Co-Operative Text Categorization And Multilabel Classification

Mohana A.L[1], Subathra M[2], S.R.Sendhil Kumar[3], R.K.Kapilavani[4]
Department of Computer Science and Engineering
[1,2] Student, Prince Shri Venkateshwara Padmavathy Engineering College
[3,4] Assistant Professor, Prince Shri Venkateshwara Padmavathy Engineering College

Abstract- *The proposed system is going to deal with a very challenging task of automatically generating presentation slides for academic papers. The wide availability of web documents in electronic forms requires an automatic technique to label the documents with a predefined set of topics, what is known as automatic Text Categorization (TC). Over the past decades, it has been witnessed a large number of advanced machine learning algorithms to address this challenging task. The generated presentation slides can be used as drafts to help the presenters prepare their formal slides in a quicker way.Documents are usually represented by the "bag-of-words": namely, each word or phrase occurs in documents once or more times is considered as a feature.It first employs the regression method to learn the importance scores of the sentences in an academic paper, and then an effective algorithm is developed for multi-label classification with utilizing those data that are relevant to the targets. The key is the construction of a coefficient-based mapping between training and test instances, where the mapping relationship exploits the correlations among the instances, rather than the explicit relationship between the variables and the class labels of data and manufactures the multilevel classifier on the adapted low-dimensional information portrayals at the same time. It initially utilizes the relapse technique to take in the significance scores of the sentences in a scholastic paper, and after that adventures the Latent Dirichlet Allocation (LDA) strategy to create very much organized slides by choosing and adjusting key expressions and sentences to a point for the slide. We prepare a sentence scoring model in light of naïve Bayes classifier and utilize the LDA technique to adjust and remove key expressions and sentences for producing the slides. Exploratory outcomes demonstrate that our strategy can produce much preferred slides over conventional strategies.*

I. INTRODUCTION

Traditional supervised learning is one of the mostly-studied machine learning paradigms, where each real-world object is represented by a single instance and associated with a single label. Although traditional supervised learning is prevailing and successful, there are many learning tasks where the above simplifying assumption does not fit well, as real-world objects might be complicated and have multiple semantic meanings simultaneously. In contrast to traditional supervised learning, in multi-label learning each object is also represented by a single instance while associated with a set of labels instead of a single label. The task is to learn a function which can predict the proper label sets for unseen instances. The emerging area of multi-label learning, fundamentals on multi-label learning including formal definition and evaluation metrics are given. Technical details of up to eight representative multi-label algorithms are scrutinized under common notations with necessary analyses and discussions. Several related learning settings are briefly summarized. Online resources and possible lines of future researches on multi-label learning are discussed. However, the generality of multi-label learning inevitably makes the corresponding learning task much more difficult to solve. Actually, the key challenge of learning from multi-label data lies in the overwhelming size of output space, i.e. the number of label sets grows exponentially as the number of class labels increases. Effective exploitation of the label correlations information is deemed to be crucial for the success of multi-label learning techniques. Existing strategies to label correlations exploitation could among others be roughly categorized into three families, based on the order of correlations that the learning techniques have considered. The task of multi-label learning is tackled in a label-by-label style and thus ignoring co-existence of the other labels, such as decomposing the multi-label learning problem into a number of independent binary classification problems. The task of multi-label learning is tackled by considering pairwise relations between labels, such as the ranking between relevant label and irrelevant label or interaction between any pair of labels. As label correlations are exploited to some extent by second-order strategy, the resulting approaches can achieve good generalization performance. The task of multi-label learning is tackled by considering high-order relations among labels such as imposing all other labels' influences on each label or addressing connections among random subsets of labels. Apparently high-order strategy has stronger correlation-modeling capabilities than first-order and second-order strategies, while on the other hand is computationally more demanding and less scalable. In traditional supervised learning, generalization performance of the learning system is evaluated with conventional metrics such as accuracy. Recent researches indicate that correlations among labels might be asymmetric, i.e.the influence of one label to the other one is not necessarily be the same in the inverse direction, or local, i.e. different instances share different label correlations with few correlations being globally applicable.

