Semi-supervised collective matrix factorization for topic detection and document clustering

Ye Wang  
Centre for Applied Informatics  
Victoria University  
Melbourne, Victoria, Australia 8001  
College of Computer  
National University of  
Defense Technology  
Changsha, Huana, P.R.China 410073  
Email: ye.wang10@live.vu.edu.au

Yanchun Zhang  
Centre for Applied Informatics  
Victoria University  
Melbourne, Victoria, Australia 8001  
Email: Yanchun.Zhang@vu.edu.au

Bin Zhou  
and Yan Jia  
College of Computer  
National University of  
Defense Technology  
Changsha, Huana, P.R.China 410073  
Email: yanchun.zhang@vu.edu.au

Abstract—Topic detection and tracking (TDT) under modern media circumstances has been dramatically innovated with the ever-changing social network and inconspicuous connections among participants in the internet communities. Apart from the inherent word features of analysing materials, such as news articles and personal or professional comments, incidental information attracts increasing attention from the research community. Meanwhile, numerous interrelations hiding in the propagated articles and network participants also promote the transfer and evolution of topics, not only apparent connections, for example having the same tags and belonging to the same party, but also weak connections which are complicated and with little causal relations. Therefore, answering the question how to exploit and use this hidden information in the social network will extend the landscape of research on TDT. In this paper, we employ the followers’ groups extracted from Twitter as the social context that accompanied the corresponding news articles and explore the interior links among them to develop a non-negative factorization methods with semi-supervised information derived from the original data. Furthermore, experiments are conducted on real and semi-synthetic data sets to test the performance of topic detection and documents clustering. The results demonstrate that the proposed method outperforms several state-of-the-art methods.

I. INTRODUCTION

Topic detection and tracking (TDT) is no doubt a well-studied research field under the circumstance of information overload, as well as the rise of new media, for example social media and we media. The various platforms of the latter, like the online social networking services, have provided the means that people can find things out with their own version of the truth for themselves and share their views instantly, which dramatically changed the way how society is informed. However, as the classic 80-20 rule was widely observed in many fields, not just business and management, traditional medium still reigns supreme, spreading the so-called most powerful ideas and only the perspectives they want people to know to the gullible public. In spite of this, the new media actually innovate the information transmission on width and depth, as well as the speed. Therefore, the research of TDT gradually attempts to combine text analysis and social link together to robustly design topic models. Many benefits come with this progress. First of all, it could alleviate some of the multiple problems caused by the lexical variation, lexical sophistication and lexical errors people used to describe a particular event or idea to some extent. And secondly, it would be a more realistic version of the analysis for text content and social context, which mutually restricts and supports each other. In addition to these, it is not in contradiction with the research of community detection and social links mining in social media, but enhance each other and open a new avenue for both of them.

Most of the traditional content-based topic discovery methods, including matrix decomposition based methods [1]–[3] and probabilistic approaches [4]–[6], mainly focus on textual content mining for latent topics, rather than other incidental information, such as the geography location, time and user related records. Nevertheless, the real implication of words is highly dependent on the context, in which they occur, not only because of the complexity of vocabulary, but also because of the various expressing forms. Different time and circumstances may give spectators different perspectives even for the same article. For example, we will have absolutely different thoughts to Donald Trump’s profiles in the past year and several years ago. An article of Trump’s profiles was merely a matter of celebrity resume several years ago. Maybe it was a little piece in a stack of commercial documents in five years ago, which might involve corporate development, investment, financing and so on. While seeing his profile in the last year, the most likely topic was about the US election under America’s political problems and it has been more of a national political issue even concern to other countries around the world after his election victory. That is to say, his profile plays a role in affecting the readers’ political stands, rather than a normal description of a celebrity nowadays. We could observe similar changes in many other occasions, most of which are always ignored by many classical topic discovery algorithms.

Normally, the internal coherence of user’s preference in a particular period helps us to find groups of people that
share same common interests and topics. Based on this "issue of common concern" view, Kalyanam etc. [7] assumes that the latent topics can also be expressed by users’ distribution apart from textual contents. Hence, by combining these two parts, their method works out more precise topics. But in the meantime, the algorithm complexity increases dramatically as the amount of users becomes very huge. The distribution of users plays a defining role, which is always changing with the arising, evolving and defusing of topics over time. Moreover, the total amount of users is immense, comparing with the quantity of textual features caused by the enormous volume of participants the online social network brings. In a twist, the previous hypothesis can be understood as the perspective of "users with common shares". In other words, the distribution of posts and shares can also represent users’ preferences. In this paper, we will use two matrices to denote the relational information of the textual content and the social context, then factorize each of them with a generalized function, bonding factors from different matrices together. To associate above two parts, we propose a novel non-negative matrix factorization (NMF) approach based on collective matrix factorization approach [8]. To our knowledge, existing NMF algorithms in the domain of topic discovery are unsupervised learning which lack features to exploit implicit associations inside of the data matrix. To retrieve and leverage these relations, we consider the constraint propagation, which has applied to spectral [9] and validate on image clustering [10] recently.

For the research of TDT, it is infeasible to utilise either fully labelled training set or a part of labelled data as the works in image processing and face recognition [9]–[11] did, not due to the expensive cost [16], but because the topic labels, which is known as “hard” or inherent constraints, is the target that we seek in its own right. This implies that all the data points we have are unconstrained with respect to each other in the initialization phase of textual content and social context. In this paper, to overcome this unfavourable precondition, we propose a constraint propagation method to explore the pairwise constraints within data points which are perceived as “weak” or potential constraints that may be influenced by certain circumstances. We construct two weight matrices respectively on the initial data matrices’ local structures in the first step and propagate the restriction between two data points over the whole matrix in both vertical and horizontal directions until convergence. New weight matrices will be developed later with the pairwise constraints for each original data matrices and will be applied as regularization terms to preserve and consolidate the geometrical structure in the low-dimensional representation space in the following collective NMF step. Our contributions are as follows:

- When discover topics in online text streams, it is inevitable that incidental information, social context in particular, be considered as a crucial factor because of the changes and challenges the online social media brings. We propose a collective NMF method for multi-domain to comply with this trend. Experimental results show that our method can improve the topic detection and documents clustering performance greatly.
- We leverage the constraint propagation scheme to adequately exploit the hidden supervised information in both text content matrix and social context matrix. The supervised information will helps preserve the geometric structure in low-dimensional representation space during the optimisation of the collective NMF objective function.
- Differing from other so called collective NMF methods, for example LETCS [7] and CMFH [12], which factorize the joint matrices as an extra feature to the same domain and ignore the fact that factors in the matrix of the social context are practically impossible complete at the beginning, our method capable of processing cases with changing and evolving groups of users.

