Interest Inference via Structure-Constrained Multi-Source Multi-Task Learning

Xuemeng Song[†], Liqiang Nie[†], Luming Zhang[†], Maofu Liu[§], Tat-Seng Chua[†]
[†] National University of Singapore, [§] Wuhan University of Science and Technology
{sxmustc, nieliqiang, zglumg}@gmail.com, liumaofu@wust.edu.cn, chuats@comp.nus.edu.sg

Abstract

User interest inference from social networks is a fundamental problem to many applications. It usually exhibits dual-heterogeneities: a user's interests are complementarily and comprehensively reflected by multiple social networks; interests are inter-correlated in a nonuniform way rather than independent to each other. Although great success has been achieved by previous approaches, few of them consider these dual-heterogeneities simultaneously. In this work, we propose a structure-constrained multi-source multi-task learning scheme to co-regularize the source consistency and the tree-guided task relatedness. Meanwhile, it is able to jointly learn the task-sharing and task-specific features. Comprehensive experiments on a real-world dataset validated our scheme. In addition, we have released our dataset to facilitate the research communities.

1 Introduction

User interest inference is the basis for many applications, such as adaptive E-learning [Abel *et al.*, 2011a] and personalized service [Pennacchiotti and Popescu, 2011]. Take target advertisement as an example. It is naturally to market cosmetics to ladies, whom are keen on beauty. On the other hand, recently we have witnessed many people with diverse interests involving in multiple social networks simultaneously. Such trend has been statistically validated by a survey result: 52% of online adults use multiple social media services[1]. Multiple social networks comprehensively convey users' interests from different view points. For instance, users may update their daily interests in Facebook, follow their interested accounts in Twitter, and ask or answer questions they are interested in Quora. Thus, fusing cues from multiple sources can potentially boost the performance of user interest inference by a large margin.

Inferencing user interests from multiple social networks, however, is non-trivial due to the following reasons. (a) **Source Integration**. Although users' footprints on heterogeneous social networks describe their interests from different views, they should characterize a same interest preference consistently. Therefore, how to effectively and comprehensively fuse them is one tough challenge. (b) **Interest Relatedness Characterization**. Interests are usually not independent but correlated in a nonuniform way. For example, given a set of interests $\mathcal{I} = \{basketball, football, travel, cooking\}$, the relatedness between *basketball* and *football* may be stronger than that between *basketball* and *cooking*. Given that in our dataset, most users who like to play basketball are more likely to spend their spare time on football than cooking. In the context of user interest inference, each interest is usually aligned with one task. Consequently, the second challenge is how to capture and characterize the relatedness among tasks and how to incorporate this into multi-task learning. (c) **Discriminant Feature Selection**. The discrimination of features is different from task to task. Learning task-sharing features and task-specific features effectively is significant to user interest inference. This thus poses another crucial challenge for us.

It is noticeable that there are three lines of researches dedicated to the problem of user interest inference. One is the single source single task learning [Pennacchiotti and Popescu, 2011]. In this context, neither the relatedness among tasks nor the complementary information across sources is explored. Another line of efforts is the multi-task learning [Xue *et al.*, 2007]. They take the task relatedness into account to boost the learning performance and alleviate the problem of insufficient training samples that the traditional single task learning is faced with. It has been observed that learning multiple related tasks simultaneously can improve the modeling accuracy and lead to a better learning performance, especially in cases where only a limited number of positive training samples exist for each task [Fei and Huan, 2013]. The third category of approaches is the multi-source learning [Abel *et al.*, 2011b; 2013]. Instead of sticking to a single source, they propose to aggregate multiple sources to infer users' interests. It should be noted that the last two categories of approaches have the weakness of: existing multi-task learning explores the relatedness among tasks, but overlooks the consistency among different sources of a single task; whereas existing multi-source learning ignores the value of the label information of the other related tasks.

[1]According to Paw Research Internet Project's Social Media Update 2014: http://www.pewinternet.org/.

As an improvement to the existing works, we propose a structure-constrained multi-source multi-task learning (*SM^{2}L*) scheme to infer users' interests. In particular, our scheme jointly regularizes two important aspects. One is the source consistency. The rationale is that interests reflected by different social networks for the same person should be similar, and hence the disagreement among the prediction results should be penalized. The other is the tree-guided task relatedness modeling. Based on prior knowledge, we organize all the tasks (interests) into a tree structure, which can effectively capture various relatedness among tasks. Specifically, the tree structure settles all tasks in leaf nodes and characterizes the relatedness among them by internal nodes. Moreover, the higher level the internal node is located, the weaker is the relatedness imposed on its children tasks. This is accomplished by a tree-guided group lasso regularizer. Meanwhile, *SM^{2}L* learns representative features for individual task and groups of related tasks. A potential benefit of sharing training instances among tasks is that the data scarcity problem can be alleviated. Extensive experiments on a real-world dataset well validated our scheme. We have released our compiled dataset[2], which will facilitate other researchers to repeat our approach and to comparatively verify their own ideas.

