneously accessed by all worker threads and holds information of all the jobs to be done.

The thread first checks if the tile has finished computing the initial topic modeling (Fig. 4(b)). The thread computes the initial topic modeling if this job has not yet started (Fig. 4(c)). Once the topic modeling is finished, $\mathcal{T}_C$ is pushed back to Q_W (Fig. 4(d)).

Another worker thread later picks up $\mathcal{T}_C$, the initial topic modeling of which is done from Q_W. It is then checked if the initial topic modeling of all the spatio-temporally neighboring tiles is completed (Fig. 4(e)). If so, the processor computes the spatio-temporally exclusive topic modeling algorithm of $\mathcal{T}_C$ (Fig. 4(f)), but otherwise, the processor pushes $\mathcal{T}_C$ back to Q_W.

When the topic modeling process of a tile is complete, the process checks whether Q_W is empty to see if any other work is left (Fig. 4(g)). The worker thread pops another tile from Q_W and repeats the process if Q_W is not empty. The entire process terminates when Q_W becomes empty.

IV. Experiments

In this section, we present both quantitative comparisons and qualitative use cases of the proposed STExNMF. After describing our experimental setup, we discuss the quantitative comparison results of our work with several baseline methods. We then provide several use cases of our algorithm using real-world data.

A. Experimental Setup

1) Datasets: We used the geo-tagged Twitter data of New York City from July 10, 2013 to July 14, 2013 (i.e., 784,414 documents with 29,223 distinct keywords), and from October 29, 2013 to November 3, 2013 (i.e., 788,604 documents with 22,994 keywords). In both cases, we divided the data into 18 regular grids or tiles with respect to its location.

2) Compared Methods: We compare STExNMF with the existing the standard NMF algorithms and LDA [1], another popular topic modeling method. We include a few different algorithms for the standard NMF, such as the alternating nonnegative least-squares method (ANLS) [9], hierarchical alternating nonnegative least-squares method (HALS) [4], and HierNMF2, which we adopted in STExNMF as explained in Sec. III-C.

As for the parameters, we set the initial number of topics per tile k as 2, the spatial window size ne_s as 1, and the temporal window size ne_t as 4. We set k_{ne} as the minimum value between $2k$ and 5 when computing W_N in Eq. (3) and W_{ex} in Eq. (6). The number of exclusive topics per tile, k_{ex}, is set as 2.

3) Evaluation Measures: We adopt the three following evaluation measures to analyze the quality of the topics generated from our model: topic coherence, topic variation score, and spatio-temporal similarity.

Topic coherence. We use the point-wise mutual information (PMI) to evaluate the quality of individual topics, PMI indicates how likely a pair of keywords would co-occur in a document set. The more the two words co-occur in the same document, the more they are semantically related. Given two words w_i and w_i, PMI is defined as

$$\mathrm{PMI}\left(w_i,\, w_j\right) = \log \frac{P\left(w_i,\, w_j\right)}{P\left(w_i\right) P\left(w_j\right)} \tag{10}$$

where $P(w_i, w_j)$ represents the probability of w_i and w_j co-occurring in the same document and $P(w_i)$ (or $P(w_j)$) denotes the probability of w_i (or w_j) appearing in a document corpus. We select the ten most representative keywords and compute the mean average value among them to compute the PMI scores of each tile.

Spatio-temporal similarity score. The spatio-temporal similarity score (ST-similarity) measures the distinctiveness of the topics with respect to those from its spatio-temporal neighbors. This score reveals the relationship between the topics extracted from a tile and those from other tiles. We define the ST-similarity score for a center tile C as

$$\text{sim}_{ST}(C) = \frac{1}{ne_s^2 + ne_t} \sum_{i=1}^{ne_s^2+ne_t} \left\| (W_{ex})^T W_{ne}^{(i)} \right\|_{1,1} . \quad (11)$$

The ST-similarity is the averaged inner product value among all topic pairs between W_{ex} and each $W_{ne}^{(i)}$. This measure indicates how distinct the topics a particular tile has against the spatio-temporally neighboring tiles. A lower ST-similarity score corresponds to more distinct topics, whereas a larger score corresponds to less distinct topics.

