Metadata-Based Clustered Multi-task Learning for Thread Mining in Web Communities

Qiang You(✉), Ou Wu, Guan Luo, and Weiming Hu

CAS Center for Excellence in Brain Science and Intelligence Technology, National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China
{qyou,wuou,gluo,wmhu}@nlpr.ia.ac.cn

Abstract. With user-generated content explosively growing, how to find valuable posts from discussion threads in web communities becomes a hot topic. Although many learning algorithms have been proposed for mining the thread contents, there are still two problems that are not effectively considered. First, the learning algorithms are usually complicated so as to deal with various kinds of threads in web communities, which damages the generalization performance of the algorithms and takes the risk of overfitting to the learning models. Second, the small sample size problem exists when the training data for learning is divided into many isolated groups and each group is trained separately in order to avoid overfitting. In this paper, we propose a metadata-based clustered multi-task learning method, which takes full use of the metadata of threads and fuses it in the multi-task learning based on a divide-and-learn strategy. Our method provides an effective solution to the above problems by finding the geometric structure or context of semantics of threads in web communities and constructing the relations among training thread groups and their corresponding learning tasks. In addition, a soft-assigned clustered multi-task learning model is employed. Our experimental results show the effectiveness of our method.

Keywords: Metadata · Thread mining · Divide-and-learn · Clustered multi-task learning · Web community

1 Introduction

With the rapid development of the Internet, more and more people would like to participate in the discussions in web communities. As a result, a large amount of user-generated content (UGC) has been accumulating, which becomes urgent to analyze so as to find useful information for decision making in different kinds of areas such as viral marketing, industry research, etc. Throughout the past decade there have been many researches on how to find valuable posts in discussion threads in web communities. The previous researches are mainly classified into content-based or structure-based. The formal method takes the posts in each

P. Perner (Ed.): MLDM 2016, LNAI 9729, pp. 421–434, 2016.
DOI: 10.1007/978-3-319-41920-6_33

discussion thread as the document set and follows the pattern of text classification [1]. In the area of text classification, probabilistic topic models [2,3] have been proved to be effective in the extraction of semantics and document summarization when the corpus to be analyzed is sufficient. However, the posts in web communities are always short and sparse, which makes the result of text classification unsatisfactory. While the structure-based method goes another way, it ignores the semantics of the content and only concentrate on the structure of a web community. Considering the reply-to graph of the posts in web communities, many random-walk-based algorithms are available for measuring the importance of web pages such as HITS [4], PageRank [5] and their successive approaches are introduced to the valuable post finding. However, the reply-to graph of the posts is not explicit or hard to extract in many web communities. What is more, the posts without link-in or link-out are common in web communities, which is not applicable in most of the random-walk-based algorithms.

In this paper, we combine the content-based method and structure-based method together by the concept of metadata. We introduce the metadata into the mining tasks in web communities. If the reply-to graph of the web community apparently exists (e.g. *Slashdot.org*[1]), it is viewed as a kind of metadata. While the reply-to graph doesn't explicitly appear (e.g. many Q&A discussion forums), we reconstruct it through semantic similarity measure. Some data that shows the quality or characteristics of the data set for learning tasks can be viewed as metadata in our consideration. Different from some previous studies, we do not take the metadata just as a kind of data directly for learning and add to the learning tasks similarly to the other data. The reason is that the distribution of the metadata is different from that of the data for learning tasks and the metadata is also not independent from the data. If we simply add it and combine with the data to the learning tasks, the performance may be degraded seriously. In this paper, we conduct a *divide-and-learn* strategy. Specifically, In a web community, the different discussion threads are not isolated from each other because the users often make discussions around several central topics. Assuming that the discussion threads are clustered according to several topics in a web community, in the *dividing* step, we model the metadata of each thread as an attributed graph, and divide all the metadata attributed graphs into several groups. In the *learning* step, we propose a metadata-based clustered multi-task learning algorithm, which takes full use of the metadata and fuses it to the multi-task learning framework. The aim is that each task may benefit from each other by an appropriate sharing of information across different tasks in the framework of multi-task learning, which may significantly reduce the risk of overfitting if we develop our learning model with respect to the adaptive data.

The remainder of this paper is organized as follows. Section 2 briefly reviews related work. Section 3 presents the characteristics of the metadata in web communities, and Section 4 shows the formation of multiple tasks based on the metadata clustering. Section 5 describes the clustered multi-task learning frame-

[1] http://slashdot.org

work in detail. Experimental results are presented in Section 6, followed by the conclusion in Section 7.