2 Related Work

The problem of user interest inference from multiple social networks exhibits dual-heterogeneities: each task (interest) corresponds to features from multiple sources. Towards this end, the most related work lies in the area of multi-view multi-task learning. [He and Lawrence, 2011] proposed a graph-based iterative framework for multi-view multi-task learning (*IteM^2*) in the context of text classification. Given task pairs, *IteM^2* projects them to a new Reproducing Kernel Hilbert Space based upon the common views they share. However, this is a transductive model, which fails to generate predictive models on independent and unknown samples. To deal with the intrinsic trouble of transductive models, [Zhang and Huan, 2012] presented an inductive multi-view multi-task learning model (*regMVMT*). It employs a co-regularization term to achieve model consistency on unlabeled samples from different views. Meanwhile, another regularization function is utilized across multiple tasks to guarantee that the learned models are similar. Noticeably, the implicit assumption that all tasks are uniformly related without prior knowledge might be inappropriate. Realizing this limitation, the authors proposed a revised model (*regMVMT+*) that incorporates a component to automatically infer the task relatedness. As a generalized model of *regMVMT*, an inductive convex shared structure learning algorithm for multi-view multi-task problem (*CSL-MTMV*) was developed in [Jin *et al.*, 2013]. *CSL-MTMV* considers the shared predictive structure among multiple tasks.

Notably, only a limited number of works have been published regarding multi-view multi-task learning and few of them have been applied to user interest inference. Distinguished from these existing methods which maximize the agreement between views using unlabeled data, *SM^{2}L* works towards supervised learning with two advantages: 1) *SM^{2}L* considers source consistency and tree-guided relatedness among tasks simultaneously; 2) *SM^{2}L* allows the learning of task-sharing features and task-specific features using weighted group lasso, where the weights can be learned from prior knowledge.

[2]The compiled dataset is currently publicly accessible via: http://msmt.farbox.com/.

3 User Interest Inference

This section details the proposed *SM^{2}L* scheme for user interest inference.

3.1 Notation

We first introduce the notations throughout this section. We use bold capital letters (e.g. $\mathbf{X}$) and bold lowercase letters (e.g. $\mathbf{x}$) to denote matrices and vectors, respectively. We adopt non-bold letters (e.g. x) to represent scalars, and Greek letters (e.g. λ) as regularization parameters. If not clarified, all vectors are in column forms.

Suppose we have a set of N labeled data samples, $S \geq 2$ sources and $T \geq 2$ tasks. Let D_s denote the number of features extracted from the s-th source. Let $\mathbf{X}_s \in \mathbb{R}^{N \times D_s}$ denote the feature matrix generated from source s, and each row represents a user sample. The feature dimension extracted from all these sources is thus $D = \sum_{s=1}^{S} D_s$. The whole feature matrix can be written as $\mathbf{X} = \{\mathbf{X}_1, \mathbf{X}_2, \cdots, \mathbf{X}_S\} \in \mathbb{R}^{N \times D}$. The label matrix can be represented as $\mathbf{Y} = \{\mathbf{y}_1, \mathbf{y}_2, \cdots, \mathbf{y}_T\} \in R^{N \times T}$, where $\mathbf{y}_t = (y_t^1, y_t^2, \cdots, y_t^N)^T \in R^N$ corresponds to the label vector regarding the t-th task.

3.2 Problem Formulations

For each task, we can learn S predictive models, each of which is generated from one source and defined as follows,

$$\mathbf{f}_{st}(\mathbf{X}_s) = \mathbf{X}_s \mathbf{w}_{st}, \tag{1}$$

where $\mathbf{w}_{st} = (w_{st}^1, w_{st}^2, \cdots, w_{st}^{D_s})^T \in \mathbb{R}^{D_s}$ represents the linear mapping function for the t-th task with respect to the s-th source. The final predictive model for task t can be reinforced via linear combination of these S models. Without the prior knowledge of source confidence, we treat all sources equally as follows,

$$\mathbf{f}_t(\mathbf{X}) = \sum_{s=1}^{S} \frac{1}{S} \mathbf{f}_{st}(\mathbf{X}_s). \tag{2}$$

In multi-class problems, tasks are usually inter-correlated. Multi-source multi-task learning is thus proposed to model their relatedness while seamlessly integrating multiple sources. To select discriminant features, group lasso is considered in the component of multi-task learning. Let $\mathbf{W} = (\mathbf{w}_1, \mathbf{w}_2, \cdots, \mathbf{w}_T) \in \mathbb{R}^{D \times T}$ denote the linear mapping block matrix, where $\mathbf{w}_t = (\mathbf{w}_{1t}^T, \mathbf{w}_{2t}^T, \cdots, \mathbf{w}_{St}^T)^T \in \mathbb{R}^D$. The multi-source multi-task learning with group lasso can be formalized as follows,

$$\mathbf{\Gamma} = \frac{1}{2N} \sum_{t=1}^{T} \left\| \mathbf{y}_t - \sum_{s=1}^{S} \tfrac{1}{S} \mathbf{X}_s \mathbf{w}_{st} \right\|^2 + \frac{\lambda}{2} \sum_{s=1}^{S} \sum_{d=1}^{D_s} \left\| \mathbf{w}_s^d \right\|, \tag{3}$$

where $\mathbf{w}_s^d = (w_{s1}^d, w_{s2}^d, \cdots, w_{sT}^d)$, $\sum_{s=1}^{S}\sum_{d=1}^{D_s}\|\mathbf{w}_s^d\| = \|\mathbf{W}\|_{2,1}$ and λ is the nonnegative regularization parameter that regulates the sparsity of the solution regarding $\mathbf{W}$. When $T \geq 2$, the weights of one feature across all tasks are first grouped by the L_2 norm, and all features are then grouped by the L_1 norm. Thus, the $L_{2,1}$ norm penalty is able to select features based on their strength over all tasks. In this way, we can simultaneously learn the task-sharing features and task-specific features. Obviously, when $T = 1$, this formulation reduces to Lasso [Tibshirani, 1996].