Topic variation score. This score measures how much the topic keywords in W_{ex} differ from their initial topic modeling result W_{ijt} by increasing α. To be specific, let C_α be the union of the topic keywords of W_{ex} computed by STExNMF at a particular α value and C_I be the set of keywords of W_{ijt} of the corresponding tile. We measure the topic variation score $T_v(C_\alpha, C_I)$ as the Jaccard distance between C_α and C_I, i.e.,

$$T_v(C_\alpha, C_I) = 1 - \frac{|C_\alpha \cap C_I|}{|C_\alpha \cup C_I|}, \quad (12)$$

where $|\cdot|$ denotes the cardinality of a set.

A higher topic variation score implies that with an increasing α, a user can expect significantly different keywords from its own initial topic keywords. In our experiment, we select the 20 most representative keywords from each topic extracted from STExNMF and compare them with the keywords extracted from the initial topic modeling.

All experiments were conducted using MATLAB 2016b and python 3.5.0 from the workstation with the following processor: Intel(R) Xeon(R) CPU E5-2687W v3 @ 3.10GHz with 384GB of memory.

B. Quantitative Comparison

This section presents topic quality results, computing times, and use cases for event detection.

1) Topic Quality: Table II shows the PMI topic coherence scores for several topic modeling methods of different k values, along with our method using three different α values. Among the standard NMF methods, the HALS- and ANLS-based standard NMF methods generally show the best performance while HierNMF2 performs slightly worse because of its greedy, successive rank-2 NMF approach. The STExNMF performances with various α are comparable to HierNMF2 because STExNMF works based on the HierNMF2 algorithm for solving core NCLS problems.

The main difference of STExNMF from HierNMF2 is that STExNMF extracts topics from the residual matrix R_C, computed through our sophisticated procedure instead of A_C. Note that the topic coherence is not degraded much but often improved as α becomes large. It may be counter-intuitive because the topic modeling on the input matrix close to the original matrix A_C is expected to yield a high topic coherence score. Our conjecture for its reason is that the process of removing the explainable part by the topics from the neighboring tiles actually works as the noise removal process by, say, removing some randomly appearing but meaningless words, which are often found in noisy social media data such as Twitter data. We also think that the reason for the poor performance of LDA compared to the NMF-based methods is because its susceptibility to a large amount of noise existing in the Twitter data.

Figs. 5(a) and (b) demonstrate the average ST-similarity scores of the tiles covering New York City of July 14, 2013 and November 3, 2013, respectively, by varying the values of α and k. Both figures indicate that the ST-similarity scores decrease as α increases. The amount of A_C explained by topics from its spatio-temporally neighboring tiles increases as α gets larger. Consequently, the residual matrix R_C becomes more spatio-temporally exclusive, making the resulting W_C also spatio-temporally exclusive.

Figs. 6(a) and (b) show the comparisons of the ST-similarity scores of STExNMF with those from the other NMF techniques (i.e., ANLS, HALS, and HierNMF2). The error bar represents the standard deviation of the ST-similarity of the tiles on the map. The graph denotes that the ST-similarity scores of ANLS, HALS, and HierNMF2 are similar to those of STExNMF with a low α value. The topical factor matrix W_C extracted using the low α represents the residual matrix R_C of which only a small portion is explained via its neighboring topics, thereby making R_C similar to its original term-document matrix, A_C.

2) Computing Times: Table III compares the running time of STExNMF using three different NMF algorithms, HALS, ANLS, and HierNMF2, using Twitter data set of July 14, 2013. The results show that HierNMF2, which we adopted in our STExNMF, works fastest compared to the other methods such as HALS and ANLS approaches.

To evaluate the performance of parallel approach of our model, we measure the speedup of its execution time. Fig. 7 shows the linear speedup as the number of threads increases, which is an ideal case.