The rest of this paper is organised as follows. Section 2 briefly reviews the previous work on topic detection and tracking. We give a preliminary definition of our model in section 3. Section 4 explains our constraints propagation scheme in detail. Section 5 proposes our multi-domains NMF with constraints propagation methods with the updating rules and computational complexity analysis. Section 6 shows the experimental results and section 7 concludes the paper.

II. Previous work

Topic detection and tracking has been a fundamental problem in a wide range of applications, such as news event analysis [13], information discovery [14], [15], social interest discovery [16], [17], social emotion learning [18] and expert system [19]. Though methods of TDT in above applications for different domains have been paid particular attentions to, nearly all of them are developed on the basis of latent topic models which represented by pLSA (probabilistic latent semantics analysis [20]) and LDA (latent dirichlet allocation [5]). Latent topic model, generally speaking, aims to model representations to indicate the latent variables, topics in particular, in underlying structure of discrete data. In this paper, we primarily concerned with two roughly categories of topic modelling approaches. One is probabilistic model based approaches [5], [6] and another one NMF-based approaches [7], [21].

Comparing to the basic latent semantic analysis (LSA) [1], pLSA [4] model each topics as multinomial distributions over words, which extend the SVD based LSA with the statistical distribution. LDA [5] developed pLSA with a Dirichlet distribution on the distribution of topics for each article and words distribution to this topic is multinomial distribution. To exploit the link structure among documents, PHITS [22] extend the pLSA by defining a generative process for citations as hyperlinks. Erosheva et al. propose a similar idea Link-LDA [23]. Nallapati et al. [24] developed pairwise Link-LDA and Link-PLSA-LDA by emphasizing directional citations and asymmetric citing links. For dynamic features in data streams, Blei et al. further extended LDA to Dynamic Topic Model (DTM) [25] which is a representation of discrete dynamic topic model (dDTM). In the counterpart of dDTM, known as continuous time dynamic topic model (cDTM),
Wang et al. [26] employed Brownian motion to model continuous evolution topics in sequential time-series data. With the prevalent trend of social media, for example posts and comments on Twitter, some works [27], [28] are proposed to overcome challenges including text shortness, low meaningful description and high velocity stream. Twevent [29] is a feature-pivot methods, in which tweet segments are used to detect and cluster events.

Non-negative matrix factorization (NMF) [30] has been a mainstream method of part-based representation for research communities of information retrieval, pattern recognition and computer vision. Previous works [31], [32] have demonstrated the close connection between NMF and pLSA. Xu [33] proposed a NMF-based document clustering method resorting to LSA [1] idea, in which it does not require the decomposed matrices are orthogonal and it guarantees that feature values of all documents are non-negative in all latent semantic directions and each document can be presented as additive combination of the base latent topics. Cao et al. proposed Online-NMF in [2] to detect and track the moving of latent factors in data streams, considering multiple topics cooccurrence. A quite similar optimization strategy to NMF was used as dictionary learning in [34], which identifies the incoming observation documents as a novel one or not with \( \ell_1 \)-penalty to measure reconstruction error and new dictionary will be learnt with novel documents. Other NMF-based dynamic approaches [35], [36] were proposed to capture the evolving set with temporal regularization terms. Recently, Suh et al. [21] impose ensemble model on NMF-based method to discover more precise local topics under noisy circumstance.

III. PRELIMINARY

Tweets always continue to show as a flow, including meta data with either various topics or the daily-life situations. We assume that a constant stream of news articles and the batch of articles at each time step form as a non-negative matrix \( X = [X_1, X_2, \cdots, X_N] \in \mathbb{R}^{M \times N_d} \), consisting of \( N_d \) articles and \( M \) texture features (terms in bag-of-words model). Each column of \( X \) is an \( M \)-dimensional data point. To associate with the social context, which is the user preference in our previous definition, we introduce a non-negative matrix \( U = [U_1, U_2, \cdots, U_N] \in \mathbb{R}^{N_d \times N_u} \), where \( N_u \) users are activated in this time slot and \( N_d \) articles are involved as features. By definition, each column in matrix \( U \) is also a data point, but the preference feature is \( N_d \)-dimension.

Non-negative matrix factorization [30] is an algorithm that decomposes the original non-negative data matrix \( X \in \mathbb{R}^{M \times N} \) to two non-negative matrices \( H \in \mathbb{R}^{M \times K} \) and \( V \in \mathbb{R}^{N \times K} \), whose linear product approximates as close to the original matrix as possible:

\[
X \approx HV^T
\]

\[
\arg\min_{H, V \geq 0} f = \mathcal{D}(X, HV^T) + \mathcal{R}(H, V)
\]

Where, loss function \( \mathcal{D}(X, HV^T) \) quantifies the cost of the approximation and \( \mathcal{R}(H, V) \) is the regularization penalty. From the definition, we can always find two entities and a relation between them. The NMF process, in essence, is a linear regression, which determines the relation description between the two entities by given that the relation exists.