However, the above optimization problem simply assumes that all the tasks share a common set of relevant input features, which might be unrealistic in many real word scenarios. For example, in our work, the tasks "basketball" and "football" tend to share a common set of relevant input features, which are less likely to be useful for the task "cooking". This consideration propels us to assume that the relatedness among different tasks can be characterized by a tree $\mathcal{T}$ with a set of nodes $\mathcal{V}$. In particular, the leaf nodes represent all the tasks, while the internal nodes denote the groupings of leaf nodes. Intuitively, each node $v \in \mathcal{V}$ of the tree $\mathcal{T}$ can be associated with group G_v, which consists of all the leaf nodes (tasks) belonging to the subtree rooted at node v. Moreover, the higher level the internal node is located at, the weaker relatedness it controls. The root of $\mathcal{T}$ is assigned the highest level. To characterize such strength of relatedness among tasks, we assign a weight e_v to each node $v \in \mathcal{V}$ according to the prior knowledge via a hierarchical agglomerative clustering algorithm [Schickel-Zuber and Faltings, 2007]. As illustrated in Figure 1, it is apparent that the tasks "basketball" and "football" are more correlated as compared to the task "cooking". Thus, in Figure 1, the tasks "basketball" and "football" are first grouped in node v_4 with a weight $e_{v_4} = 0.6$. Then these two tasks are grouped in a higher level internal node v_5, whose weight $e_{v_5} = 0.4$, together with the task "cooking".

We mathematically formulate the source integration and tree-constrained[3] group lasso into one unified model,

$$\mathbf{\Gamma} = \frac{1}{2N}\sum_{t=1}^{T}\left\|\mathbf{y}_t - \sum_{s=1}^{S}\tfrac{1}{S}\mathbf{X}_s\mathbf{w}_{st}\right\|^2 + \frac{\lambda}{2}\sum_{s=1}^{S}\sum_{d=1}^{D_s}\sum_{v\in\mathcal{V}} e_v\|\mathbf{w}_{sG_v}^d\|, \tag{4}$$

where $\mathbf{w}_{sG_v}^d$ is a vector of coefficients $\{w_{st}^d : t \in G_v\}$. In addition, we assume that the mapping functions from all sources agree with one another as much as possible. Therefore, we introduce the regularization term to model the result consistency among different sources. The final objective function $\mathbf{\Gamma}$ is restated as follows,

$$\frac{1}{2N}\sum_{t=1}^{T}\left\|\mathbf{y}_t - \sum_{s=1}^{S}\tfrac{1}{S}\mathbf{X}_s\mathbf{w}_{st}\right\|^2 + \frac{\lambda}{2}\sum_{s=1}^{S}\sum_{d=1}^{D_s}\sum_{v\in\mathcal{V}} e_v\|\mathbf{w}_{sG_v}^d\| + \frac{\mu}{2N}\sum_{t=1}^{T}\sum_{s=1}^{S}\sum_{s'\neq s}\left\|\mathbf{X}_s\mathbf{w}_{st} - \mathbf{X}_{s'}\mathbf{w}_{s't}\right\|^2, \tag{5}$$

[3]Beyond tree-structure, our model is extendable to incorporate other structures, such as graph.

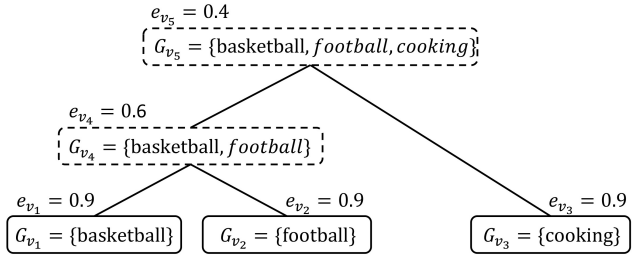

Figure 1: Illustration of inter-interests relatedness in a tree structure.

where μ is the nonnegative regularization parameter that regulates the disagreement among models learned from different sources.

3.3 Optimization

Considering that the second term in Eqn. (5) is not differentiable, we use an equivalent formulation of it, which has been proven by [Bach, 2008], to facilitate the optimization as follows,

$$\frac{\lambda}{2}\Big(\sum_{s=1}^{S}\sum_{d=1}^{D_s}\sum_{v\in\mathcal{V}} e_v\|\mathbf{w}_{sG_v}^d\|\Big)^2. \tag{6}$$

