Web Based services for Big Data Connectivity

Lei Yang[1], Yan Chu[1], Jianpei Zhang[1], Linlin Xia[2], Zhengkui Wang[3], Kian-Lee Tan[4]
[1]College of Computer Science and Technology
Harbin Engineering University, Harbin
[2]School of Automation Engineering
Northeast Dianli University, Jilin
China
[3]Information Communication and Computing
Singapore Institute of Technology
[4]School of Computing, National University of Singapore
Singapore
chuyan@hrbeu.edu.cn, prettylin521@aliyun.com, zhengkui.Wang@singaporetech.edu.sg, tankl@comp.nus.edu.sg

ABSTRACT: *Big data analytics technologies are to extract value from very large data volume, variety of data, and highly rate of data stream. With the fast deployment of cloud services with mobile devices, big data analytics is shifting from personal computer to mobile devices. But, significant limitations of mobile devices are less storage amount and processing power. This paper proposes a big data analytic platform on mobile cloud computing with efficient query execution time by developing MapReduce Transformation Process and query operation based on input query's complexity level. Furthermore, this paper presents the process of RESTful web service for providing seamless connectivity between mobile devices and cloud storage, where store the data and all of necessary processing steps are done by cloud backend. This proposed platform is evaluated by using Census dataset and compared the result with other traditional high level query languages, such as Pig, Hive, and Jaql.*

Keywords: Big Data Analytic, Mobile Cloud Computing, Hadoop Mapreduce, Restful Web Serivce

Received: 17 July 2015, Revised 21 August 2015, Accepted 25 August 2015

1. Introduction

Transfer learning is a new approach of improving the data learning result by utilizing the knowledge from different tasks and domains. The traditional machine learning or data mining approaches require the training and test data to be under the same feature space and the same distribution. Transfer learning, in contrast, allows the domains, tasks and distribution used in training and testing to be different. Specifically, when the training data in the target task are insufficient for a good data modeling, it transfers the useful knowledge from the related auxiliary data from another task to enrich the data features. In this case, more data characteristics are integrated into the data learning facilitating an improved learning results [1], [2].

Much research has been devoted into the transfer learning in the domain of analyzing the long text data. To name a few, [3] proposed source free transfer learning to transfer knowledge from long texts to the long and [4] proposed latent dirichlet allocation to analyze two sets of topics on short and long texts. As the rapid development of Internet, more and more blog-sphere and social networking applications come into being, such as Microblog, Twitter, QQ news and online advertising. These applications exhibit two important features that differs themselves from traditional applications. First, data generated from these applications contain a lot of short texts, which contains rich useful information. Second, the text data vary dramatically every day, in terms of data size and data distribution. On one hand, these new features eventually challenge the traditional data mining and machine learning approaches, as the assumptions made do not hold in these new applications.On the other hand, the existing transfer learning algorithms tailored for long text analysis can not be directly applied in these application as well. The long text data analysis aims to analyze the long text data by utilizing the knowledge learnt from other long text datasets. The techniques are designed to handle the data that is well labeled, naturally compact and structured. However, the short text differs from the long text due to the sparse nature, noise words, syntactical structure and colloquial terminologies used, which result in unsatisfactory analysis results by directly using the transfer learning algorithms in the long text analysis domain.

In order to better utilize the short text data, it is essential to develop new transfer learning techniques in short text analysis. Given the fact that the result learnt from the long text analysis is enriched, one promising approach is to transfer the long text knowledge into the short text analysis. Several algorithms have been proposed under the similar methodology of utilizing the long text information to help the short text analysis. In their work, a major assumption is that source data are provided by the problem designers. This, however, would reduce the usability of these algorithms, as it requires the designers to have a well understanding of the source data. In addition, the prior probability distribution is required. In the big data era, it is significantly difficult to obtain such a data prior probability distribution. Therefore, this calls for the new algorithms that can release the dependency of specific source data and data prior probability distribution knowledge.