Collective matrix factorization [8] aims to trade on the correlations between matrices that contain more than one relation and associates the factors involving in different relations together with a generalised-linear link function. Take a two-relation schema as an example: two data matrices \( X \in \mathbb{R}^{M \times N} \) and \( U \in \mathbb{R}^{N \times L} \) we use the factor \( V \) in both constructions and then we have the decomposition and objective function:

\[
X \approx HV^T \text{and } U \approx VG^T
\]

\[
\arg\min_{H, V \geq 0} f = \mu\mathcal{D}(X, HV^T) + (1 - \mu)\mathcal{D}(U, VG^T) + \mathcal{R}(H, V, G)
\]

Where \( \mu \in [0, 1] \) is a trade-off parameter to weight the relative importance between two relations.

IV. CONSTRAINT PROPAGATION

As the NMF is an unsupervised method which optimise the convex objective function to obtain good result, the supervised information of the data set are not being used generally. [10] mentioned that class labels and pairwise constraints are two commonly accepted sources of supervised information for a data set. The former, our target, is what fixed in the internal nature of each data point, while the latter just gives us a weak pair-wise connection of two data points, whether the two points must be linked or not. Since the later can be derived from the former, for example, points with the same label share a must-link connection and otherwise a cannot-link appears between them. However, the pairwise constraints do not afford the inverse deduction with our common sense, which embodies the weakness of the pairwise constraints. On the other hand, the pairwise constraints can be obtained more widely and enhanced by some propagation mechanism. In this section, we explain how the constraints will propagate for data points in the original matrices \( X \) and \( U \) from horizontal and vertical view.

By convention [9], [10], we begin with the construction of the initial pairwise constraints matrices \( Z_{X} = \{Z_{ij} > 0\}_{N_d \times N_d} \) and \( Z_{U} = \{Z_{ij} > 0\}_{N_u \times N_u} \). Because the constraint propagation procedures of matrices \( X \) and \( U \) will follow the same schema, we demonstrate one matrix \( Z \) instead of both of them here for concision and unique definitions will not be made for matrices \( X \) and \( U \) separately in the following. For original \( n \)-dimensional data point-feature matrix, denotes \( C_i \) as the feature set involved by data point and \( Z = \{z_{ij}\}_{n \times n} \) can be initialized as:

\[
z_{ij} = \begin{cases} 
+1, & r > \alpha \\
-1, & r < \beta \\
0, & \text{otherwise.}
\end{cases}
\]

Where \( r = \|C_i \cap C_j\| \). \( \alpha \) and \( \beta \) are the adjustable parameters for deciding whether two data points can be linked by the same
label, in our case, same topic. Jaccard similarity coefficient is used here to measure the similarity between the pair of data points.

We have directly obtained the pairwise constraints with little information loss by far. It can be found that some pairs of data points are not constrained (i.e. \( z_{ij} = 0 \)), that is, the corresponding data points \( x_i \) and \( x_j \) are initially unlabelled. Therefore, the goal is to transductive infer the labels of the unlabelled points \([37]\). Here we denote the propagated pairwise constraints matrix as \( F = \{f_{ij}\}_{n \times n} : |f_{ij}| \leq 1 \). More concretely, \( Z \) is the initial status of \( F \).

The weight matrix \( W = \{w_{ij}\}_{n \times n} \) is used to show the proximity of two data points \( x_i \) and \( x_j \). For document processing tasks in IR community, the dot-product weighting is pretty common \([38]\). Here, we use normalized dot-product, which is also known as cosine similarity of the two vectors to initially define \( W \) as

\[
\begin{align*}
\hat{w}_{ij} &= \begin{cases}
\frac{x_i \cdot x_j}{|x_i||x_j|}, & r \geq \alpha \\
0, & \text{otherwise}.
\end{cases}
\end{align*}
\]

Construct a symmetric matrix \( L = S^{-1/2}WS^{-1/2} \) with a diagonal matrix \( S = \{s_{ij} = \sum_j w_{ij}\}_{n \times n} \). The following iterations are executed from vertical and horizontal perspectives, showing how each data points takes over the information from its neighbours and keep its initial information.

1) For the vertical constraint propagation, iterate

\[
F_v(t) = \delta LF_v(t-1) + (1-\delta)Z
\]

until converge, where the parameter \( \delta \in (0, 1) \) specifies the ratio of information from itself and its neighbours.

2) For the horizontal constraint propagation, iterate

\[
F_h(t) = \delta LF_h(t-1)L + (1-\delta)F_v^*(t),
\]

where \( F_v^* = (1-\delta)(I-\delta L)^{-1}Z \) is the limit of \( \{F_v(t)\} \).

Proof. By previous definition, \( F(0) = F_v(0) = Z \). By the Equation (3), we have

\[
\begin{align*}
F_v(1) &= (\delta L)F_v(0) + (1-\delta)Z \\
F_v(t) &= (\delta L)^tZ + (1-\delta)\sum_{i=0}^{t-1}(\delta L)^iZ.
\end{align*}
\]

Since \( 0 < \delta < 1 \) and the eigenvalues of \( L \) in \([-1, 1]\), according to the principle of infinite geometric series, we have

\[
\lim_{t \to \infty} (\delta L)^t = 0 \quad \text{and} \quad \lim_{t \to \infty} \sum_{i=0}^{t-1}(\delta L)^i = (I - \delta L)^{-1}
\]

Therefore, \( F_v(t) \) converges to

\[
F_v^* = (1-\delta)(I-\delta L)^{-1}Z
\]

\( \blacksquare \)

3) Denote \( F^* = F_h^* \) is the final representation of the propagated pairwise constraints, where \( F_h^* = (1-\delta)F_v^*(I - \delta L) \) is the limit of \( \{F_v(t)\} \).

\[
\begin{align*}
\tilde{F}_{ij} &= \begin{cases}
1 - (1 - f_{ij}^*)(1 - w_{ij}), & f_{ij}^* \geq 0 \\
(1 + f_{ij}^*)w_{ij}, & f_{ij}^* < 0,
\end{cases}
\end{align*}
\]

The procedures of constraint propagation for text context and user preference matrices are summarised in Algorithm 1.

