Semi-Supervised Collaborative Text Classification

Rong Jin[1], Ming Wu, and Rahul Sukthankar[2]

[1] Michigan State University, East Lansing MI 48823, USA,
rongjin@cse.msu.edu
[2] Intel Research Pittsburgh and Carnegie Mellon University, USA,
rahuls@cs.cmu.edu

Abstract. Most text categorization methods require text content of documents that is often difficult to obtain. We consider "Collaborative Text Categorization", where each document is represented by the feedback from a large number of users. Our study focuses on the semi-supervised case in which one key challenge is that a significant number of users have not rated any labeled document. To address this problem, we examine several semi-supervised learning methods and our empirical study shows that collaborative text categorization is more effective than content-based text categorization and the manifold regularization is more effective than other state-of-the-art semi-supervised learning methods.

1 Introduction

Most studies of text categorization are based on the textual contents of documents. The most common approach for text categorization is to first represent each document by a vector of term frequency, often called the bag-of-words representation, and then to apply classification algorithms based on term frequency vectors. The classification accuracy often heavily depends on the quality of textual contents of documents.

This paper focuses on the case where the textual contents of documents are either inaccurate or difficult to acquire, which makes it difficult to apply the standard text categorization methods. To this end, we propose **Collaborative Text Categorization** (as opposed to content-based text categorization) which classifies documents using the users' feedback such as ratings and click-through data. The underlying assumption is that two documents are likely to be in one category if they share similar feedback from a large number of users.

A straightforward approach toward collaborative text categorization is to represent each document by a vector of users' feedback. The problem arises when the number of labeled documents is small, which we refer to as "*semi-supervised collaborative text categorization*". Given a small number of labeled documents, the feedback from users who gave no feedback for *any* of the labeled documents will not be incorporated into the classification model. We refer to this problem as the "*missing user*" problem. This paper focuses on how to address the missing user problem in semi-supervised collaborative text categorization by exploiting the unlabeled documents. We will examine four semi-supervised approaches including *label propagation*, *user clustering*, the *kernel* approach, and *manifold regularization*.

The remainder of this paper is organized as follows: Section 2 briefly reviews the previous work on text categorization as well as the studies on exploiting collaborative feedback for information retrieval; Section 3 describes the problem of collaborative text categorization and the four semi-supervised approaches for the missing user problem; Section 4 presents our empirical study with movie classification; Section 5 concludes this paper with the future work.

2 Related Work

This work is closely related to previous studies on exploiting user feedback information for information retrieval [5, 8, 11]. Unlike the previous studies in information retrieval, this study utilizes the user feedback information for text categorization. Our work also differs from the previous studies on adaptive information filtering (e.g., [10]) in that the adaptive information filtering employs the feedback as a class label while our work uses feedback as part of the document representation.

Our work is related to the previous research on text categorization, including decision trees [2], logistic regression [12], and support vector machines (SVM) reported as the best [7]. A number of studies have also been devoted to using semi-supervised learning techniques for text categorization, including transductive support vector machine [9], graph-based approaches [1, 13] and Bayesian classifiers [4]. Our work differs from earlier research in that it uses the users' feedback, rather than the textual content, for classification and it focuses on exploiting the unlabeled documents to alleviate the missing user problem.

3 Semi-supervised Collaborative Text Categorization

We describe the semi-supervised collaborative text categorization problem, and then present four semi-supervised learning approaches that can potentially alleviate the missing user problem in semi-supervised collaborative text categorization.

3.1 Problem Description.

Let $\mathcal{D} = (\mathbf{d}_1, \mathbf{d}_2, \ldots, \mathbf{d}_n)$ denote the document collection; the first n_l documents are labeled, $\bar{\mathbf{y}}_l = (\bar{y}_1, \bar{y}_2, \ldots, \bar{y}_{n_l})$, where each $\bar{y}_i \in \{-1, +1\}$. Let $\mathcal{U} = (u_1, u_2, \ldots, u_m)$ denote the m users who provided feedback on $\mathcal{D}$. Let $F \in \mathbb{R}^{m \times n}$ be the user feedback matrix. $F_{i,j}$ indicates the feedback of the user u_i for the document d_j. It can be a binary number, i.e., either $+1$ or -1, for binary relevance judgments, or a categorical number, such as $1, 2, 3, \ldots$, for user rating information. $F_{i,j} = 0$ if no feedback is provided by u_i for d_j. The goal of collaborative text categorization is to exploit the feedback information encoded in F to classify the documents in the document collection $\mathcal{D}$.