In this paper, we propose a novel framework, called AutoTL (Automatic Transfer Learning), which enables an automatic knowledge transferring. AutoTL differs itself by utilizing the informative online information to strengthen the short text analysis without the need of specifying the source training data, when the short text is not well labeled and without knowing the priori probability distribution. Specifically, using the latent semantic analysis techniques, it first extracts the semantic related keywords as the seed feature set between the online web (long text) data and the target data. This can be done by employing the online search engine via inputting the tags extracted from the target data to obtain the most relevant web data. It then builds one undirected graph for the online web data where the nodes represent the tags/labels. Within this graph, it further extracts one subgraph which is able to cover all the seed feature set. In addition, an improved Laplacian Eigenmaps is adopted to map the high-dimensional feature representation to a low-dimensional one. Finally, it classifies the target data through one constraint function of minimizing the mutual information between the instance and feature representation.

Our major contributions are summarized as follows:
We propose AutoTL, an transfer learning algorithm of effective short text analysis. AutoTL is superior to other algorithms, as it automatically identifies the related source data from the rich online information and does not require the system to know the priori probability distribution of the data in advance.

We provide the techniques to integrate the latent semantic analysis into the short text analysis which facilitates an effective learning.

We conduct extensive experimental evaluations and experimental result indicate that our proposed technique is effective, efficient and practical.

The reminder of the paper is organized as follows. Section 2 introduces the automatic transfer learning algorithm. In Section3, we provide experimental evaluation. Section4 presents the existing work. In Section 5, we conclude the paper.

2. Automatic Transfer Learning

2.1 Problem Statement

In the target domain we have few labeled data. $T^i = \{ (x_i, y_i) | \, i=1,2, \ldots, m\}$ is the data set with class label. x_i is the instance of the target domain and y_i is feature representation of the class label. $C = \{C_h \, | \, h = 1, 2, \ldots, N\}$ is the labels of T^i. There are a large number of unlabeled data as the same time. $T^u = \{ x_j \, | j = m+1, m+2, \ldots, m+n\}$, m and n are the numbers of samples, and m <<

n.

In this paper, we will study the problem that how to make use of online information to get a precise target classifier when there is few labeled short texts and no source data and no Priori probability distribution.

2.2 Construct the New Features Representation of the Target Data

As the target data have a few words, which can only provide a small amount of labels, the first we have to do is to expand the target label sets, called feature seed sets. The algorithm in our paper does not have to prepare source data in advance. In order to get auxiliary data, associated with the target areas, similar and useful for the target data, we take full advantage of online information, input keywords extracted from the target field to a search engine, and extract the first few pages of web as source data that semantically related to the target domain.

The next step is to identify the useful tags for the target task from the source domain. Traditional transfer learning methods often use bag of words model to represent the source and the target data. And then calculate the data similarity. According to the similarity, filter the most useful features from the source data. Although this method is simple, it views each word just as an individual, ignoring the relationship between the texts, especially the semantic relationships hidden in context keywords. However, the method of Latent semantic analysis organizes a text into a space semantic structure[5]. It assumes that there is a link, namely a potential semantic structure between the text and words. Use the Latent semantic analysis Statistical analysis of a lot of text sets, and get the context meaning of the words. And then extract the feature seed sets from the source domain.

First constructed the typical word - text matrix: $M = [a_{ij}]_{m \times n}$, a_{ij} is the logarithmic of the frequency that thei -th word appeared in the j^{th} text. Since each word appears only in a small amount of text, M is a high-level sparse matrix. Then apply the technology of singular value decomposition and map words and texts from a high-dimensional space to a low-dimensional latent semantic space. Finally, we can get a new matrix $\tilde{M}$.

$$\tilde{M} = U \tilde{\Sigma} V^T \approx U \Sigma V^T = M \qquad (1)$$

In the formula (1), U and V are orthogonal matrix. $UU^T = VV^T = \mathrm{I}$,

Σ= diag $(a_1, a_2, \ldots, a_k, \ldots, a_v)$ is a Diagonal matrix and $a_1, a_2, \ldots, a_k, \ldots, a_v$ are the singular values of M.

In the matrix $\tilde{M}$, the weight in thei-th row and **j** -th column represents the relevance of the word in the i-th row with text in the j-th column. Setting a threshold value λ. When the weight is greater than λ, the keyword can be a feature seed.