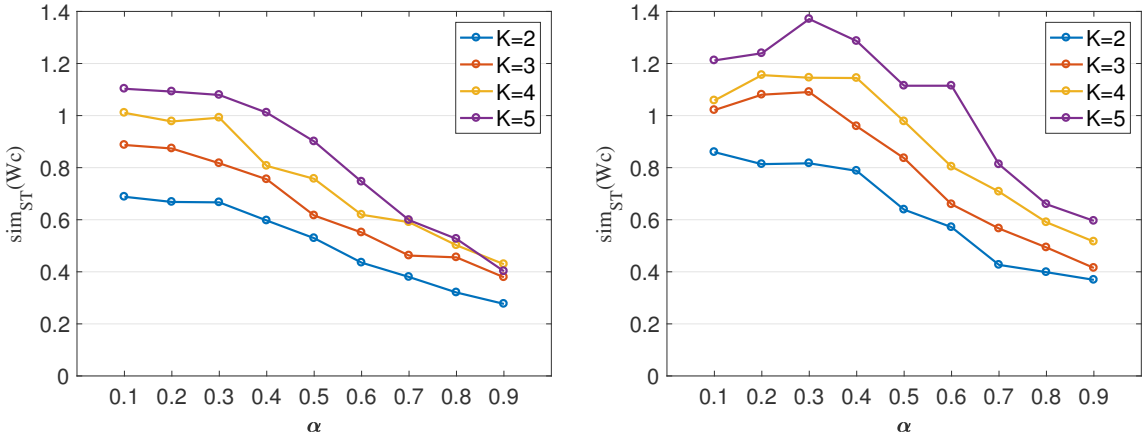
C. Use Cases for Event Detection

This section demonstrates several use cases of STExNMF in event detection using real-world datasets.

1) Northern Central Park and Southern Harlem, July 14, 2013 : The upper part of Table IV shows the two topics extracted from the tile covering the Central Park and Southern Harlem areas on July 14, 2013. The topics extracted with

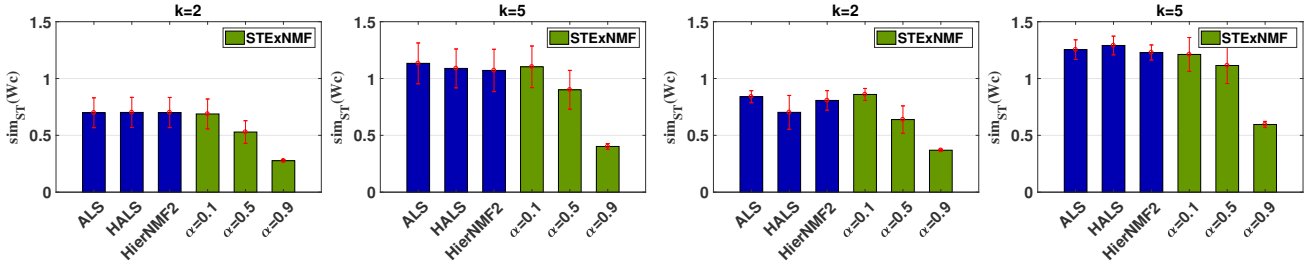
TABLE II: Comparison of topic coherence values. The following results are averaged values of PMIs of partitioned Twitter data sets of 18 tiles. The value in parentheses represent the standard deviation.

Data sets	k		NMF			STEx-NMF		
		LDA	HALS	ANLS	HierNMF2	α=0.1	α=0.5	α=0.9
Jul. 14, 2013	2	0.540 (0.334)	2.510 (1.733)	2.476 (0.861)	1.804 (1.331)	2.272 (0.933)	2.223 (1.211)	2.087 (1.162)
	3	0.510 (0.252)	2.518 (1.511)	2.460 (0.788)	1.931 (1.358)	1.936 (1.211)	2.020 (0.938)	2.001 (0.919)
	4	0.499 (0.206)	2.852 (1.165)	2.614 (0.841)	2.093 (1.521)	2.020 (1.231)	2.435 (0.866)	2.452 (0.801)
	5	0.449 (0.154)	2.615 (1.336)	2.700 (0.869)	2.031 (1.460)	2.025 (1.163)	2.401 (0.481)	2.471 (0.741)
Nov. 3, 2013	2	0.781 (0.571)	2.867 (1.255)	2.791 (1.420)	2.150 (1.754)	2.161 (0.854)	1.850 (0.701)	2.053 (0.793)
	3	0.631 (0.312)	2.886 (1.135)	2.918 (1.159)	2.248 (1.741)	2.484 (0.801)	1.835 (0.514)	1.775 (0.485)
	4	0.625 (0.263)	2.666 (1.368)	2.847 (0.948)	2.265 (1.610)	2.460 (0.670)	1.990 (0.319)	1.986 (0.628)
	5	0.594 (0.270)	2.903 (1.140)	2.737 (0.812)	2.087 (1.465)	2.307 (0.568)	2.007 (0.534)	1.841 (0.451)