Still, the $L_{2,1}$ norm in the above formulation gives rise to a non-convex function, which makes it intractable to solve directly. Therefore, we further resort to another variational formulation [Argyriou *et al.*, 2008] of Eqn. (6). According to the Cauchy-Schwarz inequality, given an arbitrary vector $\mathbf{b} \in \mathbb{R}^M$ such that $\mathbf{b} \neq \mathbf{0}$, we have,

$$\begin{aligned}\sum_{i=1}^{M}|b_i| &= \sum_{i=1}^{M}\theta_i^{\frac{1}{2}}\theta_i^{-\frac{1}{2}}|b_i| \\ &\leq \Big(\sum_{i=1}^{M}\theta_i\Big)^{\frac{1}{2}}\Big(\sum_{i=1}^{M}\theta_i^{-1}b_i^2\Big)^{\frac{1}{2}} \leq \Big(\sum_{i=1}^{M}\theta_i^{-1}b_i^2\Big)^{\frac{1}{2}},\end{aligned} \tag{7}$$

where θ_i's are introduced variables that should satisfy $\sum_{i=1}^{M}\theta_i = 1, \theta_i > 0$ and the equality holds for $\theta_i = |b_i|/\|\mathbf{b}\|_1$. Based on this preliminary, we can derive the following inequality,

$$\begin{aligned}\Big(\sum_{s=1}^{S}\sum_{d=1}^{D_s}\sum_{v\in\mathcal{V}} e_v\|\mathbf{w}_{sG_v}^d\|\Big)^2 &\leq \sum_{s=1}^{S}\frac{\Big(\sum_{d=1}^{D_s}\sum_{v\in\mathcal{V}} e_v\|\mathbf{w}_{sG_v}^d\|\Big)^2}{q_s} \\ \leq \sum_{s=1}^{S}\sum_{d=1}^{D_s}\frac{\Big(\sum_{v\in\mathcal{V}} e_v\|\mathbf{w}_{sG_v}^d\|\Big)^2}{q_{s,d}} &\leq \sum_{s=1}^{S}\sum_{d=1}^{D_s}\sum_{v\in\mathcal{V}}\frac{e_v^2\|\mathbf{w}_{sG_v}^d\|^2}{q_{s,d,v}},\end{aligned} \tag{8}$$

where we introduce the variable $q_{s,d,v}$. The equality can be attained if $q_{s,d,v}$ satisfies that,

$$q_{s,d,v} = \frac{e_v\|\mathbf{w}_{sG_v}^d\|}{\sum_{s=1}^{S}\sum_{d=1}^{D_s}\sum_{v\in\mathcal{V}} e_v\|\mathbf{w}_{sG_v}^d\|}. \tag{9}$$

Consequently, minimizing $\mathbf{\Gamma}$ is equivalent to minimizing the following convex objective function,

$$\frac{1}{2N}\sum_{t=1}^{T}\left\|\mathbf{y}_t - \sum_{s=1}^{S}\tfrac{1}{S}\mathbf{X}_s\mathbf{w}_{st}\right\|^2 + \frac{\lambda}{2}\sum_{s=1}^{S}\sum_{d=1}^{D_s}\sum_{v\in\mathcal{V}}\frac{e_v^2\|\mathbf{w}_{sG_v}^d\|^2}{q_{s,d,v}} + \frac{\mu}{2N}\sum_{t=1}^{T}\sum_{s=1}^{S}\sum_{s'\neq s}\left\|\mathbf{X}_s\mathbf{w}_{st} - \mathbf{X}_{s'}\mathbf{w}_{s't}\right\|^2. \tag{10}$$

To facilitate the computation of the derivative of objective function $\mathbf{\Gamma}$ with respect to $\mathbf{w}_{st}$, we define a diagonal matrix $\mathbf{Q}_{st} \in \mathbb{R}^{D_s \times D_s}$ as follows,

$$Q_{st}(d,d) = \sum_{v:t \in G_v} \frac{e_v^2}{q_{s,d,v}}. \tag{11}$$

Finally, we have the following objective function,

$$\frac{1}{2N}\sum_{t=1}^{T}\left\|\mathbf{y}_t - \sum_{s=1}^{S}\tfrac{1}{S}\mathbf{X}_s\mathbf{w}_{st}\right\|^2 + \frac{\lambda}{2}\sum_{t=1}^{T}\sum_{s=1}^{S}\mathbf{w}_{st}^T\mathbf{Q}_{st}\mathbf{w}_{st} + \frac{\mu}{2N}\sum_{t=1}^{T}\sum_{s=1}^{S}\sum_{s'\neq s}\left\|\mathbf{X}_s\mathbf{w}_{st} - \mathbf{X}_{s'}\mathbf{w}_{s't}\right\|^2. \tag{12}$$

We adopt the alternating optimization strategy to solve Eqn. (12) [Kim and Xing, 2010]. Particularly, we alternatively optimize $\mathbf{w}_{st}$ and $q_{s,d,v}$, where we optimize one variable with the other one fixed in each iteration and keep this iterative procedure until the objective value converges.

When $q_{s,d,v}$ is fixed, we take the derivative of objective function $\mathbf{\Gamma}$ regarding $\mathbf{w}_{st}$ as follows,

$$\frac{\partial \mathbf{\Gamma}}{\partial \mathbf{w}_{st}} = \frac{1}{NS}\mathbf{X}_s^T\Big(\sum_{s=1}^{S}\frac{1}{S}\mathbf{X}_s\mathbf{w}_{st} - \mathbf{y}_t\Big) + \lambda\mathbf{Q}_{st}\mathbf{w}_{st} + \sum_{s\neq s'}\frac{\mu}{N}\mathbf{X}_s^T(\mathbf{X}_s\mathbf{w}_{st} - \mathbf{X}_{s'}\mathbf{w}_{s't}). \tag{13}$$