The second step of our work is to build a bridge between the labels. Social media is a website and technology that allows people to write, share, evaluate, discuss and communicate with each other. It provides tools and platforms for people to share opinions, insights, experiences and perspectives among themselves. Social media can be considered as a tag cloud, in which the co-occurrence labels carry a wealth of information. In the paper, social media becomes our auxiliary means, which can help build the bridge between the source labels and the target labels. First, each label is viewed as a node and connects the co-occurrence labels. So all the labels in the social media would appear in the graph. Then extract the subgraph that contains all feature seed sets from it. This can create a bridge between the labels in the source domain and the labels inthe target domain. Finally, by improving the algorithm of Laplacian Eigenmapsmap, all the nodes in the subgraph can be mapped into a low-dimensional space. It effectively alleviates the problems such as data overfitting, low efficiency and so on, which are caused by the high-dimensional data. The next will improve the algorithm of Laplacian Eigenmapsmap so that it can comply with the requirements of our work.

The basic idea of Laplacian Eigenmapsmap is to represent the features of manifold local structure with Laplacian. High dimensional manifolds composed by the n points $x_1, x_2, \ldots, x_n$, were given mapping in the D-dimensional space:

$$f: R^D \rightarrow R^d \, (d << D) \qquad (2)$$

All the points are embedded in -dimensional space. Then in the d-dimensional space, we can get n points $y_1, y_2, \ldots, y_n$.

The Laplacian Eigenmapsmap algorithm [6] assumes that if the points are near in the high-dimensional space, the distances between them should be short when embedded into a low-dimensional space. As in the algorithm, it does not consider the classes information of the samples when calculate the neighbor distance. No matter the point inside the calss or outside the class, it gives the points same weight if the distances are the same. It is clearly unreasonable for target domain containing both labeled data and unlabeled data. In the paper, we improve the Laplacian Eigenmapsmap algorithm, using different methods to calculate the weight of labeled data and unlabeled data. And make points distance inside the class is less than the distance whose points are outside the class.

First construct the relative neighborhood graph. Use the method of unsupervised learning to calculate the distance between the unlabeled data. In the paper, we use Euclidean distance calculation method.

We use the method of supervised learning to calculate the distance between the labeled data. As follows:

$$D(x_i, x_j) = \begin{cases} \sqrt{1 - \exp(-d^2((x_i, x_j) / \beta} & c_i = c_j \\ \sqrt{\exp(-d^2((x_i, x_j) / \beta} & c_i \neq c_j \end{cases} \tag{3}$$

In the formula (3), c_i and c_j are classes of the samples x_i and x_j. $d(x_i, x_j)$is the Euclidean distance between x_i and x_j. Parameter β can prevent $D(x_i, x_j)$ too large when $d(x_i, x_j)$ become larger. It can effectively control the noises. If the distance between sample points x_i and x_j is smaller than the threshold e, the two points are neighbor points.

Then calculate the weight matrix W. If x_i and x_j are neighbor points, $W_{ij} = 1$, otherwise, $W_{ij} = 0$.

Finally calculate the Laplacian generalized eigenvectors and embed them into low-dimensional space. The problem is to solve that:

$$\begin{cases} min \sum_{i,j} || Y_i - Y_j || w_{ij} \\ s.t. \quad Y^T DY = 1 \end{cases} \tag{4}$$

In the formula (4) , D is a diagonal matrix. The formula (4) can be transformed into

$$\begin{cases} \arg min \ tr(Y^T LY) \\ s.t. \quad Y^T DY = 1 \end{cases} \tag{5}$$

In the formula(5), $L = D - W$.

With the improved Laplacian Eigenmaps algorithm, we canmget a matrix Y, each node can map to a low-dimensional space y. So each data can get a new feature representation.

2.3 Generate the Target Classifier

After getting the new feature representations of the target data, we can train a target classification with them. To complete the target task, we use the concept of mutual information.

Mutual information measures how much a random variable contains another one. It can reflect the relationship between two statistics. The higher the degree of association become, the greater the mutual information is. Mutual information is defined as follows:

$$I(P ; Q) = \sum_{x \in P} \sum_{y \in Q} p(x,y) \quad \frac{p(x,y)}{p(x) \ p(y)} \tag{6}$$

Due to the target domain are short texts, containing few characteristics and some may still be useless or disturbance character istics, in this paper we propose AutoTL algorithm which can transfer usful features from the long texts to enrich the target characteristics. The smaller of the mutual information between the new features representation and the label, the greater the degree of correlation of the data with the class. The likelihood of the data belonging to the class is greater. So the objective function is:

$$min\ I_j (y_i , c_j) \tag{7}$$

In the formula (7), y_i represents the new feature representation of data, c_j represents the j-th class.