A straightforward approach is to first represent each document $\mathbf{d}_i$ by its user feedback $\mathbf{d}_i = (F_{1,i}, F_{2,i}, \ldots, F_{n,i})$ and then apply standard supervised learning methods using the feedback information. The underlying assumption is that two documents are likely to share the same category if their user feedbacks are similar, which we refer to as the "**user feedback assumption**". We examine this assumption using the movie rating data in Sect. 4. The important challenge in collaborative text categorization is the "*missing user*" problem. For users who have not provided feedback for any labeled document, their feedback information cannot be exploited in standard supervised learning and therefore will be completely wasted. We refer to the users who provide no feedback for any labeled documents as the *missing users*.

3.2 Semi-supervised Learning Approaches

We discuss four semi-supervised approaches for collaborative text categorization.

Label Propagation. One difficulty from the missing user problem is that the feedback of the missing users cannot be used to assess the similarity between the labeled and unlabeled documents. To alleviate this problem, we can employ the label propagation approach. The key idea behind label propagation is to propagate the class labels of documents to neighbors that share similar feedback ratings from a large number of users. Thus, given two documents d_i and d_j sharing no common users, we may still be able to infer the category of d_j from d_i if there is a sequence of documents $d_i, d_{p_1}, d_{p_2}, \ldots, d_{p_l}, d_j$ such that every two consecutive documents in the sequence share large similarity. A potential problem with label propagation is that there may be a sequence of consecutively-similar documents for two documents with completely opposite user feedback. This issue becomes more serious when the similarity information is sparse, as is often the case in collaborative text categorization.

User Clustering. The second approach toward the missing user problem is to reduce the number of distinct users. We can cluster a large number of users into a relatively small number of user clusters and then represent each document by the aggregated feedback from each user cluster. In this study, we choose the probabilistic spectral clustering algorithm [6] because of its effectiveness and *soft* cluster membership assignments that is better for capturing the feedback of users with mixed interests. One difficulty in the user clustering approach is how to determine the number of clusters. A small number of user clusters may not capture the diversity of user interests while a large number of user clusters may not alleviate the missing user problem sufficiently. Cross validation may be employed, but it is unlikely that cross validation will reliably identify the optimal number with the small number of labeled documents.

The Kernel Method. The key idea of the kernel method is to improve the estimation of document similarity by exploiting the user similarity. Two documents d_i and d_j that share no common users may have zero similarity computed as $S^d_{i,j} = \mathbf{d}_i^\top \mathbf{d}_j$. Based on the intuition that d_i and d_j are similar if the two sets of users who provided feedback for d_i and d_j have similar feedback, we can improve document similarity by a kernel similarity measure as $\tilde{S}^d_{i,j} = \mathbf{d}_i^\top S^u \mathbf{d}_j$ with the user similarity matrix S^u. We refer to $\tilde{S}^d_{i,j}$ as the "*transformed document similarity*" as opposed to standard S^d. Such a kernel can then be incorporated into a support vector machine for document classification. A potential problem with the proposed kernel is the overestimated document similarities. This problem could be partially addressed by the user clustering approach, which unfortunately has its own significant weaknesses as described above.

Manifold Regularization. In a linear classifier, the most important parameters are the weights $\mathbf{w} = (w_1, w_2, \ldots, w_m)$ assigned to the m users. Given the limited number of labeled documents, a typical algorithm for maximum margin classification (e.g., SVM) would assign zero weights to these missing users and lead to a classifier that ignores the feedback from these missing users. Given the labeled documents $\mathbf{d}_1, \mathbf{d}_2, \ldots, \mathbf{d}_{n_l}$, a standard support vector machine is formulated as:

$$\begin{aligned} &\min_{\mathbf{w},\varepsilon} \frac{1}{2}\sum_{k=1}^{m} w_k^2 + C\sum_{i=1}^{n_l}\varepsilon_i \\ &\text{s. t. } y_i(\mathbf{w}^\top \mathbf{d}_i - b) \geq 1 - \varepsilon_i,\ \varepsilon_i \geq 0,\ i = 1, 2, \ldots, n_l\ . \end{aligned}$$

Clearly, zero weights are assigned to the missing users because the conventional regularizer, $l(\mathbf{w}) = \sum_{k=1}^m w_k^2$, encourages w_k to be set to zero whenever possible.