(a) July 14, 2013 (b) November 3, 2013

Fig. 5: ST-similarity scores (Eq. (11)) with respect to different values of k and α



(a) July 14, 2013 (b) November 3, 2013

Fig. 6: Comparison of ST-similarity scores (Eq. (11)) among different methods

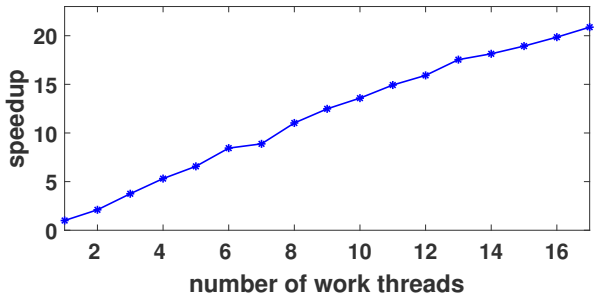


Fig. 7: Speedup results of parallelized STExNMF with respect to the number of worker threads

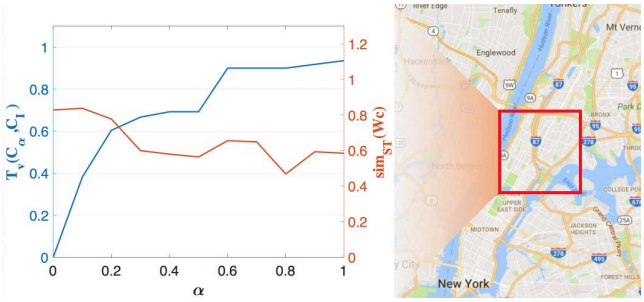
TABLE III: Comparison of running time in seconds among different NMF algorithms of STExNMF on Twitter dataset from the New York City on July 14, 2013

Rank	2	3	5	7	9
HALS	275.216	296.204	309.268	344.247	347.031
ANLS	186.253	200.468	213.372	240.322	250.532
HierNMF2	**16.455**	**52.428**	**83.380**	**96.353**	**92.002**

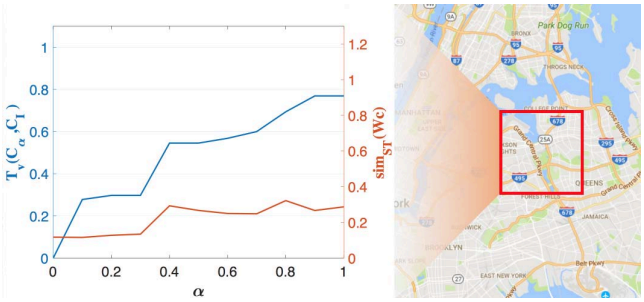
a large α value reveals interesting events from this region. Meanwhile, the topics extracted from the standard NMF and STExNMF with $\alpha = 0.1$ show almost identical topics, because the residual matrix R_c in Eq. (5) preserves a majority (e.g., 90%) of A_C as it is. The first topic contains keywords such

TABLE IV: Topic summaries from several tiles of New York City on July 14. 2013 and November 3, 2013