Algorithm AutoTL

Input the target data $T = T^1 \cup T^u$ (T^1 are labeled data, and T^u are unlabeled data), the labels C (the number of the labels id N), neighbor value k , feature seed threshold λ , parameter β, feature threshold ε.

Output the classification results of the target data according to I.

$$I (P ; Q) = \sum_{x \in P} \sum_{y \in Q} p (x, y) \frac{p (x,y)}{p (x) \ p (y)}$$

Initialize k, λ, β, ε;

For $c = 1, . . . , N$

1. Input the target labels to a search engine.

2. Extract the first 10 pages of data as the data that the most associated with the target areas.

3. Get the feature seed sets according to $\tilde{M}$, k and λ.

$$\tilde{\mathrm{M}} = U \tilde{\Sigma} V^T \approx U \Sigma V^T = \mathrm{M}$$

4. Build the undirected graph of the social media.

5. Extract a subgraph that contains all these seed feature sets.

6. Filter the features of the target representation according to $tr\ (Y^T LY)$, β and ε

$$\begin{cases} \arg min\ tr\ (Y^T LY) \\ s.t. \quad Y^T DY = 1 \end{cases}$$

3. Experiments

3.1 Data set

In order to evaluate the efficiency of the AutoTL, we use 20Newsgroups, SRAA and Reuter-21578 as data sets. The 20Newsgroups includes 18774 news reports, which consists of 7 big class, 20 small class and 61188 vocabularies. SRAAincludes more than 700,00 UseNet articles, which consists of 2 big class and 4 small class. Reuter-21578 includes 22 files, which consists of 5 class.

First of all, extract all the labels in the target domain as keywords and input them into a search engine (In the paper, we use Google). Then select the most semantically related keywords with the source domain. Since these keywords are used as the feature seeds, it may not be good if they are too many, or it may produce too many useless items in the process of these seeds expansion. Not only can it affect the efficiency of the experiment, but also can affect the accuracy. In the paper, we select the first pages as the most semantically related with the source data. Then build the undirected graph of the social media(In the paper, we use Delicous website). Extract the subgraph according to the feature seed sets. So we can extract the most relevant characteristics from the short texts again, keeping the inherent structure of the source data. This is the second expansion of the data characteristics of the target domain.

3.2 Analysis of Experimental Results

In order to get a good effect, we firstly test some possible factors that may affect AutoTL algorithm. Then set appropriate values for them. Finally, compare our algorithm with other algorithms.

1) Parameter sensitivity test: In this paper, there are two very important parameters, λ and ε. When the number of the relevant pages is definite, the parameter of λ can decide the size of the seed feature sets. When we know the features seed sets, the parameter of ε can determine how many features we can filter from the social media. In this paper, we test the sensitivity of the two parameters using online advertising for the experimental data. First of all, fix the parameter of ε , we test the experimental performance when λ are set up for 0.3, 0.5, 0.7, 0.9. The experimental result was shown in Figure 1.

Secondly, fix the parameter of λ, we test the experimental performance when ε are set up for 0.3, 0.5, 0.7, 0.9. The experimental result was shown in Figure 2.