Based on manifold regularization [3], our approach alleviates the missing user problem by replacing $l(\mathbf{w})$ with $l_m(\mathbf{w}) = \sum_{i,j=1}^m S^u_{i,j}(w_i - w_j)^2 = \mathbf{w}^\top L^u \mathbf{w}$ where the graph Laplacian $L^u = D^u - S^u$ and the diagonal matrix D^u has $D^u_{i,i} = \sum_{j=1}^n S^u_{i,j}$. The regularizer $l_m(\mathbf{w})$ measures the inconsistency between $\mathbf{w}$ and S^u. By minimizing $l_m(\mathbf{w})$, we enforce similar weights for those users sharing a large similarity in their interests. Hence, the missing users can still be assigned significant weights if they share large similarity with the users who did provide feedback for the labeled documents. $l_m(\mathbf{w})$ leads to the following problem:

$$\begin{aligned} &\min_{\mathbf{w},\varepsilon} \frac{1}{2}\mathbf{w}^\top L^u \mathbf{w} + C\sum_{i=1}^{n_l}\varepsilon_i \\ &\text{s. t. } y_i(\mathbf{w}^\top \mathbf{d}_i - b) \geq 1 - \varepsilon_i,\ \varepsilon_i \geq 0,\ i = 1, 2, \ldots, n_l\ . \end{aligned} \tag{1}$$

It is not difficult to compute the dual form of the above problem, i.e.,

$$\begin{aligned} &\max_{\alpha} \sum_{i=1}^{n_l}\alpha_i - \frac{1}{2}\alpha^\top X^\top [L^u]^{-1} X\alpha \\ &\text{s. t. } \sum_{i=1}^{n_l}\alpha_i y_i = 0,\ 0 \leq \alpha_i \leq C,\ i = 1, 2, \ldots, n_l\ . \end{aligned} \tag{2}$$

where $X = (\mathbf{d}_1, \mathbf{d}_2, \ldots, \mathbf{d}_n)$ represents the document collection. We make the transformation $\mathbf{d}_i \leftarrow [L^u]^{-1/2}\mathbf{d}_i$ $(i = 1, 2, \ldots, n_l)$ which turns the dual formulism in (2) into the dual form of the standard SVM. We add a small identity matrix $\delta I_{n_l \times n_l}$ $(\delta \ll 1)$ to L^u to avoid a singularity graph Laplacian.

4 Experiments

We evaluate four methods for semi-supervised collaborative text categorization and address two questions: (1) How effective is collaborative text categorization in comparison to content-based approaches? (2) How effective are the various proposed algorithms in the study?

We employ the MovieRating dataset[3] which consists of 1682 movies in 19 categories rated by 943 users with the integer ratings ranging from 1 (worst) to 5 (best) or 0 for unavailable ratings. We select the four most popular categories: "Action", "Comedy", "Drama", and "Thriller". The resulting dataset has 1422 movies each represented by a vector of ratings from 943 users. We also download the movie keywords from the online movie database[4] resulting in 10116 unique words for 1422 movies. We use the linear SVM as the baseline implemented in SVM-light[5]. For every category, we compute the $F1$ metric by averaging $F1$ scores over 40 independent trials. We compute both the movie similarity matrix S^d and the user similarity matrix S^u by the linear kernel similarity.

4.1 Effectiveness of Collaborative Text Categorization

The number of unique movie keywords is significantly greater than the number of users who rated the movies. This raises the concern that the user feedback representation may be less rich than the keyword representation and thus collaborative text categorization may not be as effective as content-based text categorization. We summarize in Table 1 the $F1$ results for both collaborative text categorization and content-based text categorization. In all cases, collaborative text categorization is considerably more effective.

We then verify the "user feedback assumption", which is that two documents tend to be in the same category if they have similar user ratings. Figure 1 shows the distribution of the probability for two movies to share the same category w.r.t the pairwise movie similarity based on user ratings. The high end of the distribution appears to be spiky because few document pairs are able to achieve a similarity score greater than 0.6. The overall trend proves that our assumption is reasonable for document categorization.

4.2 Semi-supervised Collaborative Text Categorization

To study the effectiveness of semi-supervised collaborative approaches, we randomly select 10, 20, 30, and 40 movies for training. To avoid a skewed number

[3] `http://www.cs.usyd.edu.au/~irena/movie_data.zip`

[4] `http://us.imdb.com/`

[5] `http://svmlight.joachims.org/`

Table 1: F1 scores of collaborative and content-based categorization.