**Algorithm 1 Constraint Propagation for Context and Users**

**Input:**

Article matrix \( X \in \mathbb{R}^{M \times N_d} \), user preference matrix \( U \in \mathbb{R}^{N_u \times N_u} \), parameter \( \delta \).

**Output:**

Weight matrices \( \tilde{W}_X \in \mathbb{R}^{M \times N_d} \), \( \tilde{W}_U \in \mathbb{R}^{N_u \times N_u} \).

1. Construct the initial pairwise constraints matrices \( Z_X \) and \( Z_U \) by Equation (5).
2. Initialize propagated pairwise constraints matrices \( F_X \) and \( F_U \) with \( Z_X \) and \( Z_U \) correspondingly.
3. Define weight matrices \( W_X \) and \( W_U \) by Equation (6).
4. Construct symmetric matrices \( L_X = S_1^{-1/2}WS_1^{-1/2} \) and \( L_U = S_2^{-1/2}WS_2^{-1/2} \), where diagonal matrix \( S_1 = \{\sum_j w_{ij}\}_{n_d \times n_d} \) and \( S_2 = \{\sum_j w_{ij}\}_{n_u \times n_u} \).
5. Update \( \tilde{F}_X \) and \( \tilde{F}_U \) from vertical and horizontal perspectives by the limit \( F^* \) in Equation (11).
6. Compute New weight matrices \( \tilde{W}_X \) and \( \tilde{W}_U \) by Equation (12).

In the following steps, \( \tilde{W} \) will be used for constrained non-negative matrix factorization. From the algorithm we designed above, \( \tilde{W} \) shows the following nice properties:

1) \( \tilde{W} \) is symmetric and non-negative.

Proof. The symmetric feature is inherited from the symmetric of both matrix \( F^* \) and \( W \). By the definition, \( w_{ij} \in [0, 1] \) and \( |f_{ij}^*| \leq 1 \), then we have

\[
\begin{align*}
\tilde{w}_{ij} &= \begin{cases}
1 - (1 - f_{ij}^*)(1 - w_{ij}), & f_{ij}^* \geq 0 \\
(1 + f_{ij}^*)w_{ij}, & f_{ij}^* < 0
\end{cases}
\end{align*}
\]

\( \blacksquare \)
2) \( \tilde{W} \) is adjusted by \( F^* \) with no distinction between \( f_{ij}^* \geq 0 \) and \( f_{ij}^* < 0 \).

\textbf{Proof.} \( \tilde{w}_{ij} \) is differentiable at \( w_{ij} = 0 \) and for all \( f_{ij}^* \),
\[ \frac{d\tilde{w}_{ij}}{dw_{ij}} = 1 - |f_{ij}^*|, \]
\( \Box \)

3) \( \tilde{W} \) shows that the pairwise constraint between two data points has been reinforced after propagation.

\textbf{Proof.} The Equation (8) is a monotonically increasing function of \( f_{ij}^* \). Therefore, the new weight matrix \( \tilde{W} \) increases \( \tilde{w}_{ij} \geq w_{ij} \) with \( f_{ij}^* \geq 0 \) and decreases \( \tilde{w}_{ij} < w_{ij} \) with \( f_{ij}^* < 0 \).

\( \Box \)

As mentioned before, more potential pairwise information gained from the constraint propagation procedure is incorporated into new weight matrices \( \tilde{W}_X \in \mathbb{R}^{N_d \times N_d} \) and \( \tilde{W}_U \in \mathbb{R}^{N_u \times N_u} \). The properties above guarantee that data points sharing the same topics have relatively larger associated scores and vice versa. Next, we will perform non-negative matrix factorization with weight matrices.

V. \textbf{NMF_CP PROCESS FOR USER PREFERENCE AND TEXT CONTENT MATRICES}

According to the traditional non-negative matrix factorization, the original non-negative data matrix will be decomposed into two non-negative matrices, whose linear product approximates as accurate to the original matrix as possible. In our case, we decompose the articles matrix \( X \) and the user preference matrix \( U \) in terms of latent topics. We firstly set the number of topics as \( k \), which is usually smaller than \( N_d \), \( N_u \) and \( M \). Then, we have:

\[
X \approx HV^T \quad \text{s.t.} \quad H, V \geq 0 \quad (13)
\]

\[
U \approx VG^T \quad \text{s.t.} \quad V, G \geq 0 \quad (14)
\]

Where, \( H = [H_{ij}] \in \mathbb{R}^{M \times K} \) is a topic matrix. Each column \( H_{ij} \) represents a latent topic expressed as a combination of several terms. The matrix \( V \) can be regarded as the low dimensional representation of \( X \) under the new basis \( H \), that is how each article is arranged in terms of the latent topics discovered in \( H \). The decomposition of matrix \( U \) is sharing the same \( V \) to fulfill our assumption that users will only be concerned with the news what interest them. Therefore, Equation (10) illustrates how users are grouped in terms of the articles having the same latent topics. Similar to the role of matrix \( V \) in Equation (9), \( G = [G_{ij}] \in \mathbb{R}^{N_u \times K} \) can be regarded as the low dimensional representation of \( U \) with respect to the new basis \( V \) and each column in \( G \) represents a community consisting of users who have the same interest at this moment.

Here, we choose Euclidean distance to measure the dissimilarity of data points under the lower dimensional representation \( V \) and \( G \). Combining with the constraint propagation information, we have the following enhanced distance terms:

\[
\phi(V) = \frac{1}{2} \sum_{i,j=1}^{N_d} \| V_i - V_j \|^2 \cdot \tilde{W}_U X_{ij} \quad (15)
\]

\[
\phi(G) = \frac{1}{2} \sum_{i,j=1}^{N_u} \| G_i - G_j \|^2 \cdot \tilde{W}_U X_{ij} \quad (16)
\]

The properties of weight matrices \( \tilde{W}_X \) and \( \tilde{W}_U \) ensure that in low dimensional representation, \( \phi(V) \) and \( \phi(G) \) can draw similar data points closer and distance data points belong to different topics in geometric space which consistent with the intrinsic relationships of data points in original data matrix.