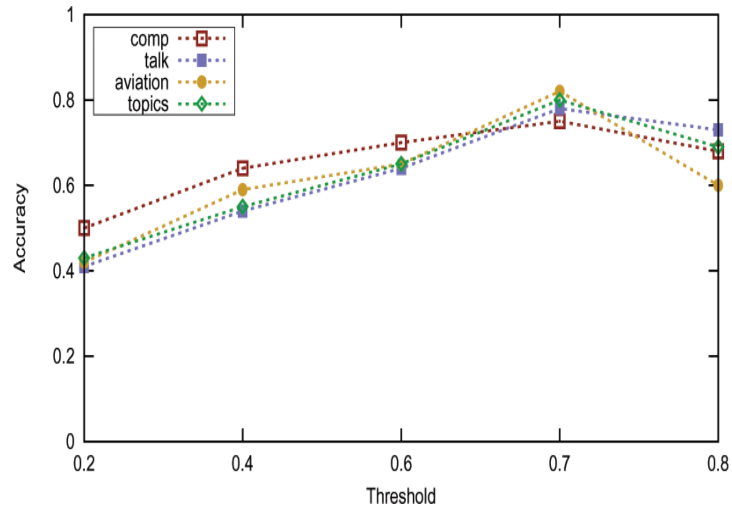


Figure 1. Impact of the mutual information threshold

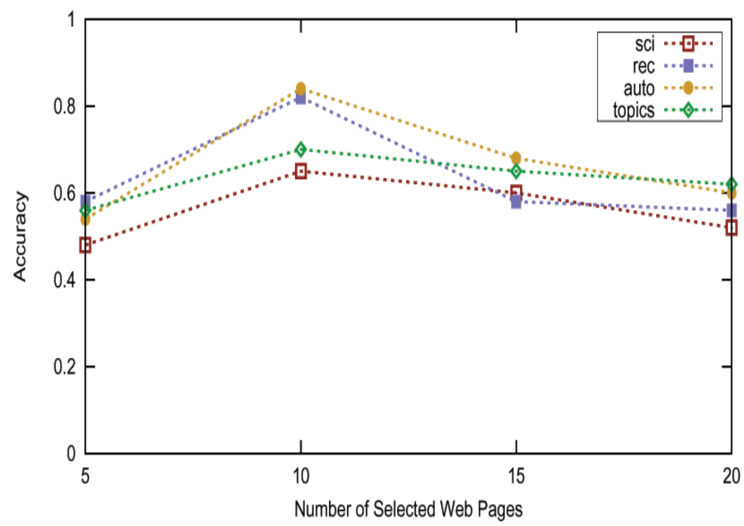


Figure 2. Impact of the number of web pages

The experimental results show that the value of λ and ε set too big or too small can both affect the accuracy of the experiment. The reason is that the value set too small will make the features associated with the target domain filtered out; However if they are set too big, some useless features would be mixed, which may exert an noise influence upon the final classification. From the experiment, we can see that when the value of λ is about 0.7, and ε is about 0.5, it can get a better experimental result. As λ decides the feature seed sets, it require a higher relevance for the keywords;ε determines the final characteristics of the target data, to be set up a small value is beneficial for filtering possible useful features.

2) The extraction of the relevant page in a search engine: As there is no need to provide the source data in this paper, all the auxiliary data come from the online information. In the traditional method of transfer learning, the quantity of the source data must be much bigger than the target data. In our algorithm, there are two steps to extract auxiliary data. The size of feature seed sets affects the characteristics extraction of the social media tags. The effect of the size of feature seed sets or the number of related pages in the search engine are shown in Figure 3.