II. RELATED WORK

Implementation is the most crucial state in achieving a successful system and giving the user's confidence that the new system is workable and effective. Implementation of a modification application is to replace an existing one. This type of conversion is relatively easy to handle, provided there are no major changes in the system. Each program is tested individually at the time of development using the data and has verified that this program linked together in the way specified in the programs specification, the computer system and its environment is tested to the satisfaction of the user. A simple operating procedure is included so that the user can understand the different functions clearly and quickly. Initially at a first step, the executable form of the application is to be created and loaded in the common server machine which is accessible to the entire user and the server is to be connected to a network. The final stage is to document the entire system which provides components and the operating procedures of the system. Implementation is the stage of the project when the theoretical design is turned out into a working system. The implementation stage involves careful planning, investigation of the existing system and it's constraints on implementation, designing of methods to achieve change over and evaluation of change over methods, implementation is the process of converting a new system design into operations. It is the phase that focuses on user training, site preparation and file conversion for installing a candidate system.

Clustering is a useful statistical tool in computer vision and machine learning. It is generally accepted that introducing supervised information brings remarkable performance improvement to clustering.However, assigning accurate labels is expensive when the amount of training data is huge.In order to alleviate the labeling burden, semi-supervised clustering algorithms are proposed [13]. Only a part of the training descriptors are labeled and the rest of the descriptors are unlabeled. Although the semi-supervised framework requires less labeling operations than the fully supervised one, the labeling work is intolerable for big data-based applications. Another shortcoming of the semi-supervised framework is that the performance is sensitive to labeled data. This leads to unstable clustering performance. It is worth noticing a previous work titled "kernel methods for weakly supervised mean shift clustering.To successfully avoid instance-level labeling burden, bag level labeling has been proposed [18]. It assumes a bag of descriptors share one and only one label. Thus, labels are provided for bags instead of the individual instance-level descriptors. a weakly supervised clustering algorithm for image semantic segmentation with image-level labels, i.e., collaboratively performing image segmentation and tag alignment with those regions is proposed. The approach is motivated by the observation that super pixels belonging to an object class usually exist across multiple images and hence can be gathered via clustering. Noticeably, this approach cannot preserve the locality and discriminative of samples in each bag. Besides, this algorithm is specifically designed for semantic image segmentation. It cannot be applied to the various computer vision applications directly.

C Multi-label learning refers to problems where an instance an be assigned to multiple classes. This differs from multi-class learning where every instance can be assigned to only one class even though the number of classes is more than two. The essential difference between multi-class learning and multi-label learning is that classes in multiclass learning are assumed to be mutually exclusive while classes in multi-label learning are often correlated. A serious problem with existing approaches is that they are unable to exploit correlations between class labels. A serious problem with existing approaches is that they are unable to exploit correlations between class labels framework is proposed. The proposed framework takes into account the simultaneous propagation of multiple labels. Multi-label learning approaches learn a ranking function of class labels from the labeled examples and apply it to order the class labels for the given test examples. The main idea is to propagate the labels from training examples to test examples through their similarities. The label information propagated from different training examples are then accumulated and used as the basis for scoring the class labels of test examples. a linear programming problem with an exponential number of constraints that cannot be practically solved using standard techniques is also formulated. Based on properties of submodular functions an algorithm is formulated that can solve this problem exactly and efficiently. Correlated label propagation is more effective than the statistical translation model for automatic image annotation is also shown.

III. PROPOSED SYSTEM

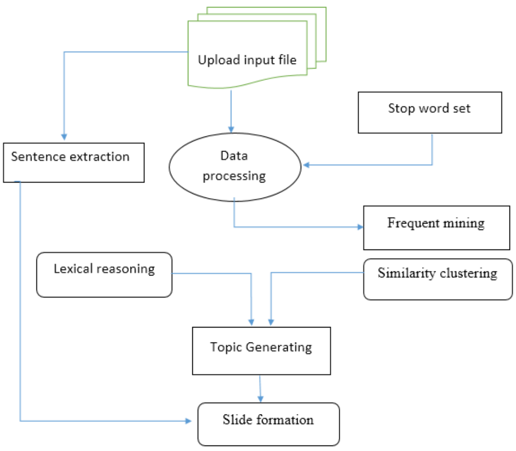
The generated presentation slides can be used as drafts to help the presenters prepare their formal slides in a quicker way.