2 Related Work

Many researchers have studied the mining tasks such as finding valuable posts or domain experts in web communities from the perspective of semantic understanding of the discussion threads and posts. As for the semantic models, Cong et al. [6] aimed at ranking answers for given questions in web forums. References such as [7] and [8] reconstructed the relationship among posts and threads based on the similarity of topics and semantics. Lin et al. [9] proposed a combination approach for simultaneously modeling semantics and structure of threaded discussions, which was used for junk detection and expert finding. The researches listed above almost all consider separated learning tasks in the whole feature space. Regardless of semantic reconstruction ([7,8]) or achievement in the optimization algorithm with respect to the relation of posts ([9]), there are still two problems commonly existing in mining tasks in web communities. First, the dimension of the whole feature space is high. An effective strategy is to partition the space into sub-regions and reduce the dimension according to the different characteristics of the data. Second, in spite of the large amount of data in the whole web, the data samples for a mining task in a web community is sparse and insufficient. Finally, we review a few previous studies involved in solving the two problems generally.

Several previous studies have implicated the concept of metadata. Researches in [10,11] extracted a large number of quality measures from the biometric traits. With the help of the quality information derived from the data, a unified framework for biometric expert fusion was constructed. The quality measures in biometric authentication can be treated as a kind of metadata in our consideration. Another kind of metadata describes the geometric characteristic of the original data. The quality measure is also adopted to web data classification [12]. In [13,14], a learning model with mixing linear SVMs was proposed to handle the problem of nonlinear classification, and to promote the efficiency while still providing a classification performance comparable to non-linear SVM. Based on the local linearly separable characteristic of the data set, the feature space can be partitioned into sub-regions. As a result, the learning model with mixing linear SVMs is available for nonlinear classification. However, the strategy that simply partitions the data set into subgroups according to the metadata and learns different models with respect to different groups largely ignores the connectivity of each group. Especially in web communities with discussion groups, the central topics are never completely isolated. What is more, the data samples for semantic analysis are more insufficient if we divide them into pieces.

Providing the sparsity and shortage of data samples in multiple related classification tasks under some circumstances, there is a growing interest in multi-task learning (MTL), where multiple related tasks are learned simultaneously by extracting appropriate shared information across tasks. The effectiveness of

MTL has been verified theoretically in researches such as [15–17]. Several methods have been proposed based on how the relatedness of different tasks is modeled. Mean-regularized MTL [18] was proposed under the assumption that the parameter vectors of all tasks are close to each other, which is simple but not hold in real applications such as mining in different topics of discussion threads in web communities. By sharing a different kind of underlying structure among multiple tasks, the relatedness can also be modeled as clusters [19,20], tree [21] or network [22,23]. In practical applications, the tasks may suggest a more sophisticated group structure where the models of tasks from the same group are closer to each other than those from a different group. There have been many researches involved in this line of research, known as clustered multi-task learning (CMTL). Bakker and Heskes [20] adopted a Bayesian approach by considering a mixture of Gaussians instead of single Gaussian priori to realize the clustered multi-task learning. Xu et al. [24] identified subgroups of related tasks using the Dirichlet process prior. Jacob et al. [25] proposed a clustered MTL framework that simultaneously identified clusters and performed multi-task inference. Given that the formulation is non-convex, they introduced a convex relaxation to the original formulation. Zhou et al. [26] found the equivalence between alternating structure optimization and CMTL formulation. They also relaxed the problem and solved it though alternating optimization method and other two gradient optimization algorithms [26]. While the previous researches of CMTL all assume the tasks are clustered into isolated groups, in this paper, we extend and propose a soft assigned CMTL in order to study the semantic context of different task groups.

3 The Metadata

The metadata has been widely used in search engine techniques where the web crawler can easily get the characteristics of the web page such as *charset*, *encoding*, *key words* and other descriptive information without crawling the whole page. Similarly, it is introduced here to show the schema of the data set to be analyzed. Let us take the popular technology-related news web community *Slashdot.org* as an example. *Slashdot.org* is a typical web community organized with threads constituted by posts which are scored by users where the score can be seen the value of the post.

There are mainly two methods to conduct the mining task. As Fig. 1a shows, with the content of the post and the score as its label, we can learn a model with each thread without much difficulty. However, we may face a small sample size problem because the posts in a thread is insufficient. On the contrary, when we take all the posts into learning without separating the posts according to the thread as shown in Fig. 1b, and use vector space models such as term frequency as feature description, the dimension of the feature space is really high and nearly all the posts are sparse, which damages the performance of the learning model. What is more, the learning process is also time-consuming. Rather than dimension reduction via feature selection by different kinds of rules, we extract the metadata to describe the characteristics of the discussion threads. The metadata of the thread in a web community consists two aspects: one is the structure