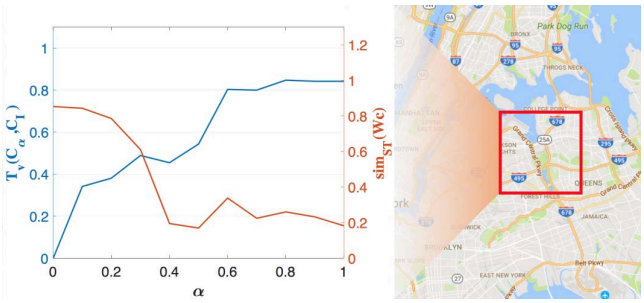
Data set	Topic	standard NMF	STExNMF		
			$\alpha = 0.1$	$\alpha = 0.5$	$\alpha = 0.9$
Northern Central Park and Southern Harlem Jul. 14, 2013	1	strong, love, parent, relationship, cover, book, companionship, people	strong, love, parent, relationship, book, cover, companionship, retweet	strong, love, parent, relationship, cover, book, companionship, retweet	harlem, powell, adam, trayvon, clayton, office, building, manhattan, protest
	2	centralpark, philharmonic, concert, gate, west, mariah, museum, nature, triathlon	centralpark, philharmonic, riverside, west, mariah, museum, triathlon, nature	black, zimmerman, people, trayvon, martin, george, guilty, harlem, free, justice	black, zimmerman, kill, trayvon, george, guilty martin, verdict, justice
Queens, Jul. 14, 2013	1	sunday, check, taco, bell, happy, brunch, baseball, nimmo, team, found, love	sunday, check, taco, bell, happy, brunch, baseball, nimmo, love, team, found	home, run, watch, night, piazza, people, mike, hit guilty, zimmerman, kill	home, run, watch, piazza, night, mike, citifield, hit, celebsoftball, people
	2	mets, citifield, fan, love, futuresgame, baseball, run random, legend, sunday	mets, citifield, fan, legend futuresgame, baseball, run random, championship	citifield, mets, sunday, fan, baseball, legend, random championship, allstargame	queens, corona, center, king, comfort, subway, park, live, hall, unisphere
Queens, Nov. 3, 2013	1	love, fall, astoria, friend, night, girl, watch, party, sunday, brownies, mary	love, live, party, night, upload, astoria, fall, girl, friend, club, watch	party, triplethreat, drink, goodtime, awesome, night jackson, girl, height, live	jackson, heights, friend, greenmarket, happy, food turn, diwali, buy, fresh
	2	music, peace, download, grace, song, track, cain, grate, sean, start, lion	music, peace, download, grace, song, track, cain, grate, lion, start, tonight	music, song, peace, grace download, track, tonight, tonight, start, wait, listen	night, home, party, girl, nurse, happy, astoria, care watch, people, movie,



(a) Northern Central Park and Southern Harlem, July 14, 2013



(b) Queens, July 14, 2013



(c) Queens, November 3, 2013

Fig. 8: ST-similarity scores (Eq. (11)) and topic variation scores (Eq. (12)) of tiles

as 'centralpark,' 'philharmonic,' 'concert,' and 'mariah.' An MLB ALL-STAR Charity Concert starring the New York Philharmonic and the singer Mariah Carey took place at Central Park on July 14, 2013. However, topics at α= 0.9 are different. The first topic contains keywords such as 'harlem,' 'powell,' 'adam,' 'clayton,' 'building,' and 'office.' On the same day the concert took place, the Trayvon Martin Protest also occurred in New York City, mainly in Union Square and Times Square. Though not as dominant as the two spots, the Adam Clayton Powell State Building Plaza, located at Harlem, was also one of the places where the protest was held.

2) Queens, July 14, 2013: The middle part of Table IV shows the two topics extracted from the tile covering Queens, a borough of New York City. Similar to the previous example, the topics extracted from the standard NMF and STExNMF at α= 0.1 exhibit similar topics. The first topic shows keywords such as 'mets,' 'citifield,' 'futuresgame,' and 'baseball' because the Citi Field stadium, which is the home of the New York Mets, is located in Queens. Trivial keywords such as 'sunday,' 'check,' 'taco,' and 'bell' appear in another topic. However, the topics extracted at α= 0.9 convey meaningful information, such as 'home,' 'run,' 'watch,' 'piazza,' and 'celebsoftball' for the first topic. Its relevant event is that Mike Piazza, a Mets legendary player, attended the Taco Bell All-Star Legends and Celebrity Softball Game in July 14, 2013. Mike Piazza hit two home runs on the match.