A. \textbf{The Objective Function}

To maximum the approximate decomposition, our optimization problem is as follow:

\[
\min_{V,H,G} f(V,H,G) = \mu(\|X - HV^T\|^2_F + \lambda_1 \phi(V)) + (1 - \mu)(\|U - VG^T\|^2_F + \lambda_2 \phi(G)) + \mathcal{R}
\]

\text{s.t.} \quad \mathbb{V}, \mathbb{H}, \mathbb{G} \geq 0 \quad (17)

Here, we use 11-norm based regularization \( \mathcal{R} = \gamma_1 \|V\|_1 + \gamma_2 \|U\|_1 + \gamma_3 \|H\|_1 \) to promote the sparsity. The regularization parameters \( \lambda_1 \) and \( \lambda_2 \) controls the contribution proportions of the supervised information parts in our objective function. Parameter \( \mu \in [0, 1] \) controls the balance between text content part and social context part. If \( \mu = 0 \), the text content would be ignored and only users be grouped together with topics.

B. \textbf{Updating rules}

The objective function \( f(V,H,G) \) is not convex in all the variables together. Therefore, we turn to find its local minima instead of the global minima with the classical multiplicative updating rules [30]. For ease of layout, we omit the parameter \( \mu \) in the following formulas. We first rewrite it as follow:

\[
f(V,H,G) = \|X - HV^T\|^2_F + \|U - VG^T\|^2_F
\]

\[
+ \frac{\lambda_1}{2} \sum_{i,j=1}^{N_d} \| V_i - V_j \|^2 \cdot \tilde{W}_X X_{ij}
\]

\[
+ \frac{\lambda_2}{2} \sum_{i,j=1}^{N_u} \| G_i - G_j \|^2 \cdot \tilde{W}_U U_{ij} + \mathcal{R}
\]

\[
= \|X - HV^T\|^2_F + \|U - VG^T\|^2_F
\]

\[
+ \lambda_1 \sum_{i=1}^{N_d} V_i^T V_i D_i V_i - \lambda_1 \sum_{i,j=1}^{N_d} V_i^T V_j D_{Xij} X_{ij}
\]

\[
+ \lambda_2 \sum_{i=1}^{N_u} G_i^T G_i D_{Gii} - \lambda_2 \sum_{i,j=1}^{N_u} G_i^T G_j D_{Uij}
\]

\[
+ \mathcal{R}
\]

\[
= \text{Tr}((X - HV^T)(X - HV^T)^T)
\]

\[
+ \text{Tr}((U - VG^T)(U - VG^T)^T)
\]

\[
+ \lambda_1 \text{Tr}(V^T \tilde{D} V V) - \lambda_1 \text{Tr}(V^T \tilde{W} XV)
\]

\[
+ \lambda_2 \text{Tr}(G^T \tilde{D} G G) - \lambda_2 \text{Tr}(G^T \tilde{W} U G)
\]

\[
+ \mathcal{R} \quad (18)
\]
Here, $\text{Tr}(\cdot)$ is the trace of a matrix and $\hat{D}_V$ and $\hat{D}_G$ are diagonal matrices whose diagonal elements $D_{Vii} = \sum_{j=1}^{N_d} \hat{W}_{Xij}$ and $D_{Gii} = \sum_{j=1}^{N_u} \hat{W}_{Uij}$. Denote $\hat{L}_X = \hat{D}_V - \hat{W}_X$ and $\hat{L}_U = \hat{D}_G - \hat{W}_U$, which are symmetric matrices. Then, the objective function can be shown as:

$$f(V, H, G) = \text{Tr}(XX^T) - 2\text{Tr}(XVHG^T) + \text{Tr}(HV^TVH^T) + \text{Tr}(UU^T) - 2\text{Tr}(UGV^T) + \text{Tr}(VG^TVG^T) + \lambda_1\text{Tr}(V^T\hat{L}_XV) + \lambda_2\text{Tr}(G^T\hat{L}_UG) + R$$

This variation applied properties $\text{Tr}(A + B) = \text{Tr}(A) + \text{Tr}(B)$, $\text{Tr}(A) = \text{Tr}(A^T)$ and $\text{Tr}(BA^T) = \text{Tr}(AB^T)$. With the Karush Kuhn Tucker condition, we have the primary feasibility $V \geq 0, H \geq 0$ and $G \geq 0$. We define the Lagrangian as:

$$\mathcal{L}(V, H, G, \Psi, \Phi, \Omega) = f(V, H, G) + \text{Tr}(\Psi V^T) + \text{Tr}(\Phi H^T) + \text{Tr}(\Omega G^T)$$

Let $\Psi, \Phi$ and $\Omega$ be Lagrange multiplier matrices and their elements $\psi_{ij}, \phi_{ij}$ and $\omega_{ij}$ are the Lagrange multipliers for constraints $V_{ij} \geq 0, H_{ij} \geq 0$ and $G_{ij} \geq 0$ respectively. The partial derivatives of the objective function in Equation 15 with respect to each variable are:

$$\frac{\partial \mathcal{L}}{\partial V} = -2X^TH + 2VHG + 2UG + 2GV^TG + 2\lambda_1\hat{L}_XV + \Psi + \gamma_1\epsilon^Te$$

$$\frac{\partial f}{\partial H} = -2XV + 2HV^TV + \Phi + \gamma_2\epsilon^Te$$

$$\frac{\partial f}{\partial G} = -2UV + 2GV^TV + 2\lambda_2\hat{L}_UG + \Omega + \gamma_3\epsilon^Te$$