		# Training examples			
Cat.	Classif.	20	40	80	100
Act.	collab.	0.291	0.337	0.353	0.344
	content	0.170	0.229	0.346	0.343
Com.	collab.	0.349	0.399	0.436	0.459
	content	0.321	0.345	0.363	0.374
Dra.	collab.	0.509	0.603	0.645	0.665
	content	0.398	0.416	0.546	0.575
Thri.	collab.	0.159	0.184	0.209	0.207
	content	0.118	0.131	0.153	0.165

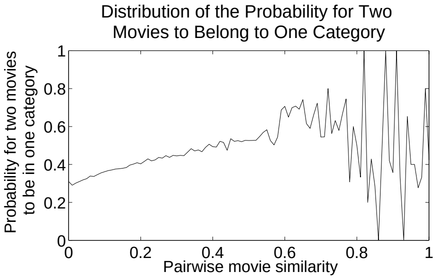

Fig. 1: Distribution of the probability that two movies are in one category.

of positively-labeled examples, we set the number of positively-labeled examples to be same as the number of negatively-labeled examples. We first examine the missing user problem. Table 2 shows the percentage of users who did not rate any of the labeled movies (i.e., the fraction of *missing users*). Clearly the missing user problem can be significant when the number of labeled examples is small.

Table 2: The fraction of “missing users” in four categories.

	# Training examples			
Category	10	20	30	40
Action	60.8%	42.2%	34.2%	20.4%
Comedy	57.4%	41.1%	33.2%	23.0%
Drama	57.0%	44.9%	34.9%	22.5%
Thriller	61.4%	41.3%	31.3%	21.1%

Table 3: F1 scores of user clustering with 10 training examples.

# Clu.	Action	Comedy	Drama	Thriller
5	0.140	0.329	0.431	0.120
10	0.117	0.308	0.396	0.121
30	0.113	0.311	0.427	0.120
50	0.119	0.290	0.436	0.112
100	0.121	0.319	0.427	0.118

We then examine the classification accuracy of the four discussed methods. Tables 4(a) to 4(d) summarize the $F1$ results of the four methods and linear SVM for the chosen categories. Our implementation of label propagation is based on [13]. We set the number of user clusters to be 5 for the user clustering approach. From the results in Tables 4(a) to 4(d), we first observe that among the four approaches, the manifold regularization approach is the *only* one that consistently improves the performance of the linear SVM. For a number of cases, manifold regularization yields considerable improvements.

The second observation drawn from Tables 4(a) to 4(d) is that the other three methods: user clustering, the kernel method, and label propagation, all perform significantly worse than the linear SVM for all categories but Comedy. The failure of *label propagation* may be attributed to a sparse similarity matrix in which more than 2/3 of the pairwise similarity is less than 0.1 and only 0.5% percentage of the pairwise similarity is significantly large (i.e., > 0.5).

Table 4: F1 measure of the four semi-supervised learning methods for chosen categories.

(a) Action Category

	# Training examples			
Classifier	10	20	30	40
SVM	0.219	0.291	0.308	0.344
Manifold Reg.	0.264	0.341	0.375	0.381
User Cluster.	0.140	0.140	0.140	0.140
Kernel	0.146	0.178	0.204	0.207
Label Prop.	0.155	0.147	0.142	0.135

(b) Comedy category

	# Training examples			
Classifier	10	20	30	40
SVM	0.308	0.349	0.394	0.400
Manifold Reg.	0.338	0.370	0.421	0.423
User Cluster.	0.329	0.339	0.343	0.350
Kernel	0.322	0.351	0.355	0.386
Label Prop.	0.300	0.296	0.296	0.295

(c) Drama Category

	# Training examples			
Classifier	10	20	30	40
SVM	0.507	0.509	0.562	0.603
Manifold Reg.	0.519	0.568	0.589	0.643
User Cluster.	0.431	0.484	0.493	0.534
Kernel	0.354	0.496	0.530	0.537
Label Prop.	0.353	0.353	0.360	0.353

(d) Thriller Category

	# Training examples			
Classifier	10	20	30	40
SVM	0.144	0.159	0.171	0.184
Manifold Reg.	0.161	0.169	0.196	0.201
User Cluster.	0.120	0.121	0.122	0.127
Kernel	0.124	0.125	0.125	0.126
Label Prop.	0.136	0.129	0.125	0.129