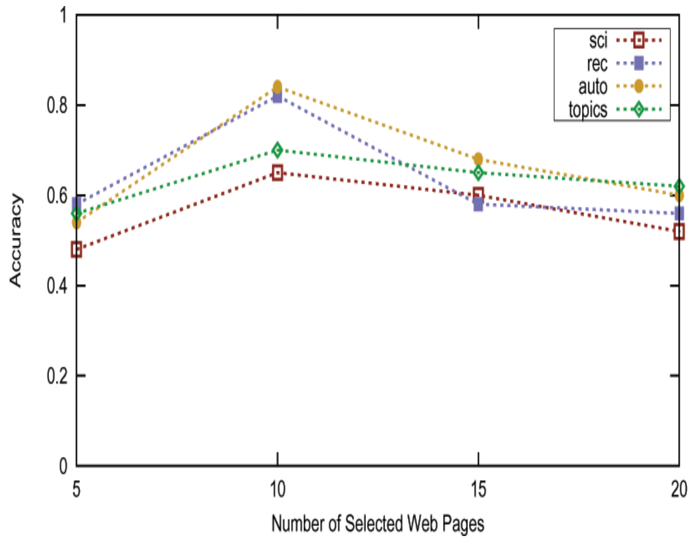


Figure 3. Impact of the number of web pages

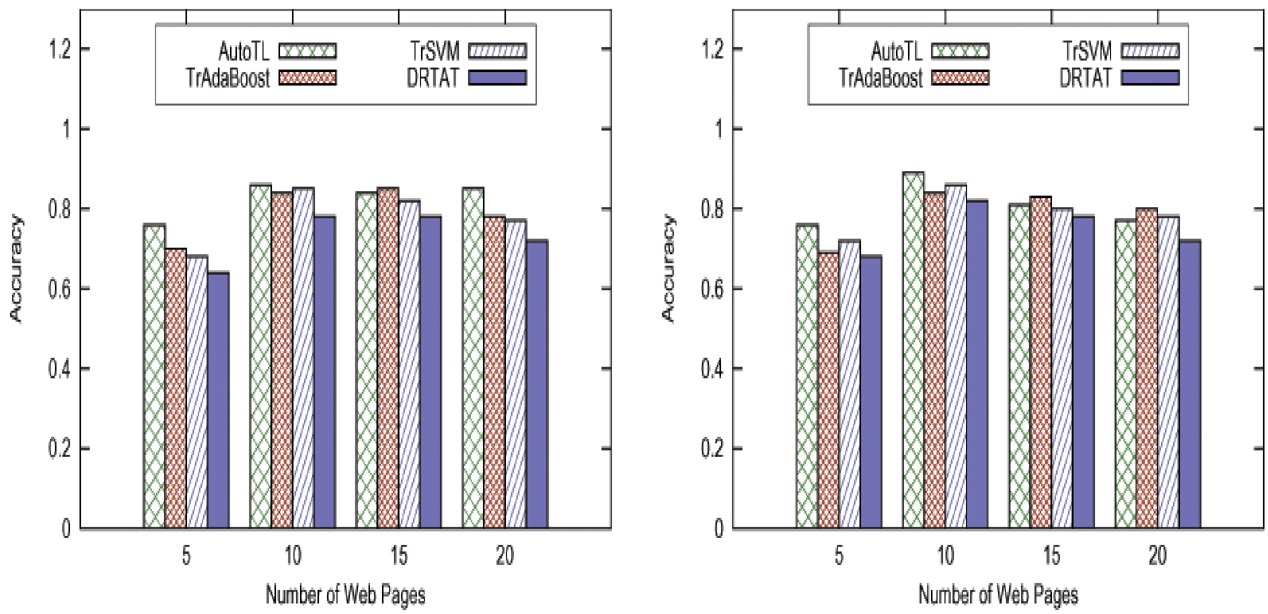
The experimental results show that the fewer the relevant pages in the search engine are, the fewer feature seeds it contains. Further, the features selected in the social media are fewer which will reduce the final classification accuracy. However, as it is shown in Figure 3, it is not the best to select many relevant pages in the search engine. Filtering many useless feature seeds will introduce more useless feature items, which can affect the accuracy and efficiency of the algorithm. Through several tests, to extract about the first 10 semantically related webs, it can get a better experimental result.

3. Compared Algorithms: In order to verify the effectiveness of the algorithm in the paper, we compared the AutoTL with SVM which was a classification algorithm without transfer learning, Tr _ SVM which was used to transfer knowledge from the long texts to the long, and DLDA that could transfer from the long texts to the short. In the algorithm of Tr _SVM, C and C_s were set 1.2. In the algorithm of DLDA, $\gamma^{aux}_{tar} = \gamma^{tar}_{aux} = \gamma_{small} = 0{:}2$; $\gamma^{tar}_{tar} = \gamma^{aux}_{aux} = \gamma_{big} = 0{:}5$. In the AutoTL, k was set 7, λ was set 0.75, ε was set 0.5, β was set 2.

This experiment was completed in the environment of linux, using the C language. In order to make the experiment not affected by random factors and more authentic, we did 10 times of the experiments and took the average of the experimental results. In the traditional transfer learning, when the target data was constant, the quantity of the data with labels can affect the algorithm. We compared the results of each algorithm by doing experiments on 4 data sets with different proportions of labeled data. Table.1 showed the accuracy changes of each algorithm when proportion of the labeled data increased from 5% to 15%. From the table, we could see the classification result of SVM without transfer learning is very poor. However, although TrSVM transferred knowledge from other domain, it ignored the differences between the long texts and the short, leading to not good result. The algorithms of DLDA and AutoTL, took into account the different data structures of the two fields, the quality selecting useful features was higher than other algorithms. Their accuracy could be up to 70% -80%. In addition, compared with DLDA, AutoTL did not to provide the source data and priori probability distribution of data before, so it reduced the influence caused by subjective factors. The feature seeds are the most semantically related to the target areas, and the expansion of the feature is the most relevant to the feature seeds. It greatly improved the characteristics' quality and increased the accuracy of the target classifier.