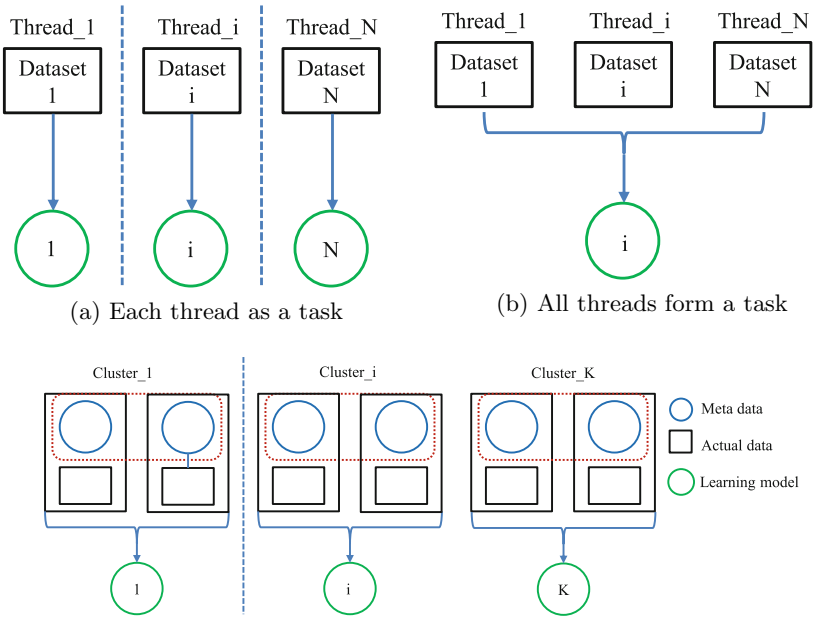


Fig. 1. The three methods for learning in a typical web community

of the threads modeled as a reply-to graph, the other is the topic distribution of the posts in the thread. The formal shows the context of the related posts, while the latter suggests the geometric characteristics and quality of themselves. With respect to the two aspects of the metadata of the thread in the web community, we model it as an attributed graph.

We assume that a web community is constituted by N threads, and the i-th thread is represented as a directed graph $G_i(V, E)$. The node $v \in V$ is associated with a post. There is an attribute vector $\mathbf{a}_{iv}$ described the characteristic of the post v in the i-th thread. Inspired by [27,28], we design a clustering algorithm with respect to the metadata as shown in Fig. 1c. Our algorithm is based on the assumption that rather than isolated from each other, the threads are clustered according to several central topics.

3.1 Metadata Modeling

As mentioned above, there are N threads in the web community. The metadata of the i-th thread can be modeled as an attributed graph $G_i(V, E, A)$ where V is the set of nodes, E the set of edges, and A the set of m attributes with nodes in V for describing node properties. The attribute vector is represented as $\mathbf{a}_{iv} = [a_1(v), ..., a_j(v), ..., a_n(v)]$ where $a_j(v)$ is the attribute value of node v on attribute a_j.

3.2 Metadata in Web Communities

Metadata is a very general concept in our description, which shows the characteristics of the original data. While the characteristics are hard to obtain from the original data because neither do we know the distribution of the data set nor do we assume too much in order to avoid decreasing the generalization performance, we mainly focus on two aspects of the metadata in a web community. One is the reply-to graph which we think supplies the semantic context of a thread. The other is the weighted topic distribution of each post in the thread, which shows the geometric characteristics and quality of each post.

Reply-To Graph. The reply-to graph directly exists in some web communities, while in other web communities it does not explicitly appear. To tackle the latter issue, we propose a semantic reconstruction algorithm to create the reply-to structure based on the semantic similarity measure.

Given a thread with m posts $\{\mathbf{L}_i\}_{i=1}^m$, their time stamps $\{ts_i\}_{i=1}^m$ where $ts_i < ts_j$ if $i < j$ and the similarity measure function $S(\mathbf{L}_i, \mathbf{L}_j)$, we reconstruct the reply-to structure through the following method. In our similarity computation, we define the similarity measure function as the weighted sum of two parts: The first part $S_{\cos}(\mathbf{L}_i, \mathbf{L}_j)$ is the cosine similarity which measures the similarity between the directions of two feature vectors, which shows the consistency of the semantics between two posts.

$$S_{\cos}(\mathbf{L}_i, \mathbf{L}_j) = \frac{\mathbf{L}_i \mathbf{L}_j^T}{2\|\mathbf{L}_i\| \, \|\mathbf{L}_j\|} \tag{1}$$

The second part $S_{str}(\mathbf{L}_i, \mathbf{L}_j)$ is the similarity between two posts with respect to the strength of the semantics.

$$S_{str}(\mathbf{L}_i, \mathbf{L}_j) = \frac{\|\mathbf{L}_i\| \, \|\mathbf{L}_j\|}{\|\mathbf{L}_i\|^2 + \|\mathbf{L}_j\|^2} \tag{2}$$