A novel system is proposed to address this task. It first employs the regression method to learn the importance scores of the sentences in an academic paper, and then exploits the LDA method to generate well-structured slides by selecting and aligning key phrases and sentences.We train a sentence scoring model based on naïve Bayes classifier and use the LDA method to align and extract key phrases and sentences for generating the slides. Experimental results show that our method can generate much better slides than traditional methods Each sentence in a paper is learned by using the support vector regression (SVR) model.

The presentation slides for the paper are generated by using the integer linear programming (ILP) model.The slides are generated automatically from the academic papers.The generated slides for the presentation will have only important points and all necessary figures, tables and graphs.

IV. ARCHITECTURE

In this section, we present the main components of multilabel learning has been analysed and then th e steps involved in converting pdf document to presentation slides has been discussed.



FiFig. 1a

The generated presentation slides can be used as drafts to help the presenters prepare their formal slides in a quicker way.

A novel system is proposed to address this task. It first employs the regression method to learn the importance scores of the sentences in an academic paper, and then exploits the LDA method to generate well-structured slides by selecting and aligning key phrases and sentences.

We train a sentence scoring model based on naïve Bayes classifier and use the LDA method to align and extract key phrases and sentences for generating the slides. Experimental results show that our method can generate much better slides than traditional methods

V.SUPERVISED LEARNING ALGORITHMS

Let Y = {1,...,K} denote the label set and B = {Bi}i=1,...,n denote the training bags. Each bag Bi contains multiple descriptor instances Bi = {xij}j=1...ni . On one hand, each bag Bi is associated with multiple human-specified bag-level labels Yi = {yk i }k=1,...,K, where yk i is either 0 or 1. We call a label valued one as legal label and illegal label if the label is not one. We denote the set of all the instances as X = {xij}i=1,...,n,j=1,...,ni . Note that the instance-level labels {yij} are unavailable at the training stage. Our goal is to cluster the instances into discriminative codebooks without violating the bag-level constraints. Aiming at this, we learn the clustering function C(·) which returns the cluster that an instance belongs to. C(·) preserves the locality of data such that if pairwise instances are in the same cluster, they should be neighbors in a certain metric. Based on C(·), we can assume a matching function F(·) between an instance and a class. F(x0, k) is the probability that the instance x0 matches the kth class. It is straightforward to define F(x0, k) as the density of class labels of the cluster x0 belonging to F(x0, k) = xij∈C(x0) yij = k |C(x0)| . (1) Based on the definition of F(·), we further define L(·) as the discrimination that predicts the label of an instance L(x0, y0) = F(x0, y0) − arg max k∈Y,k =y0 F(x0, k). (2) In weakly supervised clustering circumstance, the groundtruth instance-level labels are unknown. Thus, the clustering procedure leads to the following objective function: y∗ ij, C(·) ∗ = arg min yij,C(·) n i=1 ni j=1 −L xij, yij s.t. ∀y j i = 1, ni j=1 [yij = k] > 0 ∀y j i = 0, ni j=1 [yij = k] = 0. (3) A graphical illustration of the two constraints in (3) is shown in Fig. 1. Based on the bag-level labels, we can divide the entire sample into K partitions, each of which contains ni samples. The first constraint means that if any bag-level label k from {1, 2,...,K} corresponding to the ith bag is legal, then there is at least one sample in this bag belonging to label k ∈ {1, 2,...,K}. The second constraint means that if any bag-level label from {1, 2,...,K} corresponding to the ith bag is illegal, then there is definitely no sample in this bag belonging to label k ∈ {1, 2,...,K}. The optimization procedure of (3) aims to minimize the global loss [the first term of (3)] in discrimination by considering the weakly supervised constraints [the second and the third terms of (3)]. The optimization extends the multiple instance constraints in two aspects: 1) multiple (instead of single) labels are associate with a bag and 2) positive and This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination. Clustering is an indispensable component in machine learning and computer vision. In this paper, we propose a WSRF that allows semantic clustering under bag-level supervision. We first formulate the weakly supervised clustering into a constrained objective function minimization problem. Thereby, a discriminative codebooks can be achieved without violating the bag-level labels. Then, a WSRF is proposed that formulates the weakly supervised clustering into a constrained margin maximization problem. And a DA algorithm is employed to solve the margin maximization problem. Finally, the instances are clustered with guidance of the given bag-level labels. The experimental results demonstrate that the resulted clusters preserve data locality and semantic discrimination. Moreover, our clustering algorithm can facilitate three real-world computer vision task remarkably.instead of the weakly supervised clustering problem.