Setting Eqn. (13) to zero and rearranging the terms, we derive that all $\mathbf{w}_{st}$'s can be learned jointly by the following linear system given a task t,

$$\mathbf{L}_t\mathbf{w}_t = \mathbf{b}_t,$$

$$\begin{bmatrix} \mathbf{L}_{11} & \mathbf{L}_{12} & \mathbf{L}_{13} & \cdots & \mathbf{L}_{1S} \\ \mathbf{L}_{21} & \mathbf{L}_{22} & \mathbf{L}_{23} & \cdots & \mathbf{L}_{2S} \\ \mathbf{L}_{31} & \mathbf{L}_{32} & \mathbf{L}_{33} & \cdots & \mathbf{L}_{3S} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \mathbf{L}_{S1} & \mathbf{L}_{S2} & \mathbf{L}_{S3} & \cdots & \mathbf{L}_{SS} \end{bmatrix} \begin{bmatrix} \mathbf{w}_{1t} \\ \mathbf{w}_{2t} \\ \mathbf{w}_{3t} \\ \vdots \\ \mathbf{w}_{St} \end{bmatrix} = \begin{bmatrix} \mathbf{b}_{1t} \\ \mathbf{b}_{2t} \\ \mathbf{b}_{3t} \\ \vdots \\ \mathbf{b}_{St} \end{bmatrix}, \tag{14}$$

where $\mathbf{L}_t \in \mathbb{R}^{D\times D}$ is a sparse block matrix with $S \times S$ blocks, $\mathbf{w}_t \in \mathbb{R}^D$ and $\mathbf{b}_t \in \mathbb{R}^D$ are both sparse block matrices with S blocks. $\mathbf{L}_{ss}$, $\mathbf{L}_{ss'}$ and $\mathbf{b}_{st}$ are defined as,

$$\begin{cases} \mathbf{L}_{ss} & = \frac{1}{NS^2}\mathbf{X}_s^T\mathbf{X}_s + \frac{\mu(S-1)}{N}\mathbf{X}_s^T\mathbf{X}_s + \lambda\mathbf{Q}_{st}, \\ \mathbf{L}_{ss'} & = \frac{1}{NS^2}\mathbf{X}_s^T\mathbf{X}_{s'} - \frac{\mu}{N}\mathbf{X}_s^T\mathbf{X}_{s'}, \\ \mathbf{b}_{st} & = \frac{1}{NS}\mathbf{X}_s^T\mathbf{y}_t. \end{cases} \tag{15}$$

According to the definition of positive-definite matrix, $\mathbf{L}_t$ can be easily proven to be positive definite and invertible. Then we can derive the closed-form solution of $\mathbf{w}_t$ as follows,

$$\mathbf{w}_t = \mathbf{L}_t^{-1}\mathbf{b}_t. \tag{16}$$

Furthermore, we notice that $\mathbf{w}_t$ can be computed individually, which saves considerable space and time cost. On the other hand, we optimize $q_{s,d,v}$ according to Eqn. (9) with fixed $\mathbf{w}_t$.

3.4 Construction of Interest Tree Structure

We aim to employ the hierarchical agglomerative clustering algorithm to construct the tree structure. One challenge is that an interest is usually represented by a single concept, which makes it hard to measure the similarities among interests and apply the hierarchical agglomerative clustering algorithm. Towards this end, two types of prior knowledge are utilized.

1) **External source**. We exploit an external source—the Web, where a huge amount of prior knowledge about interests are encoded implicitly. We transform each interest into a query and submit it to Google search engine. We collect the top 10 webpages, and then employ the library of BoilerPipe[4] [Kohlschütter *et al.*, 2010] to extract clean main contents from the returned webpages. Therefore, each interest can be represented by a document, based on which the Bag-of-words model [Mitchell, 1997] with TF-IDF term weighting scheme [Salton and McGill, 1983] can be applied and the similarities among interests can be evaluated.

2) **Internal source**. Although the external source provides us the general prior knowledge, we believe that the internal prior knowledge stored in our dataset also plays a vital role in user interest inference. Driven by this consideration, we propose to measure the similarities among interests based on their co-occurrence in users' LinkedIn profiles in our dataset[5]. It deserves attention that we exploit all available LinkedIn profiles that exhibit users' personal interests rather than that of the subset of users selected for the task of interest inference. Suppose we have a set of interests $\mathcal{I} = \{In_1, In_2, \cdots, In_T\}$, and a set of documents $\mathcal{DD} = \{d_1, d_2, \cdots, d_N\}$, where d_l contains all interests of user l. Let $c(j,k,l) = 1$ if and only if interests In_j and In_k both occur in d_l, and otherwise $c(j,k,l) = 0$. Then the co-occurence matrix $\mathbf{H}$ is defined as follows,

$$H(j,k) = \begin{cases} \frac{\sum_l c(j,k,l)}{\sum_j\sum_l c(j,k,l)} & \text{if } j \neq k; \\ 1 & \text{otherwise.} \end{cases} \tag{17}$$

Each row of $\mathbf{H}$ corresponds to the co-occurrence of an interest with others. Then we use the JensenShannon divergence [Bordag, 2008] to measure the similarities among interests.

Then it is suggested to apply the hierarchical agglomerative clustering algorithm on these enriched interests and build the tree structure. To assign appropriate weights to nodes, we choose to utilize the normalized height h_v of subtree rooted at node v to characterize its weight e_v, where $e_v = 1 - h_v$. Such assignment guarantees the aforementioned condition that the higher node corresponds to the weaker relatedness. It is noted that we normalize the heights for all nodes such that the root node is at height 1. We thus derive two models *SM^{2}L-e* and *SM^{2}L-i* based on two types of prior knowledge, respectively.