The parameter λ here weights the two parts. Now we get the whole similarity function

$$S(\mathbf{L}_i, \mathbf{L}_j) = \lambda S_{\cos}(\mathbf{L}_i, \mathbf{L}_j) + (1 - \lambda) S_{str}(\mathbf{L}_i, \mathbf{L}_j) \tag{3}$$

As for post $\mathbf{L}_j$, we choose one post as its predecessor from the ahead posts. The predecessor should have the most similarity with $\mathbf{L}_j$. We choose the predecessor of $\mathbf{L}_j$ according to the following maximum problem, where $\mathbf{L}_*$ is the most suitable predecessor.

$$\mathbf{L}_* = \arg_{\mathbf{L}_i} \max_{1 \le i \le j-1} S(\mathbf{L}_i, \mathbf{L}_j) \tag{4}$$

Let j decrement from m to 2, then the reply-to structure is reconstructed.

Weighted Topic Distribution. We deem each thread with many posts in a web community as a document with many paragraphs and thus the whole community with many threads can be seen as a document set. After that, we apply

the latent Dirichlet allocation [3] algorithm to the document set and extract n hottest topics. In every post of a thread, we can map each word in the post to one of the n hottest topics with a relevance weight. Therefore, the n dimensional weighted topic distribution of every post is extracted, which is the attribute vector in the attributed graph representing the metadata.

4 Formation of Multiple Tasks

In the following, we first define a metric to measure the distance between attributed graphs. Based on the similarity measures of each graph, we can cluster the metadata into different groups. Consequently, the multiple tasks are naturally created with respect to the different groups.

4.1 The Similarity Measure

As aforementioned, the metadata of the i, j-th thread is modeled as an attributed graph $G_i(V, E, A), G_j(V', E', A')$. Their directed product $G_\times$ is a graph with vertex set $V_\times = \{(v_r, v'_r) : v_r \in V, v'_r \in V'\}$ and edge set $E_\times = \{((v_r, v'_r), (v_s, v'_s)) : (v_r, v_s) \in E \wedge (v'_r, v'_s) \in E'\}$. The k-th order subgraph of the graph G_i is defined as $G_i^k(V^k, E^k)$ where $V^k \subseteq V, E^k \subseteq E$ and $|V^k| = k \leq |V|$. As we only want to extract the semantic context similarity of the attributed graph without comparing the two whole graphs, unlike the time-consuming calculation of the similarity is conducted from the 1-st order to the full order subgraphs [28], we only calculate the similarity between the second order subgraphs (edges). The similarity measure between G_i and G_j can be defined as the graph kernel:

$$k(G_i, G_j) = \frac{\sum_{e \in E} \sum_{e' \in E'} k(e, e')}{|E_\times|} \tag{5}$$

Assume $e = (v_r, v_s), e' = (v'_r, v'_s)$ and the node v_r with the attribute $\mathbf{a}_r$, then $k(e, e')$ is defined as the attribute similarity

$$k(e, e') = \phi(\mathbf{a}_r, \mathbf{a}'_r) \times \phi(\mathbf{a}_s, \mathbf{a}'_s) \tag{6}$$

where in our calculation, we use the Gaussian similarity function defined as follows:

$$\phi(\mathbf{x}, \mathbf{y}) = \exp(-\gamma \|\mathbf{x} - \mathbf{y}\|^2) \tag{7}$$

where γ is a scalar parameter determined the width of the Gaussian kernel. In the following metadata clustering algorithm, we set $\gamma = 1$ for simplicity.

4.2 The Metadata Clustering

Given the metadata set for the whole N threads in a web community, the similarity matrix is $S \in \mathcal{R}^{N \times N}$ where the element $s_{ij} = k(G_i, G_j)$. The diagonal matrix is D where $d_i = \sum_{j=1}^{m} s_{ij}$. Accordingly, the graph Laplacian matrix is represented as $L = D - S$. Once we get the graph Laplacian matrix, referring to the spectral clustering algorithm [29] and the efficient clustering method based on the data fragments [30], we can easily cluster the metadata set into several groups.

5 The Learning Algorithm

The multiple learning tasks are automatically constructed after we cluster the metadata into several groups. Not only the semantic context of the feature space is considered, we also want to study the semantic context of the task graphs. Unlike the previous studies in CMTL, we think the learning tasks are softly clustered in groups instead of independently assigned to each group. We propose the softly clustered multi-task learning (sCMTL) algorithm with Gaussian mixture models.