VI. EFFICIENT LEARNING ALGORITHMS

Efficient Learning Algorithm section, we show that when the label kernel function Ω(x) is a concave function, there is a simple and greedy algorithm for finding the optimal solution to the problem in Equation 8. Furthermore, the solution only depends on the relative order of weights {αk}m k=1, and is independent of their exact values. The algorithm for estimating label confidence scores z is summarized in Figure 2. This greedy algorithm is based on the following theorem from discrete optimization [16]: Input • xt: the test example • α1 ≥ α2 ≥ ... ≥ αm > 0 Output: optimal label scores (zt,1,...,zt,m) for xt For k = 1,. . . ,m • Let class label set Tk = {1, 2,...,k}. • f(Tk) = n i=1 K(xi, xt)Ω(tT (Tk)t(Si)) • zt,k = f(Tk) − f(Tk−1) Figure 2. Algorithm for finding the optimal solution to Equation 8 Given: (1) a finite set N , (2) a set function f : 2N → R with f(φ) ≥ 0, and (3) a weight vector w ∈ R|N |. Then, the linear programming problem: max w∈R|N | wT x s. t. ∀A ⊆ N , e∈A x(e) ≤ f(A) ∀e ∈ N , x(e) ≥ 0 can be solved by the following greedy algorithm if the set function f is submodular: • Sort elements of N as w(e1) ≥ w(e2) ≥ ... ≥ w(en) • Let V0 = φ For i =1,. . . ,n, let Vi = Vi−1 + ei, and x(ei) = f(Vi) − f(Vi−1). The validity of applying the above theorem to our problem defined in Equation 8 relies on the fact that the function f in our algorithm, i.e., f(u) = n i=1 K(xt, xi)Ω(uT ti) is interesting that the kernel-based k Nearest-Neighbor is a special case of the algorithm given by setting Ω(x) = x. 1 This is because zt,k = f(Tk) − f(Tk−1) = n i=1 K(xt, xi) t(Tk) − t(Tk−1))T t(Si)) = n i=1 K(xt, xi) eT k t(Si)) = n i=1 K(xt, xi)I(k ∈ Si)

VII. EXPERIMENTAL STUDY

Experimental Settings

1) Datasets: We carried out the comparison experiments on eight multilabel datasets with different types and sizes. They are Arts, Education, Entertainment, Health, Recreation, Reference, Science, and Social. These datasets were frequently used to validate performance of multilabel classification models in the literature one may observe that the multilabel datasets vary from the quantities of labels and differ greatly in the sizes of variables.

2) Evaluation Criteria: For the traditional learning algorithms, their effectiveness or performance is often evaluated by the criterion of precision or accuracy, which simply denotes the number of correctly predicted instances relative to the total number of instances in a test dataset. However, it is not appropriate to the case of multilabel learning, because the output of a multilabel classifier involves multiple class labels at the same time. In our experiments, we adopted four different criteria to evaluate the performance of the multilabel learning methods. They are Hamming loss (HL), ranking loss (RL), one error (OE), and average precision (AP) [1]. 1) Hamming Loss: This criterion is defined as the percentage of labels which are classified incorrectly by the multilabel classifiers. The misclassified labels include the relevant labels that have not been predicted and the irrelevant labels that have been predicted. 2) Ranking Loss: Ranking the predicted labels is important in multilabel learning, especially for the label ranking learning methods, because we always expect the relevant labels would be outputted first and their ranks should be higher then those irrelevant ones. RL is such a criterion, which refers to the mis-ordered degree that the irrelevant labels are ranked higher than the relevant ones in the predicted results.