The next experiment, doing on the data of online advertising, studied the accuracies of different algorithms, when the target labeled data is constant but the target data gradually increased. Suppose the quantity of the target data was M in the above experiments. In this experiment, we would test the result when target data change from $0.1\times M$ to $1.5\times M$. The experimental result was shown in Figure 4.

From Figure 4, we can see that with the gradual increase of the target data, the accuracy of the algorithms increased. However, the classification accuracy of our algorithm proposed in the paper had always been better than the other algorithms. Especially when the target data were very few, the accuracy of AutoTL is much higher than other algorithms.



a) comp dataset (b) sci dataset

Figure 4. Performance comparison over comp and sci datasets, when the number of web pages changes

4. Related Work

In this section, we summarize our previous representative work on transfer learning. Dai et al. [7] proposed a Tradaboost algorithm, which improved the boosting technology in order to create an automatic weight adjustment mechanism. It can filter out most of the data similar to the target areas from the source field so that it can enrich the training data to improve the accuracy of the classifier; Mei et al. [8] proposed a WTLME algorithm. It bases on maximum entropy model, using instance weighted technology. The algorithm transfers model parameters studied from the original field to the target domain. It can reduce the time of re-collection and marking a large number of target data as well as the time of training model, achieving domain adaptation; Hong et al. [9] proposed a Tr_SVM algorithm. It requires weak similarity, which is defined by themselves, between two fields meeting certain constraints. Contact this constraint with the target classification, and embed the source data is into the support vector machine training. These transfer learning algorithms are based on instance, which can get a good result, but requires that the source data and the target data must be very similar. When the Source field and target area have a big difference, at the characteristic level, there are often some intersections which will become a bridge to connect two different domains, helping realize the knowledge transfer. Dai et al. [10] proposed a CoCC algorithm, in which the co-occurrence of words in the source field and the target domain were used as a bridge. The tag structures of the source field and the target domain were collaborative clustering at the same time. By minimizing the mutual information between words and samples, it can achieve the goal that transfer the tag structure of the source domain to the target domain; Xue et al. [11] proposed a TPLSA algorithm. As the source and target domain are related, there may be some common topics, which can be used as a bridge to connect the two areas. From the source domain, transfer the useful information, similar to the target subjects to the target field and help the target classification task; Long et al. [12] proposed a GTL algorithm, which think that the document is a collection of hidden topics and has its inherent geometry between keywords and the document. By extracting the potential common themes between source domain and target domain, optimizing maximum likelihood function, and maintaining the geometric structure of the documents as the same time, it can make the process of transfer learning smoother; The above algorithm is mainly used in the same language of the text fields. When the languages of the two fields were different, Ling et al. [13] found that although the data has a different text feature, but it may have a great relevance on semantic. Therefore they proposed an information bottleneck model. First of all, the Chinese web pages were translated into English. Then put these web pages and English web pages into the information bottleneck model for information coding. Finally filtered out the common information in the two areas and solved the problems of translation inaccuracy, label drift and so on. It made cross-language transfer learning possible.

All these algorithms above are transfer learning methods in long text domain. It may get good results when transfer knowledge from long text to the long one. But it is not applicable for short text as there are too many differences such as data structures, forms et al. between the two domains.

Most existing works focus on this kind of transfer learning that transfer knowledge from long text to the long. Although rarely, there also some people did the work that transfer methods or skills form the long text to the short. For example, Jin et al. [4] proposed a DLDA model. It extracted two sets of topics from the source and target domains and used a binary switch variable to control the forming process of the documents. As it allowed the two areas share the topic structures, transferring part related and useful data from the long texts to the target domain. However, in the algorithm, the source data were required to provide before, and the priori probability distribution of the data were also needed. Source data would affect the target task. It put a great burden on the designer to find the right data. How to filter the similar, relevant, and useful data, and how to accurately calculate the datas prior probability distribution became a big problem.