Given K (clusters of the metadata) learning tasks $\{T_i\}_{i=1}^{K}$, for the i-th task T_i with its feature space $\mathbf{F}_i \subseteq \mathbb{R}^{d_i}$ where d_i is the dictionary dimension of the $i-th$ thread, the training set consists of n_i sample points $\{(\mathbf{x}_j^i, y_j^i)\}_{j=1}^{n_i}$, with $\mathbf{x}_j^i \in \mathbf{F}_i$ and its corresponding output $y_j^i \in \mathbb{R}$ if it is a regression problem. The linear function for T_i is defined as $f_i(\mathbf{x}) = \mathbf{w}_i^T\mathbf{x} + b_i$. The loss function is defined as $l(f(\mathbf{x}), y) = (f(\mathbf{x}) - y)^2$. The basic model is to find an optimal value of $W = \{(\mathbf{w}_i, b_i)\}_{i=1}^{m}$ through minimizing the loss function

$$\mathcal{L}(W) = \frac{1}{K}\sum\nolimits_{i=1}^{K}\sum\nolimits_{j=1}^{n_i} l(f(\mathbf{x}_j^i), y_j^i) \tag{8}$$

As for Eq. (8), there is nothing much different from the single-task learning for K tasks respectively. In order to learn the K tasks simultaneously, we follow the regularized form to minimize the empirical risk where the regularized part $\Omega(W)$ can be designed from priori knowledge to constrain some sharing of information between tasks. The learning framework can be represented as the minimum optimization problem with respect to the learning weight matrix W.

$$\min_{W} \mathcal{L}(W) + \lambda\Omega(W) \tag{9}$$

The whole regularization Ω can be divided into two partial regularization parts, namely, the clustered regularization part $\mathcal{C}$ and the parameter penalty part. Now we have

$$\Omega(W) = \alpha\mathcal{C}(W) + \beta[tr(W^T W)] \tag{10}$$

Suppose that the i-th task with learning weight $\mathbf{w}_i$ can be assigned to the j-th task cluster with the probability p_{ij} and there are $k \leq K$ task clusters, the clustered regularization part can be written as follows

$$\begin{aligned} \mathcal{C}(W) &= \sum\nolimits_{j=1}^{k}\sum\nolimits_{i=1}^{m} p_{ij} \left\|\mathbf{w}_i - \overline{\mathbf{w}}_j\right\|_2^2 \\ s.t. &\sum\nolimits_{j=1}^{k} p_{ij} = 1; i = 1, ..., K \end{aligned} \tag{11}$$

The softly specified probability matrix is simply written as $\mathcal{P} \in [0,1]^{K \times k}$ as the element p_{ij} is the assigned probability. The clustered regularization part can be simplified as the following form

$$\begin{aligned} \mathcal{C}(W) &= tr(W^T W) - tr(\mathcal{P}^T W^T W \mathcal{P}) \\ s.t. &\sum_{j=1}^{k} p_{ij} = 1; i = 1, ..., K \end{aligned} \tag{12}$$

where $tr(.)$ is the trace of a matrix. To learn both the soft assigned matrix $\mathcal{P}$ and learning weight matrix W, the whole regularization Ω can be written as:

$$\Omega(W, \mathcal{P}) = \alpha[tr(W^T W) - tr(\mathcal{P}^T W^T W \mathcal{P})] + \beta[tr(W^T W)] \tag{13}$$

Let $\eta = \beta/\alpha > 0$. Since $tr(W^T W) = tr(W W^T)$,

$$\begin{aligned} \Omega(W, \mathcal{P}) &= \alpha \left((1+\eta) tr(W^T W) - tr(\mathcal{P}^T W^T W \mathcal{P}) \right) \\ &= \alpha \left((tr(W^T \left((1+\eta) I - \mathcal{P}\mathcal{P}^T \right) W) \right) \end{aligned} \tag{14}$$

As for $\mathcal{P}$, it is concave, and the formulation in Eq. (14) is non-convex. We conduct an alternating optimization method to inference the parameters. If the softly specified matrix $\mathcal{P}$ is fixed, Eq. (14) is convex with respect to W. It can be solved using gradient methods. After we find the optimal W^* to minimize the loss function with the whole penalty regularization, we simply fix W^*, and recompute the soft assigned matrix $\mathcal{P}$ with the Gaussian mixture models. We repeat the alternating optimization procedure until the constraints (e.g. the constraint steps, or the minimal error rate) achieved.