Such a sparse similarity matrix is unlikely to reveal any clustering structure of movies. One major problem with the *user clustering* method is the difficulty in determining the appropriate number of clusters. Table 3 shows the F1 scores of user clustering with different cluster numbers. Regardless of cluster numbers, the algorithm is unable to consistently outperform the linear SVM model. The failure of *the kernel method* may be explained by the overestimated movie similarity which can lead to the skewed spectrum of the similarity matrix. Figure 2 shows the top 100 eigenvalues of the transformed similarity matrix and the original similarity matrix. Clearly the spectrum of the original similarity matrix is much flatter than the transformed one. This is consistent with our hypothesis.

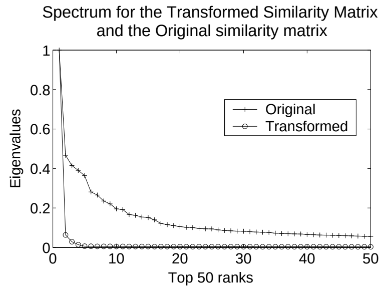

Fig. 2: Spectrum of the original and transformed movie similarity matrix

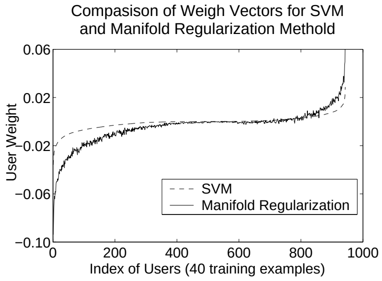

Fig. 3: User weights by linear SVM and manifold regularization

Finally, we examine the missing user problem. Figure 3 shows the weights of 943 users computed by the linear SVM and the manifold regularization method. The horizontal axis (i.e., the user index) is sorted in the ascending order of their weights that are computed by the linear SVM. Evidently most users are assigned zero weights by the linear SVM because of the missing user problem while most users are assigned non-zeros weights by the manifold regularization method which is more effective in alleviating the missing user problem.

5 Conclusions

In this paper, we study the problem of collaborative text categorization by using user feedback as the basis for classifying documents. Our experiments validate the basic assumption behind collaborative text categorization. Moreover, this work evaluated four algorithms for semi-supervised collaborative text categorization and our empirical finding is that manifold regularization is the most effective among the four competitors and is a considerable improvement over traditional content-based categorization.

References

1. R. Angelova and G. Weikum: Graph-based text classification: learn from your neighbors. Proceedings of SIGIR. (2006)
2. C. Apte, F. Damerau and S. Weiss: Automated Learning of Decision Rulesfor Text Categorization. ACM Transactions on Information Systems. **12(3)** (1994)
3. M. Belkin, P. Niyogi and V. Sindhwani: Manifold Regularization: a Geometric Framework for Learning from Examples. Technical Report. (2004)
4. A. Dayanik, D. Lewis, D. Madigan, V. Menkov and A. Genkin: Constructing informative prior distributions from domain knowledge in text classification. SIGIR'06.
5. C. H. Hoi and M. R. Lyu: A novel log-based relevance feedback technique in content-based image retrieval. Proceedings of ACM Multimedia. (2004)
6. R. Jin, C. Ding and F. Kang: A Probabilistic Approach for Optimizing Spectral Clustering. Advances in NIPS. **18** (2006)
7. T. Joachims: Text categorization with support vector machines: learning with many relevant features. Proceedings European Conference on Machine Learning. (1998)
8. T. Joachims: Optimizing search engines using clickthrough data. Proceedings of the eighth ACM SIGKDD. (2002)
9. T. Joachims: Transductive Inference for Text Classification using Support Vector Machines. Proceedings of ICML. (1999)
10. S. Robertson and J. Callan: Routing and filtering. TREC: Experiment and Evaluation In Information Retrieval. (2006) MIT Press
11. G. Xue, H. Zeng, Z. Chen, W. Ma, H. Zhang and C. Lu: Implicit link analysis for small web search. Proceedings of SIGIR. (2003)
12. J. Zhang, R. Jin, Y. Yang and A. Hauptmann: Modified Logistic Regression: An Approximation to SVM and its Applications in Large-Scale Text Categorization. Proceedings of ICML. (2003)
13. X. Zhu, Z. Gharahmani and J. Lafferty: Semi-supervised learning using Gaussian fields and harmonic functions. Proceedings of ICML (2003).