5.Conclusion

Transfer learning is a technique that finds useful knowledge and skills in the previous tasks and apply them to new tasks or domains. In this paper, we proposed AutoTL, an automatic transfer learning framework to analyze the short text data by utilizing the long text knowledge such as the web data. AutoTL shows its superiority than other algorithms from different perspectives. First, it does not enforce the user to provide a specific source data for training, but conducts an automatic source data selection. Second, no priori probability distribution is required in advance. Third, AutoTL integrates the rich online information in the short text learning task, which highly increases the learning accuracy. Extensive experimental evaluation indicated that AutoTL is practical, efficient and effective.

Acknowledgement

The authors would like to thank Dr. Wei Fengmei for her help and support in making this work possible. The work is funded by the Heilongjiang Scientific Research Foundation for Returned Scholars(Grant No.LC2015025), supported by Fundamental Research Funds for Central Universities(Grant No.HEUCFD1508 and HEUCF100602) and a special study of technological innovation fund of Harbin(Grant No.2013RFQXJ113 and 2013RFQXJ117). This work is also partially supported by China NSF(Grant No. 61402126, 61370083) and Postdoctoral Scientific Research Foundation of Heilongjiang Province.

References

[1] Pan, S. J., Qiang, Y. (2010) . A Survey on Transfer Learning. *IEEE Transactions on Knowledge and Data Engineering*. 22 (10) 1345–1359 .

[2] Sinno Jialin Pan and Qiang Yang (2008). An Introduction to Transfer Learning. Advanced Data Mining and Applications Conference, Hong Kong.

[3] Lu, Zhongqi., Zhu, Yin., Jialin Pan, Sinno (2014). Source Free Transfer Learning for Text Classification. *In*: Proceedings of the 28^{th} AAAI Conference on Artificial Intelligence, p.122–128. AAAI Press, Qu´ebec, Canada.

[4] Jin, Ou., Liu, Nathan, N. et al (2011). Transferring Topical Knowledge from Auxiliary Long Text for Short Text Clustering. *In*: Proceedings of the 20th ACM Conference on Information and Knowledge Management, p.775–784. ACM Press, Glasgow, UK.

[5] Dumais, S T., Furnas, G .W., et al: (1988). Using Latent Semantic Analysis to Improve Information Retrieval. *In*: Proceedings of the ACM Conference on Human Factors in Computing Systems, p.281-285. ACM Press, Washington D. C., USA.

[6] Belkin M., Niyogi P.: (2003). Laplacian eigenmaps for dimensionality reduction and data representation. *Neural Computation*. 10 (5) 1373-1396 .

[7] Dai, W., Yang, Q., Xue,G., et al. (2007). Boosting for transfer learning. *In*: Proceedings of The 24th Annual International Conference on Machine Learning, p.193–200. Corvallis, Oregon, USA.

[8] Canhua, Mei., Yuhong, Zhang., Xuegang, Hu., Li (2011). Peipei: Transfer learning algorithms based on maximum entropy model. *Computer Research and Development,* 48 (9) 1722-1728.

[9] Jiaming, Hong., Jian, Yin., Yun, Huang., Yubao, Yiu., Jiahai, Wang (2011). TrSVM: A transfer learning method based on the similarity of domains. *Computer Research and Development*. 48 (10) 1823-1830.

[10] Dai, Wenyuan., Xue, Gui-Rong et al: (2007). Co-clustering based Classification for Out-of-domain Documents. *In*: Proceedings of the Thirteenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, p.210–219. San Jose, California, USA.

[11] Xue, Guirong., Dai, Wenyuanet al: (2008). Topic-bridged PLSA for Cross- Domain Text Classification. *In*: Proceedings of the 31st Annual International ACM SIGIR Conference, p.627–634. ACM Press, Singapore.

[12] Long, Mingsheng., Wang, Jianmin., Ding, Guiguang., Shen, Dou., Qiang Yang (2012). Transfer Learning with Graph Co-Regularization. *In:* Proceedings of the 26th AAAI Conference on Artificial Intelligence, AAAI Press, Toronto, Ontario, Canada.

[13] Ling, Xiao., Dai, Wenyuan et al (2008). Can Chinese Web Pages be Classified with English Data Source. *In*: Proceedings of the 17th International Conference on World Wide Web 2008, p.969–978. Beijing, China.