6 Experimental Results

We collect two data sets over a period of time by a web crawler designed for the threaded discussion communities. One is from the iPad Q&A board in the apple discussion forum; the other is from the technique community *Slashdot.org*. These two data sets are chosen because of the following reasons: (1)The two data sets are from two kinds of typical threaded discussion communities. The first is the Q&A forum, and the second is an open discussion forum where everyone can participate and judge the comments. Both of them have interested hot topics and the reply-to structure can be extracted or reconstructed without much difficulty. (2) Both data sets contain labeled information. The iPad Q&A data set can label the answers "Helpful" by other users or "Solved" by the questioner, which we quantize as 2, 1, 0 respectively, while *Slashdot.org* can give each comment a score ranging from -1 to 5 by all the participators. In the preprocess of text feature extraction, we first remove the stop words, and then collect the terms whose number is no less than 3. For each data set, we select 5 hottest topics and ignore the unqualified threads that have posts fewer than three or without labels or ratings. The basic statistic results are shown in Table 1. The two kinds of threaded discussion communities are quite different in average thread

Table 1. The basic statistics of the data sets

Data set	iPad Q&A	Slashdot.org
Number of threads	1130	664
Number of posts	8489	146569
Number of users	2175	14241
Average thread length	7.51	220.74
Average words per post	63.09	76.33
Average posts per user	3.90	10.29
Number of topics	5	5

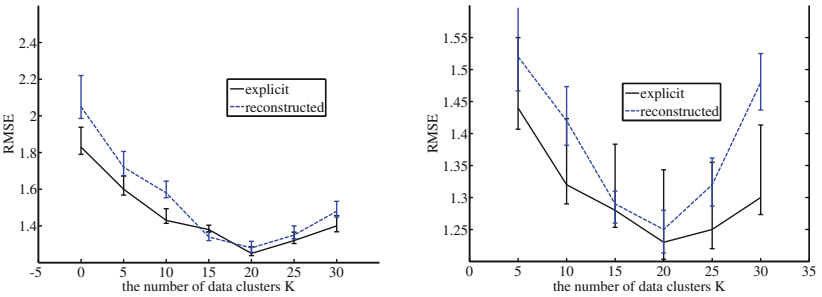
lengthes, user active degrees and so on. However, by computing the similarity in content and structure organization, we can obtain valuable answers to the questions or recommend the popular comments in our clustered multi-task learning framework.

Throughout the experiments, we use the root mean square error (RMSE) across the tasks as a criterion. The performance is better when RMSE is lower. In the learning process, the results are evaluated by 5-fold cross validation. In the following, first we conduct the experiments of the performance between the explicit reply-to graph and the reconstructed reply-to graph for *Slashdot.org*. Due to the lack of explicit reply-to graph in iPad Q&A data set, we reconstruct the reply-to graph with semantic similarity measure. Second, we discuss the two significant hyper parameters: one is the number of original multiple tasks K automatically formed by the metadata clustering, and the other is the number of task clusters k used for CMTL. We compare our sCMTL with the classical single-task learning methods such as the linear SVM, the Gaussian kernel SVM, and the hard-assigned CMTL. In the inference of W in multi-task learning algorithms, the logistic loss function is unified chosen for simplicity.

6.1 Evaluation for the Explicit v.s. Reconstructed Reply-to Graph

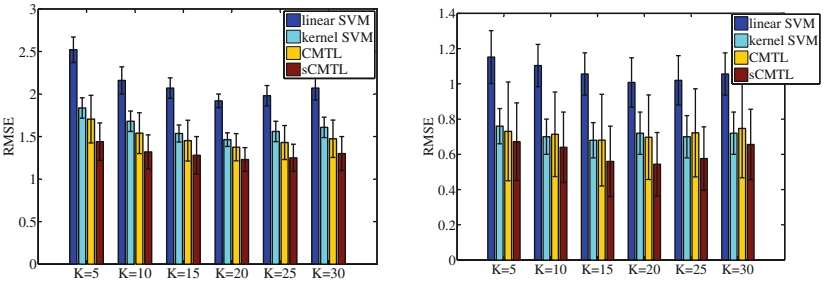
The experiments are only conducted on the *Slashdot.org* data set because the other data set is lack of explicit reply-to graph. To better measure the performance influenced by the partial metadata which shows the semantic context of each discussion thread, after the metadata clustering procedure, we compare them in the framework of the single-task learning and multi-task learning separately. In single-task learning, we choose the linear SVM for regression; while in MTL, we select our sCMTL with $k = 3$. The number of data clusters $K \geq k$ is changed from 5 to 30.

As shown in Fig. 2, it suggests that whatever in single-task learning or MTL, the performance the metadata modeling based on reconstructed reply-to graph is comparable to that based on the explicit graph when we set the suitable K. Because the explicit reply-to graph truly shows the semantic interaction between the web users, throughout the semantic reconstruction method, the reconstructed graph can basically suggest the realistic semantic interaction.



(a) RMSE vs. K using linear SVM (b) RMSE vs. K using sCMTL

Fig. 2. The comparison between explicit and reconstructed reply-to graphs



(a) RMSE vs. K for Slashdot.org (b) RMSE vs. K for iPad Q&A

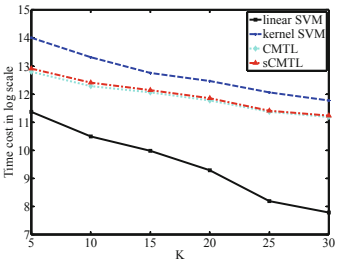
Fig. 3. The learning algorithms comparison on two data sets

Accordingly, in the following experiments, when the explicit graph does not exist, the reconstruction method is available as alternative.