3) Queens, November 3, 2013: The lower part of Table IV shows another set of the two topics extracted from the tile covering the same area as the previous example. Both topic sets extracted from the standard NMF and STExNMF at α= 0.1 exhibit trivial topics. However, keywords such as 'jackson,' 'heights,' and 'diwali' appear in one topic extracted at α= 0.9. Jackson Heights is a neighborhood of the borough of Queens in New York City, which is famous for its multi-ethnic demographics. Diwali is one of India's biggest festivals celebrated every year in autumn in the northern hemisphere. This festival was held on November 3, 2013.

Finally, in Fig. 8, all the above-mentioned cases show an increase in the topic variation score and a decrease in the ST-similarity as α increases, indicating that a large α sheds lights on minor, but interesting event information.

V. Conclusion and Future Work

In this paper, we proposed a novel topic modeling algorithm called STExNMF, which extracts spatio-temporally exclusive topics using nonnegative matrix factorization designed for a tile-based map interface. Our STExNMF reveals exclusive, thus meaningful topics specific to a particular region within a tile and a time point by discarding the explainable part from the topics of its spatio-temporally neighboring tiles. We also proposed a parallel algorithm of STExNMF, which can efficiently and simultaneously handle multiple tiles with no performance bottleneck. Our quantitative analysis demonstrated the advantages of our approach compared to several baseline methods. Moreover, we presented a few use cases with Twitter data, which revealed interesting event details that would have been otherwise not detected by the other methods.

As our future work, we plan to integrate our approach with a tile-based spatio-temporal visual analytics system, with novel user interactions and real-time interactive algorithms such that the system can serve various real-world needs for anomalous event detection given streaming geo-tagged text data. Furthermore, we plan to improve our event detection model by associating our topic analysis with other types of spatio-temporal data such as GPS or mobile phone data.

Acknowledgments

This manuscript has been authored by UT-Battelle, LLC under Contract No. DE-AC05-00OR22725 with the U.S. DOE and supported by the National Research Foundation of Korea (NRF) grant funded by the Korean government (MSIP) (No. NRF-2016R1C1B2015924). The DOE will provide public access to these results of federally sponsored research in accordance with the DOE Public Access Plan (http://energy.gov/downloads/doe-public-access-plan).

References

[1] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent dirichlet allocation. *Journal of machine Learning research (JMLR)*, 3(Jan):993–1022, 2003.

[2] N. Cao, C. Shi, S. Lin, J. Lu, Y. R. Lin, and C. Y. Lin. Targetvue: Visual analysis of anomalous user behaviors in online communication systems. *IEEE Transactions on Visualization and Computer Graphics (TVCG)*, pages 280–289, 2016.

[3] S. Chen, X. Yuan, Z. Wang, C. Guo, J. Liang, Z. Wang, X. L. Zhang, and J. Zhang. Interactive visual discovering of movement patterns from sparsely sampled geo-tagged social media data. *IEEE Transactions on Visualization and Computer Graphics (TVCG)*, 22(1):270–279, 2016.

[4] A. Cichocki, R. Zdunek, and S.-i. Amari. *Hierarchical ALS Algorithms for Nonnegative Matrix and 3D Tensor Factorization*, pages 169–176. 2007.

[5] G. Cormode and S. Muthukrishnan. An improved data stream summary: the count-min sketch and its applications. *Journal of Algorithms*, 55(1):58 – 75, 2005.

[6] L. Hong, A. Ahmed, S. Gurumurthy, A. J. Smola, and K. Tsioutsiouliklis. Discovering geographical topics in the twitter stream. In *Proc. the International Conference on World Wide Web (WWW)*, pages 769–778, 2012.