3.5 Complexity Discussion

To analyze the complexity of *SM^{2}L*, we need to solve the time cost in terms of constructing $\mathbf{Q}$, $\mathbf{L}_t$ and $\mathbf{b}_t$, defined in Eqn. (11) and Eqn. (15), as well as computing the inverse of $\mathbf{L}_t$. Assuming $D \gg S$, the construction of diagonal matrix $\mathbf{Q}$ has a time complexity of $O(DT)$, and the construction of matrix $\mathbf{L}_t$ has a time complexity of $O(ND^2)$. Due to the fact that the time cost of matrix multiplication $\mathbf{X}_s^T\mathbf{X}_{s'}$ and that of constructing $\mathbf{b}_t$ involved in Eqn. (15) remain the same for all iterations and $\mathbf{L}_t$ is symmetric, we can reduce

[4] https://code.google.com/p/boilerpipe/.

[5] Users may list a set of personal interests in their LinkedIn profiles.

the practical time consumption remarkably. In addition, computing the inverse of $\mathbf{L}_t$ has the complexity of $O(D^3)$ by the standard method. Then the total complexity should be $O(D^3T)$. We notice that the speed bottleneck lies in the number of features and the number of tasks instead of the number of data samples. As D is usually small, *SM*^{2}*L* should be computationally efficient.

4 Experiments

In this paper, we cast the problem of user interest inference as the structure constrained multi-source multi-task learning problem. In particular, we explored four popular social networks: Twitter, Facebook, Quora and LinkedIn.

4.1 Dataset Construction

To construct the benchmark dataset, we need to first tackle the problem of "social account alignment", which aims to identify the same users across different social networks by linking their multiple social accounts [Abel *et al.*, 2013]. To accurately establish this mapping, we employed the emerging social service—Quora, which encourages users to explicitly list their multiple social accounts in their Quora profiles[6]. We collected candidates from Quora by the breadth-first-search method. In the end, we harvested $172,235$ Quora user profiles and only retained those who provided their Facebook, Twitter and LinkedIn accounts in their Quora profiles. Based on these mappings, we launched a crawler to collect their historical social contents, including their basic profiles, social posts and relations.

To build the ground truth, we employed the structural information of users' linkedin profiles: "Additional Information", which usually contains information about users' personal interests. Users' interests listed in their LinkedIn profiles are usually represented by phrases separated by comma, which facilitates the ground truth construction to a large extent. To obtain the representative interests, we filtered out the interests that are liked by less than 15 users. Finally, we obtained 74 interests[7]. Then we only retained those users who expressed these interests in their LinkedIn profiles and obtained $1,607$ users ultimately. Figure 2 shows the user frequency distribution with respect to the number of interests over our dataset.

4.2 Feature Extraction

To informatively describe users, we extracted two kinds of features: user topics and contextual topics.

User topics. We explored the topic distributions of users' social posts to infer users' interests. We generated topic distributions using Latent Dirichlet Allocation (LDA) [Blei *et al.*, 2003], which has been widely found to be useful in latent topic modeling [Cimiano *et al.*, 2009; Iwata *et al.*, 2009]. Based on perplexity [Li *et al.*, 2010], we ultimately obtained

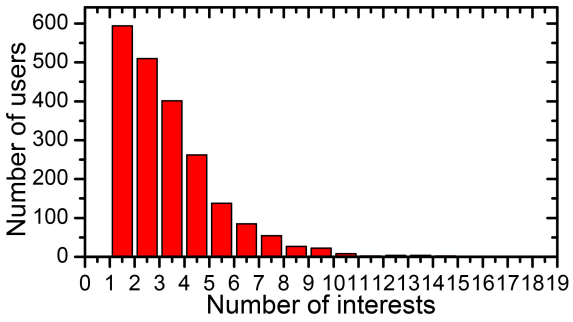

Figure 2: Distribution of user frequency distribution with respect to the number of interests over our dataset.

$89, 24, 119$ dimensional topic-level features respectively over users' Twitter[8], Facebook[9] and Quora[10] data.

Contextual topics. We define users' contextual topics as the topics of users' connections. As it goes that "birds of a feather flock together", we believe that the contextual topics intuitively reflect the contexts of users and further disclose users' interests. Particularly, we studied followee connections in Twitter because of their intuitive reflection of topics that users are concerned with. As the bio descriptions are usually provided by users to briefly express themselves and may indicate users' summarized interests, we merged the bio descriptions of a user's followees into a document, on which we further applied LDA model. We utilized the perplexity to tune the dimensions of topic-level features over these bio documents and obtained a 64 dimensional feature space. In this work, we only explored the contextual topics in Twitter, since the bio descriptions are usually missing in Facebook and Quora.

4.3 On Evaluation Matrics

For the task of user interest inference, precision is of more importance as compared to recall. We thus validated our scheme via two metrics: $S@K$ and $P@K$.

S@K: It represents the mean probability that a correct interest is captured within the top K recommended interests.

P@K: It stands for the proportion of the top K recommended interests that are correct.

All the experiments were conducted over a server equipped with Intel(R) Xeon(R) CPU X5650 at 2.67GHZ on 48GB RAM, 24 cores and 64-bit CentOS 5.4 operating system.