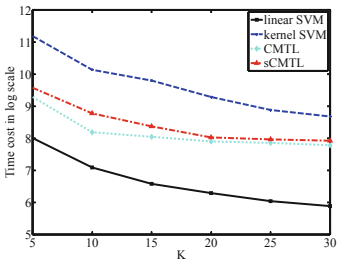
6.2 Evaluation for the Number of Data Clusters K

The hyper parameter K is significant as it shows how many semantic subspaces can be set apart in the whole feature space. On one hand, we can develop the corresponding learning model which needs not to meet all kinds of conditions; on the other hand, the semantics can be shared in each cluster, which alleviates the small sample size problem. Similar to the previous experiment settings, we set $k = 3$ for MTL algorithms, and change K from 5 to 30. We compare the single-task learning with MTL algorithms.

As shown in Fig. 3, both on two data sets, both MTL (CMTL and sCMTL) algorithms are generally more effective than the classical single-task learning algorithms. The proposed sCMTL outperforms the other three methods. Besides the comparison of performance, we also record the time cost for the learning methods on both of the two data sets. Through overall comparing the results in



(a) Time cost vs. K for Slashdot.org



(b) Time cost vs. K for iPad Q&A

Fig. 4. The comparison of time costs among the learning methods on two data sets

Table 2. Evaluation for CMTL vs sCMTL

Data set	iPad Q&A		Slashdot.org	
Learner	CMTL	sCMTL	CMTL	sCMTL
$k = 2$	0.76 ± 0.24	0.62 ± 0.24	1.38 ± 0.16	1.34 ± 0.16
$k = 4$	0.72 ± 0.20	0.58 ± 0.18	1.31 ± 0.16	1.23 ± 0.14
$k = 6$	0.72 ± 0.20	0.61 ± 0.20	1.85 ± 0.14	1.51 ± 0.14
$k = 8$	0.77 ± 0.20	0.68 ± 0.21	2.28 ± 0.14	1.96 ± 0.13
$k = 10$	0.81 ± 0.21	0.75 ± 0.20	2.47 ± 0.13	2.31 ± 0.10

Fig. 3 and Fig. 4, the linear SVM is much faster than the other methods but with worst performance even inapplicable to the thread mining tasks. Although the kernel SVM performs well comparable to the CMTL, it is much more time-consuming than the other algorithms. The proposed sCMTL costs almost equally with the CMTL, but performs much better.

6.3 Evaluation for the Number of Task Clusters k

The hyper parameter k shows the geometric characteristics of the learning tasks. With the CMTL, the learning weight can be shared among tasks. We set $K = 20$, and change k from 2 to 10. We compare the two CMTL algorithms based on the two data sets. As shown in Table 2, the sCMTL is more effective than CMTL on both of the two data sets.

7 Conclusion

In this paper, we have studied the metadata-based clustered multi-task learning for thread mining in web communities, which takes use of the metadata and fuses it in the framework of multi-task learning based on the divide-and-learn strategy. We divide the data set according to the metadata clustering, and learn multiple tasks simultaneously in the framework of softly assigned clustered multi-task

learning. We have conducted the experiments on two real data sets from two kinds of web communities. The experimental results show that our method is more effective than many previous learning algorithms, and the moderate time cost makes the propose method acceptable to the thread mining tasks.

Acknowledgments. This work is partly supported by the 973 basic research program of China (Grant No. 2014CB349303), the Natural Science Foundation of China (Grant No. 61472421 and No. 61379098), the Project Supported by CAS Center for Excellence in Brain Science and Intelligence Technology, and the Project Supported by Guangdong Natural Science Foundation (Grant No. S2012020011081).