3) One Error: Like the criterion of accuracy in the traditional learning tasks, OE also simply summarizes the ratio of how many times the most relevant label (i.e., the top-ranked label in each predicted results) is irrelevant to the true labels of the instances. 4) Average Precision: This evaluation measurement places more emphasis on the relevant labels. It refers to the percentage of relevant labels among all labels that are ranked above. 3) Comparison Methods: To make a comparison roundly, we carried out three different groups of experiments. The first group compared SWIM to the statistical multilabel learning algorithms. The statistical learning techniques recently have been extensively studied in multilabel learning. Typical examples include PLS and CCA. Both of them are good at measuring the correlations of two sets of variables. We took PLS [44], sPLS [16], PPLS-MD [10], CCA [31], and

sCCA [15] as our comparing methods. In the second group experiment, SWIM is used to compare with the instance-based learning methods. As mentioned above, kNN has been extensively studied in multilabel learning. SWIM is a general framework of the instance-based learning methods. Thus, this group aims at showing the effectiveness of SWIM in comparing with the kNN-based ones, including LPkNN [17], BRkNN [6], and MLkNN [5]. The third group made a comparison of our method to classical multilabel learning ones, such as BP-MLL [18], AdaBoost.MH [4], HOMER [49], MLStacking [3], pruned problem transformation [50], and ClassifierChain [51]. These learning algorithms stand for different learning techniques and have relatively better performance. For instance, BP-MLL [18] and AdaBoost.MH [4] extend the traditional neural network and AdaBoost learning algorithms to fit the multilabel cases, while the rest ones belong to the problem transformation kind of multilabel learning. More details of these learning methods are provided in the related work section or references therein. The proposed algorithm and the statistical ones were implemented with MATLAB. For the rest multilabel classifiers, i.e., the instance-based and classical ones, we compared them under the MULAN package [52], where the off-the-shelf learning algorithms are contained. All experiments were conducted on a Pentium IV, with a CPU clock rate of 1.7 GHz, 1 GB main memory. During the whole experiments, the tenfold crossvalidation was adopted, and the final results were the average values over the ten rounds. Experimental Results and Discussion

1) Comparing to the Statistical Learning Methods: Since SWIM exploits the penalized PLS to construct models, in the first group experiment we made a comparison of SWIM to the statistical learning methods, including PLS, sPLS, PPLS-MD, CCA, and sCCA. The reason of choosing CCA and its variants is that they can also be used to obtain a common and latent space between two sets of variables. The difference of CCA to PLS is that the variables in the latent space identified by CCA have maximal correlations, rather than the covariances for PLS. It should be pointed out that the statistical learning methods explore the correlations between the variable space and the label space while SWIM extracts the mapping function g from the instances.During the whole experimental procedures, the parameters involved within the learning algorithms were assigned to the same values for the sake of impartial comparison. For example, in the sparse variants of CCA and PLS (i.e., SWIM, sCCA, PPLS-MD, and sPLS), the regularization parameters λw and λc were equal to 0.1. In addition, all learning algorithms chose the same number of latent variables (i.e., $m = 25$) to build classification models. The experimental results on the evaluation criteria where the notation "↓" (or "↑") indicates that the lower (higher) of the curve, the better performance of the classifier.We know that SWIM surpassed the statistical learning algorithms, except PPLS-MD, in most cases. For example, SWIM had the lowest HL on the datasets. Besides, SWIM achieved the best performance of RL, OE, and AP on seven datasets in comparing to other statistical learning methods. For the evaluation criterion of RL (OE), SWIM was slightly worse than sPLS on Education (Arts). Even so, the differences between them were very small, and they were not significantly different to each other if a statistical t-test was considered. Comparing to PPLS-MD, SWIM achieved slightly poor performance. For example, the AP of PPLS-MD was higher than SWIM on five datasets. Similarly, PPLS-MD outperformed SWIM for the criteria of OE and RL in several cases. Even so, SWIM had lower HL than PPLS-MD on six over eight datasets. The underlying reason is that PPLS-MD exploits the sparse property of the label space to explore the mapping relationship f , while SWIM does not consider the sparse property of the label space. It just models the mapping relationship g of the instances in the variable space, and then applies g to the label space directly. Intuitively, SWIM, PPLS-MD, and sPLS have similar properties, because they are the sparse variants of PLS, notwithstanding their purposes are different. This assertion was demonstrated by the experiments. SWIM and sPLS exhibited analogous behaviors. For instance, both SWIM and sPLS achieved better or worse performance on the evaluation criteria in most cases. Comparing with PLS, sPLS performed relatively worse on the Entertainment dataset. This is reasonable because sPLS may lose some information when making the model sparse. An interesting fact is that CCA and sCCA had relatively poor performance in comparing with PLS and its variants. The underlying reason is that CCA and its variants aim at identifying the latent variables such that their correlations are maximal, while PLS tries to discover the latent variables with maximal covariances. Indeed, the maximal correlations do not stand for good discriminant capabilities in classification and prediction.