Using the complementary slackness: $\Psi V = 0, \Phi H = 0$ and $\Omega G = 0$, the update equations are derived as follows:

$$V \leftarrow V \odot \frac{X^TH + UG + \lambda_1\hat{W}_XV - \gamma_1\epsilon^Te}{V(H^TH + G^TG) + \lambda_1\hat{D}_VV}$$

$$H \leftarrow H \odot \frac{XV - \gamma_2\epsilon^Te}{HV^TV}$$

$$G \leftarrow G \odot \frac{UG + \lambda_2\hat{W}_UG - \gamma_3\epsilon^Te}{GV^TV + \lambda_2\hat{D}_GG}$$

Obviously, when $\lambda_1, \lambda_2 \to 0$, the above updating rules reduce to the updating rules in origin NMF [30]. From the proofs in [30], [36], [38], we know that the objective function in Equation 15 is non-increasing above updating rules and the Lagrangian $\mathcal{L}$ is invariant under these updating rules if and only if $V, H$ and $G$ are at a stationary point of the function.

The algorithm is summarized in Algorithm 2.

### C. Computational Complexity Analysis

Next, we analyse the computational complexity of our algorithm NMF_CP comparing with the origin NMF. Intuitively, we assume the multiplicative update iteration stops at time $t$, the overall cost of the origin NMF is $O(tMNK)$, where the input data matrix is an $M$ dimensional matrix with $N$ data points and $K$ is set as the number of latent factors. In our case, the multiplicative updates cost for non-negative factorization is $O(tMN^2d + tN^2NdK)$. Before this, the NMF_CP needs $O(N^2dM + N^2d^2)$ to construct the weight matrices for $X$ and $U$, respectively. Therefore, the maximum overall cost for NMF_CP is $O(MN^2K + tN^2dK + N^2d^2M + N^2d^2Nd)$ .

### VI. EXPERIMENTS

We conduct the experiments focusing on two tasks: detecting the on-going topics and clustering the documents to corresponding topic. Four evaluation metrics are employed from these two perspectives, which are Normalized Discounted Cumulative Gain, Mean Average Precision, Accuracy and Normalized Mutual Information. We compare our proposed method with five other NMF-related algorithms, i.e., NMF [30], GNMF [38], JPP [36], LETCS [7] and CNMF [8] on two types data sets.

#### A. Data sets

To evaluate the effectiveness of introducing the social content information as collective NMF and applying constraint propagated weight on unsupervised NMF, we select two types of data sets. The first one is provided by [7] consisting of all the articles published by 80 international news sources in a period of 14 days in April, 2013 and a list of all tweets which are linked to each articles within 12 hours after the corresponding article’s publication. The hashtags (＃) quoted by tweets were treated as ground truth topics of the documents which were associated with those tweets and the links between tweets and articles were used to construct the social context matrix. In our experiments, we selected two categories hashtags 5 topics which are defined as: 1) Content-stable hashtags are those that did not evolve too much in terms of text content, but keep attracting varied attention during the period of collecting; 2) Community-stable hashtags are relatively stable for their community, but the real events they referring to vary a lot. For example, #WorldCup and #GRAMMYs. We will abbreviate them to TS and CS in the following experiments.

#### Algorithm 2 NMF_CP algorithm

**Input:**

- Article matrix $X \in R^{M \times N_d}$, user preference matrix $U \in R^{Nd \times N_u}$, weight matrices $\hat{W}_X \in R^{Nd \times Nd}$, $\hat{W}_U \in R^{Nd \times Nu}$, $1 \leq k \leq \min(N_d, N_u, M)$, regularization parameters $\lambda_1, \lambda_2, \gamma_1, \gamma_2$ and $\gamma_3$, trade-off parameter $\mu$.

**Output:**

- $H \in R^{M \times K}$, $V \in R^{Nd \times K}$, $G \in R^{Nd \times K}$.

1. Initialize $H, V, G$ by random matrices.
2. Construct weight matrices $\hat{W}_X$, $\hat{W}_U$ by using Algorithm 1.
3. **repeat**
   4. Fix $H$ and $G$, update $V$ by Equation (24);
   5. Fix $V$ and $G$, update $H$ by Equation (25);
   6. Fix $H$ and $V$, update $G$ by Equation (26);
7. **until** Coverage.
The second is a serious of semi-synthetic data sets generated from NIST Topic Detection and Tracking (TDT2) text corpus. The full TDT2 corpus is composed of data collected from 6 news sources during 180 days of 1998. Here, we use a short version provided by Cai\textsuperscript{2}, keeping documents labelled with only one topic. We chose the given topic number 5,7,10,15,20 and 25 and for each of them, generated 20 random non-repeat datasets to tests with algorithms NMF, GNMF and our NMFCP. Since the continuous streaming feature does not exist on generated synthetic datasets, JPP and LETCS are not suitable and without social context domain, CMF reduced to NMF.

B. Evaluation metrics

To evaluate the algorithms’ capability of detecting the ongoing topics, we use Normalized Discounted Cumulative Gain (NDCG) and Mean Average Precision (MAP) as metrics. We use the top 10 and ranking words as the relevant words to express the ground truth topics as well as the topics obtained by the algorithms, mapping the latter to the former with the cosine similarity and setting the ground truth relevance values as binary values. For topics clustering performance, we use Accuracy and Mutual Information with a k-means clustering algorithm applying on the returned documents-words matrix to compare with ground truth topic clusters. The ground truth topic of rth article $V_{true}(i)$ is extracted from hashtags appearing in the associated tweets and the ground truth topics-words distribution $H_{true}$ is calculated by $H_{true} = XV_{true}$.