References

1. Pang, B., Lee, L., Vaithyanathan, S.: Thumbs up?: sentiment classification using machine learning techniques. In: Proceedings of the ACL-02 Conference on Empirical Methods in Natural Language Processing-Volume 10, pp. 79–86. Association for Computational Linguistics (2002)
2. Hofmann, T.: Probabilistic latent semantic indexing. In: Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 50–57. ACM (1999)
3. Blei, D.M., Ng, A.Y., Jordan, M.I.: Latent dirichlet allocation. The Journal of Machine Learning Research **3**, 993–1022 (2003)
4. Kleinberg, J.M.: Authoritative sources in a hyperlinked environment. Journal of the ACM (JACM) **46**(5), 604–632 (1999)
5. Page, L., Brin, S., Motwani, R., Winograd, T.: The pagerank citation ranking: bringing order to the web (1999)
6. Cong, G., Wang, L., Lin, C.Y., Song, Y.I., Sun, Y.: Finding question-answer pairs from online forums. In: Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 467–474. ACM (2008)
7. Blei, D.M., Moreno, P.J.: Topic segmentation with an aspect hidden markov model. In: Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 343–348. ACM (2001)
8. Shen, D., Yang, Q., Sun, J.T., Chen, Z.: Thread detection in dynamic text message streams. In: Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 35–42. ACM (2006)
9. Lin, C., Yang, J.M., Cai, R., Wang, X.J., Wang, W.: Simultaneously modeling semantics and structure of threaded discussions: a sparse coding approach and its applications. In: Proceedings of the 32nd International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 131–138. ACM (2009)
10. Poh, N., Kittler, J., Bourlai, T.: Quality-based score normalization with device qualitative information for multimodal biometric fusion. IEEE Transactions on Systems, Man and Cybernetics, Part A: Systems and Humans **40**(3), 539–554 (2010)
11. Poh, N., Kittler, J.: A unified framework for biometric expert fusion incorporating quality measures. IEEE Transactions on Pattern Analysis and Machine Intelligence **34**(1), 3–18 (2012)

12. Wu, O., Hu, R., Mao, X., Hu, W.: Quality-based learning for web data classification. In: Twenty-Eighth AAAI Conference on Artificial Intelligence (2014)
13. Fu, Z., Robles-Kelly, A., Zhou, J.: Mixing linear svms for nonlinear classification. IEEE Transactions on Neural Networks **21**(12), 1963–1975 (2010)
14. Gu, Q., Han, J.: Clustered support vector machines. In: Proceedings of the Sixteenth International Conference on Artificial Intelligence and Statistics, pp. 307–315 (2013)
15. Ben-David, S., Schuller, R.: Exploiting task relatedness for multiple task learning. In: Schölkopf, B., Warmuth, M.K. (eds.) COLT/Kernel 2003. LNCS (LNAI), vol. 2777, pp. 567–580. Springer, Heidelberg (2003)
16. Torralba, A., Murphy, K.P., Freeman, W.T.: Sharing features: efficient boosting procedures for multiclass object detection. In: Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, CVPR 2004, vol. 2, p. II-762. IEEE
17. Ando, R.K., Zhang, T.: A framework for learning predictive structures from multiple tasks and unlabeled data. The Journal of Machine Learning Research **6**, 1817–1853 (2005)
18. Evgeniou, T., Pontil, M.: Regularized multi-task learning. In: Proceedings of the tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 109–117. ACM (2004)
19. Thrun, S., O'Sullivan, J.: Clustering learning tasks and the selective cross-task transfer of knowledge. Springer (1998)
20. Bakker, B., Heskes, T.: Task clustering and gating for bayesian multitask learning. The Journal of Machine Learning Research **4**, 83–99 (2003)
21. Kim, S., Xing, E.P.: Tree-guided group lasso for multi-task regression with structured sparsity. In: Proceedings of the 27th International Conference on Machine Learning (ICML 2010), pp. 543–550 (2010)
22. Chen, J., Liu, J., Ye, J.: Learning incoherent sparse and low-rank patterns from multiple tasks. ACM Transactions on Knowledge Discovery from Data (TKDD) **5**(4), 22 (2012)
23. Chen, J., Tang, L., Liu, J., Ye, J.: A convex formulation for learning shared structures from multiple tasks. In: Proceedings of the 26th Annual International Conference on Machine Learning, pp. 137–144. ACM (2009)
24. Xue, Y., Liao, X., Carin, L., Krishnapuram, B.: Multi-task learning for classification with dirichlet process priors. The Journal of Machine Learning Research **8**, 35–63 (2007)
25. Jacob, L., Bach, F., Vert, J.P., et al.: Clustered multi-task learning: a convex formulation. In: NIPS, vol. 21, pp. 745–752 (2008)
26. Zhou, J., Chen, J., Ye, J.: Clustered multi-task learning via alternating structure optimization. In: NIPS, pp. 702–710 (2011)
27. Zhou, Y., Cheng, H., Yu, J.X.: Graph clustering based on structural/attribute similarities. Proceedings of the VLDB Endowment **2**(1), 718–729 (2009)
28. Cheng, H., Zhou, Y., Yu, J.X.: Clustering large attributed graphs: A balance between structural and attribute similarities. ACM Transactions on Knowledge Discovery from Data (TKDD) **5**(2), 12 (2011)
29. Von Luxburg, U.: A tutorial on spectral clustering. Statistics and Computing **17**(4), 395–416 (2007)
30. Wu, O., Hu, W., Maybank, S.J., Zhu, M., Li, B.: Efficient clustering aggregation based on data fragments. IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics **42**(3), 913–926 (2012)