

The effectiveness of using methods two-stage for cross-domain sentiment classification

Hoan Manh Dau^{*}, Ning Xu

School of Computer Science, Wuhan University of Technology, China

Received 14 May 2014, www.tsi.lv

Abstract

Traditional sentiment classification approaches perform well in sentiment classification but traditional sentiment classification approaches does not perform well with learning across different domains. Therefore, it is necessary to build a system which integrates the sentiment orientations of the documents for every domain. However, this needs much labelled data involving and much human labour as well as time consuming. Thus, the best solution is using labelled data in one existed in source domain for sentiment classification in target domain. In this paper, a two-stage approach for cross-domain sentiment classification is presented. The First Stage is building a bridge between the source domain and the target. The Second Stage is following the structure. The study shows that the mining of intrinsic structure of the target domain brings a considerable effectiveness during the process of sentiment transfer. This is a typical mining approach comparing to previous approaches basing on information from the source domain to address the task of sentiment transfer, which does not depend on intrinsic structure of the target domain. Experimental results on sentiment classification with a two-stage approach indicate that the effectiveness outperforms other traditional methods.

Keywords: cross-domain, sentiment classification, sentiment transfer, opinion mining

1 Introduction

Sentiment classification is attracting more and more people's attention because of its great benefits to social and human life. Automatic sentiment classification aims to predict automatically sentiment polarity (e.g., positive or negative) of users who publish sentiment data. Sentiment classification is the area of sentiment analysis can help human analyse, synthesize, organize, summarize and forecast for determining the sentiment orientation of subjective text. This is an important sub-task of sentiment analysis. It plays an important role in numerous applications like opinion mining, market analysis and opinion summarization. Today, when internet services bloom as mushrooms with many social networking sites, handsets can connect to the network and many people create sentiment data to share on the Web. Users express and share their opinions about many topics on Websites and blogs. Researching of sentiment classification has contributed to text classification research and therefore it has an important significance for those who want to forecast information from text document data. Researchers have pointed out that sentiment classification has been applied effectively, such as [1-4].

In most cases, supervised learning methods for sentiment classification have been studied popularly and applied rather successfully. Researches have showed that standard machine learning techniques definitively outperform human-produced baselines. However, a disadvantage of sentiment is expressed differently in different domains, and it is the requirement for labelled in

domain data for training. Supervised learning methods for sentiment classification require two conditions to ensure the accuracy in classification. The first condition is that training data is sufficient and labelled well; and the second is that training data and test data should have the same distribution. However, in reality these two conditions cannot be met. The main reason is that labelling data involves much human labour and it is time-consuming; apart from that the labelled and unlabelled data are often from different domains, and often have different distributions. Therefore, the problem is how to use labelled sentiment data in source domain for sentiment classification in target domain. This is the main task of cross-domain sentiment classification. When performing cross-domain sentiment classification (or sentiment transfer), many researchers gave some techniques to improve them in order to make the process of sentiment transfer more effective such as [5, 6]. However, the two problems of sentiment transfer are that the distribution of the target domain is not same with that of the source domain and the intrinsic structure of the target domain is static. To solve these two problems, a bridge needs building to share information between source domain and target domain and the intrinsic structure needs used to carry out for target domain. Selecting technique to build a bridge between the source and the target domain will impact on the effectiveness of the process of sentiment transfer. Transfer learning aims to use data from other domains to help current learning task. Transfer learning plays important role in research field in machine learning. There were some typical

^{*} *Corresponding author* e-mail: daumanhhoan@yahoo.com

researches such as: introducing a statistical formulation to domain adaptation in terms of a simple mixture model [7] introducing a two-stage approach to domain adaptation for statistical classifiers [8] proposing a bridged algorithm, which takes the mixture distribution of the training and test data as a bridge to better transfer from the training data to the test data [9] presenting an adapting naive Bayes to domain adaptation for sentiment analysis [10]. However, some researchers based only on the labelled documents to improve the performance of sentiment transfer [11, 12]. Most of these researchers used information from the source domain to address the task of sentiment transfer but ignored the intrinsic structure of the target domain.

In this paper, technique for transfer learning in the context of sentiment classification is presented. The effectiveness of applying the SentiRank algorithm in the process of sentiment transfer and mining intrinsic structure of the target domain which brings feasible effectiveness are shown. The testing results are presented and showed that the effectiveness of this approach when making sentiment transfer is considerable.