Normalized discounted cumulative gain (NDCG): Similar to the assumption of highly relevant documents, we have the following assumption:

**Assumption 1:** Highly relevant words are more useful when ranking higher in the topics-words distribution.

**Assumption 2:** The relevance scores of relevant words are set as equivalent value 1.

The NDCG is defined as:

$$NDCG(H, H_{true}) = \frac{\sum_{i=1}^{k} DCG_k}{IDCG_k}$$

Here, $k$ is the number of latent topics. $DCG_k = \sum_{j=1}^{r} \frac{rel_j}{\log(1+j)}$ is the discounted cumulative gain of the kth topic. $rel_j$ is the relevant score of jth relevant word and $r$ is the number of relevant words for the kth topic detected by algorithms. $IDCG_k = \sum_{j=1}^{R} \frac{\frac{rel_j}{\log(1+j)}}{\max(rel_j)}$ is the ideal DCG, R is the number of relevant words in ground truth. By the top 10 ranking word representing method, we have $r \leq R = 10$. A higher NDCG score indicates a closer approximation to the ground truth.

**Mean average precision (MAP)** calculate the mean of average precision (AP) for all the topics which reflect the algorithms global performance.

$$AP_i = \frac{\sum_{j=1}^{R} \frac{j}{\text{rank}_j}}{R}$$

, where $R = 10$ is the number of relevant words for the topic in ground truth and rank$_j$ is the sequence number of relevant word j in retrieved words list. $\frac{1}{\text{rank}_j} = 0$ if relevant word j is not retrieved by algorithms in topic i. The formula of mean average precision is shown as follows:

$$MAP = \frac{\sum_{i=1}^{k} AP_i}{k}.$$ 

The metrics Accuracy and Normalized Mutual Information is used to evaluate the clustering performance of documents. We compare the documents-words distribution matrix $V$ obtained by algorithms to the real distribution, ground truth $V_{true}$.

Before testing the clustering performance, a k-means clustering algorithm is applied to each data point in $V$ (that is, the article $d_i \in N_d$) to get the exact cluster label $l_i$ of $d_i$. The ground truth label is $l_{gnd}$ given by $V_{true}$. To some extent, the results would depend on the clustering results.

**Accuracy (AC)** simply measures the percentage of correct labels for all the articles. It defines as:

$$AC = \frac{\sum_{i=1}^{N_d} \delta(l_i, l_{gnd})}{N_d}$$

$\delta(l_i, l_{gnd})$ is the delta function that equals to 1 if $l_i = l_{gnd}$ and equals to 0 otherwise. $N_d$ is the total number of data points.

**Normalized Mutual Information (NMI):** Mutual information of two variables $X$ and $Y$ describes the mutual dependence between the two variables, for example in our case, the cluster labels $l$ of data points in $V$ and the ground truth cluster labels $l_{gnd}$ of $V_{true}$. More specifically, it is the average reduction of the uncertainty about the label of a data point in $V$ when knowing its ground truth label in $V_{true}$. Normally, the uncertainty of a set of clusters $C = \{c_1, c_2, ..., c_k\}$ corresponding to $X$ is represented by entropy as $H(X) = -\sum_{c_i \in C} P(c_i) \log P(c_i) = -\sum_{c_i \in C} \frac{n_i}{N} \log \frac{n_i}{N}$.

Analogously, a set of cluster $C' = \{c'_1, c'_2, ..., c'_k\}$ of another variable $Y$ has the similar entropy $H(Y)$ to quantify its uncertainty. Sometimes they are also written as $H(C)$ and $H(C')$. Intuitively, the value of the mutual information of these two variables $MI(X, Y) \in [0, \min(H(X), H(Y))]$ with the following formula:

$$MI(X, Y) = \sum_k \sum_{c_i \in C} P(c_i, c'_j) \log_2 \frac{P(i, j)}{P(c_i)P(c'_j)} = \sum_k \sum_{c_i \in C} \frac{|c_i \cap c'_j|}{N_d} \log_2 \frac{|c_i \cap c'_j|}{|c'_j|}$$

Where $P(c_i) = \frac{|c_i|}{N_d}$ or $P(c'_j) = \frac{|c'_j|}{N_d}$ denote the probabilities that a data point belongs to cluster $c_i$ in $C$ or $c'_j$ in $C'$. $P(i, j) = \frac{|c_i \cap c'_j|}{N_d}$ denotes the joint probability distribution that a data point belongs to cluster $c_i$ in $C$ and to $c'_j$ in $C'$. As $MI(X, Y)$ is bounded by their entropies, rather than a constant value, it is necessary to use normalized mutual information for easy comparison between the results of different variable pairs.

C. Detection results

The detection results are examined by the matrices Normalized discounted cumulative gain and Mean average precision. We set $\delta = 0.2$ and regularization parameters $\lambda_1 = \lambda_2 = 10^3$ for our NMFCP algorithm. Similar, we set regularization parameter $\lambda = 10^3$ and neighbour $p = 5$ for geometric structure of GNMF. The left two lists in Table I shows the detection results on TDT2 data set. Our proposed method almost outperforms other two algorithms on all NDCG and MAP value. The comparisons with NMF and GNMF indicate that the constraint propagation on the potential links of data points is effect, which will be discussed in the later section. What is interesting is that, the origin NMF algorithm performs well for some case, even better than GNMF. This is consistent with some of the experiments results in [10], [11] with image datasets. The figures below also show that NMFCP algorithm can get better results than others in most cases...