4.4 On Model Comparison

We compared *SM*^{2}*L* with the following five baselines.

SVM: The first baseline is a traditional single source single task learning method—support vector machine (SVM) [Cortes and Vapnik, 1995], which simply concatenates the features generated from different sources into a single feature vector and learns each task individually. We chose the learning formulation with the kernel of radial-basis function, implemented based on LIBSVM [Chang and Lin, 2011].

[6]One representative example can be seen via https://www.quora.com/Martijn-Sjoorda.

[7]These interests are available at http://msmt.farbox.com/.

[8]Users' Twitter data refers to users' historical tweets.

[9]Users' Facebook data refers to users' historical timelines.

[10]Users' Quora data refers to users' historical questions and answers.

RLS: The second baseline is the regularized least squares (RLS) model [Kim *et al.*, 2007], which also learns each task individually and aims to minimize the objective function of $\frac{1}{2N}\left\|\mathbf{y}_t - \sum_{s=1}^{S}\frac{1}{S}\mathbf{X}_s\mathbf{w}_{st}\right\|^2 + \frac{\lambda}{2}\|\mathbf{w}_t\|^2$.

regMVMT: The third baseline is the regularized multi-view multi-task learning model, introduced in [Zhang and Huan, 2012]. This model regulates both the source consistency and the task relatedness. However, it simply assumes the uniform relatedness among tasks.

SM^{2}L-eu: The fourth baseline is a derivation of *SM^{2}L-e*. This method constructs the tree structure based on external source in the same manner as *SM^{2}L-e* but assigns uniform weights to all nodes.

SM^{2}L-iu: The fifth baseline is a derivation of *SM^{2}L-i*, which constructs the tree structure using internal source but weights all nodes uniformly.

We adopted the grid search strategy to determine the optimal values for the regularization parameters among the values $\{10^r : r \in \{-12, \cdots, -1\}\}$. Experimental results reported in this work are the average values over 10-fold cross validation. Noticeably, we tuned the K in $S@K$ and $P@K$ from 1 to 10 and reported the optimal performance for each fold. Generally, the $S@K$ reaches the maximum at $K = 10$, while $K = 1$ is much preferable regarding $P@K$.

Table 1: Performance comparison among various models.

Approaches	$P@K$ (%)	$S@K$ (%)
SVM	8.69	54.69
RLS	24.32	73.86
regMVMT	24.69	74.54
SM^{2}L-eu	25.50	73.80
SM^{2}L-iu	24.56	74.11
SM^{2}L-e	**25.72**	**74.57**
SM^{2}L-i	**26.50**	**74.85**

Table 1 shows the performance comparison between baselines and our proposed scheme. We observed that *SM^{2}L*-i and *SM^{2}L*-e both outperform the single source single task learning *SVM* and *RLS*. This verifies the significance of considering source consistency and task relatedness simultaneously. Moreover, it is not unexpected that *SVM* achieves the worst performance. A possible explanation might be the insufficient positive training samples for certain interests. For example, only 24 positive training samples are available for the interest "surfing". In addition, the less satisfactory performance of *regMVMT*, as compared to *SM^{2}L*-i and *SM^{2}L*-e, confirms that it is advisable to characterize the task relatedness in a tree structure instead of correlating all tasks uniformly. Besides, *SM^{2}L-i* and *SM^{2}L-e* show superiority over *SM^{2}L-iu* and *SM^{2}L-eu* respectively, which enables us to draw a conclusion that modeling the relatedness strength among tasks merits our particular attention. Last but not least, *SM^{2}L*-i performs better than *SM^{2}L*-e. This finding demonstrates the importance of prior knowledge extracted from our internal source.

Based on the practical results, the time complexity of *regMVMT* is remarkably higher than that of *SM^{2}L*. In particular, *regMVMT* costs about 562 seconds to execute, 114 times of that taken by *SM^{2}L* for each iteration. This is mainly attributed to the computation of the inverse of a matrix with dimension of DT, which requires a time complexity of $O(D^3T^3)$. Compared to *SM^{2}L*, it is rather time consuming using *regMVMT*.

4.5 On Source Comparison

To shed light on the descriptiveness of multiple social network integration, we conducted experiments over various source combinations.

Table 2 shows the performance of *SM^{2}L-i* over individual social network and their various combinations. We noted that the more sources we incorporated, the better the performance can be achieved. This suggests the complementary relationships instead of mutual conflicting relationships among the sources. Moreover, we found that aggregating data from all these three social networks can achieve better performance as compared to each of the single source. Interestingly, we observed that *SM^{2}L* over Twitter alone achieves a much better performance, as compared to that using Quora or Facebook alone. This may be caused by that we additionally extracted contextual topics apart from user topics in Twitter, which can reveal users' interests more directly. It is far from incomprehensible that *SM^{2}L* would degenerate to multi-task learning when the context problem involves only one single source.

Table 2: Contribution of individual social network and their various combinations.