2 The proposed approach

2.1 PROBLEM DEFINITION

There are two document sets in this paper: D^U denotes the test data, and D^L denotes the training data. Assign every document a sentiment score ("1" denotes positive, and "-1" denotes negative) to represent its degree of sentiment orientation and call it sentiment score. S^U denotes the sentiment score set of D^U , and S^L denotes the sentiment score set of D^L . It is assumed that the training dataset D^L is from the related but different domain with the test dataset D^U . The aim is to maximize the accuracy of assigning a label in D^U utilizing the training data D^L in another domain.

2.2 OVERVIEW

A two-stage approach for sentiment transfer is given.

The implementations of this approach are shown in [6]. The process consists of two stages (two-stage):

- the first stage: building a bridge;
- the second stage: following the structure.

In the first stage, there are 3 steps:

- the first step is to use SentiRank algorithm to get the sentiment scores of the target domain documents.
- the second step is to get Initial Sentiment Scores of the Target Domain Data.
- the third step is to choose Seed Set of confidently labelled documents as high-quality.

In the second stage, there are two steps:

- the first step is to apply a manifold-ranking algorithm to follow the structure of the target domain.
- the second step is to use the manifold-ranking scores to label the target-domain data.

2.3 THE FIRST STAGE: BUILDING A BRIDGE

In this stage, firstly the SentiRank algorithm is used to build a bridge between the source domain and the target domain. In order to get the labels of the target domain documents, the information of the source domain is used. The SentiRank method [13] is an algorithm for sentiment transfer and it is used to get the sentiment orientations of the target-domain documents using the similarity between the documents from both the source domain and the target domain. The implementation of the algorithm is that if one document has a strong relationship with positive documents or negative documents, it can probably be positive or negative. The implementation of SentiRank is described as follow: A weighted graph is built from the data, and a sentiment score is assigned for every labelled and unlabelled document to denote its extent to "negative" or "positive", then the score is iteratively calculated making use of the accurate labels of source domain data as well as the "pseudo" labels of target domain data via the weighted graph. The final score for sentiment classification is achieved when the algorithm is converged, so the target domain data can be labelled based on these scores. The SentiRank process is described in details in [6].

In this algorithm, α and β show the relative importance of source domain and target domain to the final sentiment scores, and $\alpha + \beta = 1$. Algorithm achieves the convergence when the changing between the sentiment scores computed at two successive iterations for all documents in the target domain falls below a given threshold. Secondly, in order to find high quality documents from the target domain, sentiment score needs creating and using to denote the "negative" or "positive" correlation of documents. Next, the target domain documents is sorted in descending order according to their sentiment scores. So the more forward the document is sorted, the more likely it is positive; the more backward the document is sorted, the more likely it is negative. Then, the first K documents and last K documents as the high quality documents are chosen. Thirdly, Seed Set of confidently labelled documents as high quality is chosen. This algorithm proves that results produce high quality seeds. Sorting the target domain documents according to their opinion extent is effective and the proof is shown in Table 1.

TABLE 1 Seed accuracies on six tasks [6]

Do main	K						
	50	90	130	170	210	250	290
B→H	0.9500	0.9222	0.9230	0.9294	0.9333	0.9340	0.9240
B→N	0.8200	0.8778	0.8923	0.8912	0.8905	0.8820	0.8860
H→B	0.8000	0.8055	0.8115	0.8117	0.8024	0.7540	0.7431
H→N	0.9300	0.9277	0.9230	0.9235	0.9214	0.9100	0.9086
N→B	0.7400	0.7500	0.7461	0.7264	0.7142	0.7120	0.6810
N→H	0.9167	0.9111	0.9000	0.8976	0.8990	0.8980	0.8972

TABLE 2 Accuracy comparison of different methods [6]

Domain	Proto	TSVM	SentiRank	EM based on Proto	Manifold based on Proto	Main Approach
B→H	0.735	0.749	0.772	0.765	0.761	0.790
B→N	0.651	0.769	0.714	0.667	0.745	0.776
H→B	0.645	0.614	0.671	0.723	0.677	0.683
H→N	0.729	0.726	0.749	0.657	0.784	0.784
N→B	0.612	0.622	0.638	0.763	0.665	0.650
N→H	0.724	0.772	0.764	0.765	0.779	0.791
Average	0.683	0.709	0.718	0.723	0.735	0.746