[7] B. Hu, M. Jamali, and M. Ester. Spatio-temporal topic modeling in mobile social media for location recommendation. In *Proc. the IEEE International Conference on Data Mining (ICDM)*, pages 1073–1078, 2013.

[8] H. Kim, J. Choo, J. Kim, C. K. Reddy, and H. Park. Simultaneous discovery of common and discriminative topics via joint nonnegative matrix factorization. In *Proc. the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, pages 567–576, 2015.

[9] H. Kim and H. Park. Nonnegative matrix factorization based on alternating nonnegativity constrained least squares and active set method. *SIAM Journal on Matrix Analysis and Applications*, 30(2):713–730, 2008.

[10] J. Kim, R. D. C. Monteiro, and H. Park. *Group Sparsity in Nonnegative Matrix Factorization*, pages 851–862.

[11] J. Kim and H. Park. Sparse nonnegative matrix factorization for clustering. 2008.

[12] D. Kuang and H. Park. Fast rank-2 nonnegative matrix factorization for hierarchical document clustering. In *Proc. the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, pages 739–747, 2013.

[13] S. Lacoste-Julien, F. Sha, and M. I. Jordan. Disclda: Discriminative learning for dimensionality reduction and classification. In *Advances in Neural Information Processing Systems (NIPS)*, pages 897–904. 2009.

[14] D. D. Lee and H. S. Seung. Learning the parts of objects by non-negative matrix factorization. *Nature*, 401(6755):788–791, 1999.

[15] C. X. Lin, Q. Mei, J. Han, Y. Jiang, and M. Danilevsky. The joint inference of topic diffusion and evolution in social communities. In *Proc. the IEEE International Conference on Data Mining (ICDM)*, pages 378–387, 2011.

[16] K. Matsutani, M. Kumano, M. Kimura, K. Saito, K. Ohara, and H. Motoda. Combining activity-evaluation information with nmf for trust-link prediction in social media. In *Proc. the IEEE International Conference on Big Data (Big Data)*, pages 2263–2272, Oct 2015.

[17] S. Suh, J. Choo, J. Lee, and C. K. Reddy. L-ensnmf: Boosted local topic discovery via ensemble of nonnegative matrix factorization. In *Proc. the IEEE International Conference on Data Mining (ICDM)*, pages 479–488, 2016.

[18] T. Vu and V. Perez. Interest mining from user tweets. In *Proc. the ACM International Conference on Information & Knowledge Management (CIKM)*, pages 1869–1872, 2013.

[19] Y. Wang, E. Agichtein, and M. Benzi. Tm-lda: Efficient online modeling of latent topic transitions in social media. In *Proc. the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, pages 123–131, 2012.

[20] F. Wei, S. Liu, Y. Song, S. Pan, M. X. Zhou, W. Qian, L. Shi, L. Tan, and Q. Zhang. Tiara: A visual exploratory text analytic system. In *Proc. the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, pages 153–162, 2010.

[21] X. Wen, Y.-R. Lin, and K. Pelechrinis. Pairfac: Event analytics through discriminant tensor factorization. In *Proc. the ACM International on Conference on Information and Knowledge Management (CIKM)*, pages 519–528, 2016.

[22] H. Wu, J. Bu, C. Chen, J. Zhu, L. Zhang, H. Liu, C. Wang, and D. Cai. Locally discriminative topic modeling. *Pattern Recognition*, 45(1):617 – 625, 2012.

[23] W. Xie, F. Zhu, J. Jiang, E. P. Lim, and K. Wang. Topicsketch: Real-time bursty topic detection from twitter. *IEEE Transactions on Knowledge and Data Engineering (TKDE)*, (8):2216–2229, 2016.

[24] Z. Yin, L. Cao, J. Han, C. Zhai, and T. Huang. Lpta: A probabilistic model for latent periodic topic analysis. In *Proc. the IEEE International Conference on Data Mining (ICDM)*, pages 904–913, 2011.

[25] J. Zhu, A. Ahmed, and E. P. Xing. Medlda: Maximum margin supervised topic models for regression and classification. In *Proc. the International Conference on Machine Learning (ICML)*, pages 1257–1264, 2009.