Social network combinations	$P@K$ (%)	$S@K$ (%)
Twitter	24.75	73.05
Facebook	19.59	69.74
Quora	20.97	68.19
Twitter+Facebook	25.51	74.98
Twitter+Quora	24.89	74.41
Facebook+Quora	22.52	71.80
Twitter+Facebook+Quora	**26.50**	**74.85**

5 Conclusions and Future Work

This paper presented a structure-constrained multi-source multi-task learning scheme in the context of user interest inference. In particular, this scheme takes both the source consistency and the tree-guided task relatedness into consideration by introducing two regularizations to the objective function. Moreover, the proposed model is able to effectively select the task-sharing features and task-specific features by employing the weighted group lasso. Notably, the weights can be learned from two kinds of prior knowledge: external source and internal source. Experimental results demonstrate the effectiveness of our proposed scheme. Currently, we only consider studying users' distributed textual data. In the future, we will extend our work to investigate users' visual information on social media services.

Acknowledgments

This research is supported by the Singapore National Research Foundation under its International Research Centre @ Singapore Funding Initiative and administered by the IDM Programme Office.

References

[Abel *et al.*, 2011a] Fabian Abel, Ilknur Celik, Claudia Hauff, Laura Hollink, and Geert-Jan Houben. U-sem: Semantic enrichment, user modeling and mining of usage data on the social web. *arXiv preprint arXiv:1104.0126*, 2011.

[Abel *et al.*, 2011b] Fabian Abel, Eelco Herder, and Daniel Krause. Extraction of professional interests from social web profiles. In *UMAP*, 2011.

[Abel *et al.*, 2013] Fabian Abel, Eelco Herder, Geert-Jan Houben, Nicola Henze, and Daniel Krause. Cross-system user modeling and personalization on the social web. *UMUAI*, 2013.

[Argyriou *et al.*, 2008] Andreas Argyriou, Theodoros Evgeniou, and Massimiliano Pontil. Convex multi-task feature learning. *Machine Learning*, 2008.

[Bach, 2008] Francis R Bach. Consistency of the group lasso and multiple kernel learning. *JMLR*, 2008.

[Blei *et al.*, 2003] David M Blei, Andrew Y Ng, and Michael I Jordan. Latent dirichlet allocation. *JMLR*, 2003.

[Bordag, 2008] Stefan Bordag. A comparison of co-occurrence and similarity measures as simulations of context. In *CICLing*. 2008.

[Chang and Lin, 2011] Chih-Chung Chang and Chih-Jen Lin. Libsvm: a library for support vector machines. *TIST*, 2011.

[Cimiano *et al.*, 2009] Philipp Cimiano, Antje Schultz, Sergej Sizov, Philipp Sorg, and Steffen Staab. Explicit versus latent concept models for cross-language information retrieval. In *IJCAI*, 2009.

[Cortes and Vapnik, 1995] Corinna Cortes and Vladimir Vapnik. Support-vector networks. *Machine learning*, 1995.

[Fei and Huan, 2013] Hongliang Fei and Jun Huan. Structured feature selection and task relationship inference for multi-task learning. *KAIS*, 2013.

[He and Lawrence, 2011] Jingrui He and Rick Lawrence. A graph-based framework for multi-task multi-view learning. In *ICML*, 2011.

[Iwata *et al.*, 2009] Tomoharu Iwata, Shinji Watanabe, Takeshi Yamada, and Naonori Ueda. Topic tracking model for analyzing consumer purchase behavior. In *IJCAI*, 2009.

[Jin *et al.*, 2013] Xin Jin, Fuzhen Zhuang, Shuhui Wang, Qing He, and Zhongzhi Shi. Shared structure learning for multiple tasks with multiple views. In *ECML/PKDD*. 2013.

[Kim and Xing, 2010] Seyoung Kim and Eric P Xing. Tree-guided group lasso for multi-task regression with structured sparsity. In *ICML*, 2010.

[Kim *et al.*, 2007] S Kim, Kwangmoo Koh, Michael Lustig, Stephen Boyd, and Dimitry Gorinevsky. A method for large-scale l1-regularized least squares problems with applications in signal processing and statistics. *J-STSP*, 2007.

[Kohlschütter *et al.*, 2010] Christian Kohlschütter, Peter Fankhauser, and Wolfgang Nejdl. Boilerplate detection using shallow text features. In *WSDM*, 2010.

[Li *et al.*, 2010] Daifeng Li, Bing He, Ying Ding, Jie Tang, Cassidy Sugimoto, Zheng Qin, Erjia Yan, Juanzi Li, and Tianxi Dong. Community-based topic modeling for social tagging. In *CIKM*, 2010.

[Mitchell, 1997] Tom M Mitchell. Machine learning. 1997. *McGraw Hill*, 1997.

[Pennacchiotti and Popescu, 2011] Marco Pennacchiotti and Ana-Maria Popescu. A machine learning approach to twitter user classification. *ICWSM*, 2011.

[Salton and McGill, 1983] Gerard Salton and Michael J McGill. Introduction to modern information retrieval. *McGraw Hill*, 1983.

[Schickel-Zuber and Faltings, 2007] Vincent Schickel-Zuber and Boi Faltings. Using hierarchical clustering for learning theontologies used in recommendation systems. In *KDD*, 2007.

[Tibshirani, 1996] Robert Tibshirani. Regression shrinkage and selection via the lasso. *JRSS B*, 1996.

[Xue *et al.*, 2007] Ya Xue, Xuejun Liao, Lawrence Carin, and Balaji Krishnapuram. Multi-task learning for classification with dirichlet process priors. *JMLR*, 2007.

[Zhang and Huan, 2012] Jintao Zhang and Jun Huan. Inductive multi-task learning with multiple view data. In *KDD*, 2012.