Table 1 shows that the effectiveness of the algorithm carried out from domains $B \rightarrow H$, $H \rightarrow N$ and $N \rightarrow H$ has the accuracy of above 89%, and the effectiveness from domains $B \rightarrow N$ and $H \rightarrow B$ has the accuracy of above 75%. This high accuracy demonstrates that the effectiveness of algorithm is enough to choose high-quality seeds. In the case of transfer from domain $N \rightarrow B$, the accuracy is not particularly good. The main reason is due to the too big difference between the two domains notebook (N) and book reviews (B). However, this shortcoming can be overcome and can improve the performance of sentiment transfer exploiting these seeds.

2.4 THE SECOND STAGE: FOLLOWING THE STRUCTURE

In this stage, although the algorithm can build a bridge between the source domain and the target domain, the distribution of the target domain is not used but the intrinsic structure of the target domain is used for sentiment transfer. It starts with a small amount of high quality seed set, this is the number of seeds representing for intrinsic structure of the target domain. Manifold-ranking method is used to make better use of the seeds, and it can improve the performance of sentiment transfer.

The manifold-ranking method [14] is a universal ranking algorithm and it is initially used to rank data points along their underlying manifold structure. The prior assumption of manifold-ranking is nearby points which are likely to have the same ranking scores and points on the same structure are likely to have the same ranking scores. The implementation of the algorithm is as follow: a weighted network is formed on the data, and a positive rank score is assigned to each known relevant point and zero to the remaining points which are to be ranked. All points then spread their ranking score to their nearby neighbours via the weighted network. The spread process is repeated until a global stable state is achieved, and all points obtain their final ranking scores [6].

With a high quality seed set, first, the weighted network whose points denote documents in D^U is built.

And then integration the sentiment scores of the seeds into the manifold-ranking process is carried out. Then the sentiment manifold-ranking process is implemented. Finally, label the documents in target domain according to their ranking score vector. Each document is labelled with positive or negative labels.

3 Experiments

3.1. BASELINE SYSTEMS

In this part, testing results of chosen method is shown and compared to the results of other baseline methods.

Table 2 shows that accuracy comparison of different methods [6]:

- Method Proto: the results from column 2 show that the accuracy ranges from 61.25% to 73.5%. It is result of method which applies a traditional supervised classifier, prototype classifier for the sentiment transfer [15]. This technique only uses source domain documents as training data.

- Method Transductive Support Vector Machine (TSMV): the results from column 3 show that the accuracy ranges from 61.42% to 77.17%, which is better than that of method Proto. This method applies transductive SVM for the sentiment transfer [16]. This is a widely used method for improving the classification accuracy. This method uses both source domain data and target domain data.

- Method SentiRank: column 4 shows the results that the accuracy ranges from 63.7% to 77.2%, which is much better than method Proto and TSVM. The implementation of this method is to run SentiRank algorithm at places initializing the sentiment scores by prototype classifier.

- Method Expectation Maximization (EM) based on Proto: the column 5 shows that results of the method of EM algorithm [17] based on prototype classifier is similar to the above apart from changing the training classifier from SentiRank to prototype classifier, and its results are accuracy ranges from 65.7% to 76.5%, better than the first three baselines.

- Method Manifold based on Proto [14]: the column 6 shows that the accuracy ranges from 66.5% to 78.4%, which is better than all other baselines. This method begins by training a prototype classifier on the training data, then by use the similarity scores between the documents and the positive central vector and the similarity scores between the documents and the negative central vector to separately initial the ranking score vectors of the test data. Finally, it is carried out to choose KM documents that are most likely to be positive and KM documents that are most likely to be negative as seeds for manifold-ranking.