2) Comparing to the kNN-Based Learning Methods: As mentioned above, SWIM takes the weights of instances into account when predicting the class labels. It is a general framework of the instance-based learning methods to some extent.

VIII. CONCLUSION

We train a sentence scoring model based on naïve Bayes classifier and use the Latent Dirichlet Allocation method to align and extract key phrases and sentences for generating the slides. Experimental results show that our method can generate much better slides than traditional methods. We only consider one typical style of slides that beginners usually use. In the future, we will consider more complicated styles of slides such as styles that slides are not aligned sequentially with the paper and styles that slides have more hierarchies. We will also try to extract the slide skeletons from the human-written slides and apply these slide skeletons to the automatic generated slides. Furthermore, our system generates slides based on only one given paper. Additional information such as other relevant papers and the citation information can be used to improve the generated slides. We will consider this issue in the future.

REFERENCES

[1] Y. Xia et al., "Weakly supervised multilabel clustering and its applications in computer vision," IEEE Trans. Cybern., vol. 46, no. 12, pp. 3220–3232, Dec. 2016.

[2] F. Kang, R. Jin, and R. Sukthankar, "Correlated label propagation with application to multi-label learning," in Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognition., New York, NY, USA, 2006, pp. 1719–1726.

[3] M.-L. Zhang and Z.-H. Zhou, "A review on multi-label learning algorithms," IEEE Trans. Knowl. Data Eng., vol. 26, no. 8, pp. 1819–1837, Aug. 2014.

[4] J. Read, B. Pfahringer, and G. Holmes, "Multi-label classification using ensembles of pruned sets," in Proc. IEEE Int. Conf. Data Min., Pisa, Italy, 2008, pp. 995–1000.

[5] H. Liu, X. Li, and S. Zhang, "Learning instance correlation functions for multilabel classification," IEEE Trans. Cybern., vol. 47, no. 2, pp. 499–510, Feb. 2017.

[6] J. Huang, G.-R. Li, Q.-M. Huang, and X.-D. Wu, "Learning label specific features for multi-label classification," in Proc. IEEE Int. Conf. Data Min., Atlantic City, NJ, USA, 2015, pp. 181–190.

[7] Y.-K. Li, M.-L. Zhang, and X. Geng, "Leveraging implicit relative labelling-importance information for effective multi-label learning," in Proc. IEEE Int. Conf. Data Min., Atlantic City, NJ, USA, 2015, pp. 251–260.

[8] J. Huang, G.-R. Li, Q.-M. Huang, and X.-D. Wu, "Learning label specific features and class-dependent labels for multi-label classification," IEEE Trans. Knowl. Data Eng., vol. 28, no. 12, pp. 3309–3323, Dec. 2016.

[9] X. Zhu, X. Li, and S. Zhang, "Block-row sparse multiview multilabel learning for image classification," IEEE Trans. Cybern., vol. 46, no. 2, pp. 450–461, Feb. 2016.

[10] K. Nag and N. R. Pal, "A multiobjective genetic programming-based ensemble for simultaneous feature selection and classification," IEEE Trans. Cybern., vol. 46,no.2,pp.499–510,Feb.2016.