---

2 http://www.cad.zju.edu.cn/home/dengcai/Data/TextData.html
TABLE I
DETECTION AND CLUSTERING PERFORMANCE

<table>
<thead>
<tr>
<th>#Topic</th>
<th>NDCG</th>
<th>MAP</th>
<th>AC</th>
<th>NMI</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>NMF</td>
<td>GNMF</td>
<td>NMFCP</td>
<td>NMF</td>
</tr>
<tr>
<td>5</td>
<td>0.746</td>
<td>0.777</td>
<td>0.806</td>
<td>0.765</td>
</tr>
<tr>
<td>7</td>
<td>0.716</td>
<td>0.731</td>
<td>0.769</td>
<td>0.725</td>
</tr>
<tr>
<td>10</td>
<td>0.636</td>
<td>0.676</td>
<td>0.679</td>
<td>0.633</td>
</tr>
<tr>
<td>15</td>
<td>0.588</td>
<td>0.584</td>
<td>0.592</td>
<td>0.571</td>
</tr>
<tr>
<td>20</td>
<td>0.54</td>
<td>0.518</td>
<td>0.556</td>
<td>0.506</td>
</tr>
<tr>
<td>25</td>
<td>0.472</td>
<td>0.493</td>
<td>0.495</td>
<td>0.423</td>
</tr>
</tbody>
</table>

(a) TS

(b) CS

Fig. 1. The NDCG performance versus parameter \( \mu \)

Fig. 2. The MAP performance versus parameter \( \mu \)

Fig. 3. The AC performance versus parameter \( \mu \)

Fig. 4. The NMI performance versus parameter \( \mu \)

Fig. 5. The NDCG performance versus parameter \( \lambda_1 \)

Fig. 6. The MAP performance versus parameter \( \lambda_1 \)

Fig. 7. The AC performance versus parameter \( \lambda_1 \)

Varying parameters, especially when facing the complicated social context (CS dataset). However, the detection results on TS dataset (e.g. Figs. 1(a) and 2(a)) are always unsatisfactory comparing to GNMF, which is probably due to the interference caused by social context part.

D. Clustering results

The detection results are examined by the matrices Accuracy and Normalized Mutual Information. From Table I, we can see our NMFCP performs almost as well, dramatically better than NMF and GNMF when number of topics below 15. When topic number increases, the accuracy and normalized mutual information decrease on all algorithms as a common trend. On CS dataset, LETCS obtained outstanding clustering results (Figs. 3(b), 4(b), 7(b), 8(b), 11(b), 12(b)) because it treats the whole set of users as a feature set of the documents, which is equivalent of clustering documents with users’ distribution. However, as we mentioned above, this is impractical for real application, data streams in our daily social network in particular. Besides LETCS, algorithms involving pairwise links between data points (i.e. GNMF and NMFCP) outperforms others in almost all cases. The performance varies with parameters will be discussed in the following section. What is interesting is that, LETCS detection performance is relatively low on CS dataset (Figs. 1, 2, 5 and 6), on the contrary, the corresponding clustering results are surprisingly good (Figs. 3(b), 4(b), 7(b) and 8(b)).

E. Parameters selection

Our NMFCP algorithm has three basic parameters. The trade-off parameter \( \mu \) determines the relatively importance of collective domains, therefore it only works in methods LETCS, CNMF and NMFCP. The regularization parameter \( \lambda_1 \) and \( \lambda_2 \) controls how much the supervised information works, corresponding to matrix \( \mathbf{U} \) of algorithms during period of 14 days with varying domains, therefore it only works in methods LETCS, CNMF and NMFCP. The regularization parameter \( \lambda_1 \) and \( \lambda_2 \) controls how much the supervised information works, corresponding to matrix \( \mathbf{U} \) and matrix \( \mathbf{X} \) in shown in 18. Figs. 1-12 show the average performance of algorithms during period of 14 days with varying \( \mu \), \( \lambda_1 \) and \( \lambda_2 \), respectively.

As we can see in Figs. 1 to 4, \( \mu \) influences the results of LETCS, CNMF and NMFCP, even in the content-stable dataset (TS), the social context information helps promote the performance. LETCS is affected most obviously since it is highly related to the social information as features. What is unsatisfactory is that the changing
trends of detection and clustering performance on TS and CS datasets are inconsistent. For example, the best detection performance on CS was achieved for (Figs. 1(b) and 2(b)), but from Figs. 2(b) and 3(b) we can find that the performance degrade in that range. Nevertheless, this inconsistent between detection and clustering results are not unique, the other two collective methods (LETCS and CNMF) show the similar contradiction to some extent.

Figs. 9 to 12 show the performance of NMFCP versus parameter $\lambda_2$, which is unique to NMFCP, controlling how much the supervised information of social context would contributes. It can be seen that the detection performance of NMFCP is stable with respect to its clustering performance, especially on the TS dataset, that indicates $\lambda_2$ influence detection performance slightly. The clustering performance of NMFCP improved as $\lambda_2$ increases in the range of $[10^2, 10^3]$ and keep relatively stable after $\lambda_2$ reaches $10^3$. Therefore, we can also set $\lambda_2$ to $10^3$.

Figs. 5 to 8 show the performance of NMFCP versus parameter $\lambda_1$, which controls how much the supervised information of text content would contributes. It seems that both detection and clustering performances of NMFCP are relatively outstanding when $\lambda_1$ is in the range of $[10^2, 10^3]$ and tends to decrease after reached the peak, except the clustering performance on CS dataset. But comparing to others, the results of $\lambda_1 = 10^3$ are still acceptable. So the value of $\lambda_1$ can be set to $10^3$.

VII. CONCLUSION

In this paper, we present a semi-supervised collective non-negative factorization method for topic detection and tracking, which leverage not only the basic text content, but also the concomitant social context as incidental information to face the ever-changing social network and varied vocabulary challenge. The model firstly applies constraint propagation technique to reveal the interrelations among the data points in each domain and the propagated constraints are incorporated as regularization term to help optimize the NMF objective function. The experimental results demonstrate that the propagation can improve the topic detection and document clustering performance effectively. Although, we conducted our methods on streams dataset and obtained relatively better performances in most cases comparing to other methods, the information about evolution rule has not been involved. This will be our future work.

ACKNOWLEDGEMENT

This work is supported in part by National Key fundamental Research and Development Program of China (No.2013CB329601, No.2013CB329604, No.2013CB329606), National Natural Science Foundation of China (No. 61502517, No.61372191, No. 61572492) and Australia Research Council Project (DP140100841). This work is also funded by the major pre-research project of National University of Defense Technology.
REFERENCES