3.2 THE MAIN APPROACH

The proposed approach is compared with 5 baseline methods. The column 7 in Table 2 shows the mentioned approach [6]. The approach recommended in this paper is better performed than all the method baselines. Table 2 shows that greatest increase of accuracy is achieved by about 12.7%, when implementing $H \rightarrow N$ compared to method EM based on Proto. The second greatest increases of accuracy is achieved by about 12.5%, when performing $B \rightarrow N$ and the third greatest increases of accuracy is achieved by about 6.7%, when implementing $N \rightarrow H$ compared to method Proto respectively. The greatest average increase of accuracy is achieved by about 6.3% compared to method Proto. The experiment results show that this method can dramatically improve the accuracy when transferred to a new domain. The results in Table 2 also show that the average accuracies of method SentiRank and TSVM are higher than method Proto. The problem is that method SentiRank and TSVM use information of both source domain and target domain while method Proto not. This proves that using the

information of two domains is better than using the information of only one domain for improving the accuracy of sentiment transfer. In addition, it is clear that the average accuracies of three last methods are higher than that of the three first methods. Three last methods use two-stage approaches, while three first methods do not, which proves that two-stage transfer approach is more effective for sentiment transfer. The above results indicate that the approach which is recommended in this paper has feasible effectiveness.

4 Conclusions



In this paper, the effectiveness of implementing two-stage approach for sentiment transfer is presented. The effectiveness of this approach is proved by comparing its testing results to other basic approaches' results. In order to carry out this approach, a bridge between the source domain and the target domain is built and then the intrinsic structure of the target domain to improve the performance of sentiment transfer is used. The typical characteristic of this approach is using the "pseudo" labels technique to create sentiment scores of the target-domain documents by applying the SentiRank algorithm, then using sentiment scores to identify the best domains with labelled documents as high-quality seeds, in the meanwhile using manifold-ranking algorithm for ranking score for every unlabelled document, finally implementing label the target-domain data based on these scores. Testing results on data prove that this approach improves the accuracy, and can be employed as a high-performance sentiment transfer system. Exploiting good points and advantages and extending this approach for other text classification tasks are potential for further research.

References

- [1] Pang B, Lee L, Vaithyanathan S 2002 Thumbs up? sentiment classification using machine learning techniques *Proceedings of EMNLP* 79–86
- [2] Pang B, Lee L 2004 A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts *Proceedings of ACL* 217–78
- [3] Pang B, Lee L 2008 Opinion mining and sentiment analysis *Foundations and Trends in Information Retrieval* 2(1-2) 1–135
- [4] Hu M, Liu B 2004 Mining and summarizing customer reviews *Proceedings of the 10th ACM SIGKDD international conference on knowledge discovery and data mining* Seattle WA USA ACM 168–77
- [5] Liu K, Zhao J 2009 Cross-Domain Sentiment Classification Using a Two-Stage Method *Proceedings of the 18th ACM conference on information and knowledge management* 1717–20
- [6] Qiong Wu, Songbo Tan 2011 A two-stage framework for cross-domain sentiment classification *Proceedings Expert Systems with Applications* 38(11) 14269–75
- [7] Daumé III H, Marcu D 2006 Domain adaptation for statistical classifiers *Journal of Artificial Intelligence Research* 26 101–26
- [8] Jiang J, Zhai C 2007 A Two-Stage Approach to domain adaptation for statistical classifiers *Proceedings of the 16th ACM conference on Conference on information and knowledge management* Pages 401–10
- [9] Xing D, Dai W, Xue G, Yu Y 2007 Bridged refinement for transfer learning *Proceedings of the 11th European Conference on Practice of Knowledge Discovery in Databases (PKDD)* Springer 324–35
- [10] Tan S, Cheng X, Wang Y, Xu H 2009 Adapting naive Bayes to domain adaptation for sentiment analysis *Proceedings of the 31st European Conference on IR Research (ECIR)* Toulouse France April 2009 337–49
- [11] Tan S, Wang Y, Wu G, Cheng X 2008 Using unlabelled data to handle domain-transfer problem of semantic detection *Proceedings of the 2008 ACM symposium on Applied computing* 896–903
- [12] Dasgupta S, Ng, V 2009 Mine the easy, classify the hard: A semi-supervised approach to automatic sentiment classification *ACL-IJCNLP 2009 Proceedings of the Main Conference* 701–9
- [13] Wu Q, Tan S, Cheng X 2009 Graph ranking for sentiment transfer *Proceedings of the ACL-IJCNLP 2009 Conference Short Papers* 317–320
- [14] Zhou D, Weston J, Gretton A, Bousquet O, Schölkopf B 2003 *Ranking on data manifolds* Advances in neural information processing systems (NIPS) 16 169–76
- [15] Han E, Karypis G 2000 Centroid-based document classification: Analysis & experimental results *Proceedings of the 4th European Conference on Principles of Data Mining and Knowledge Discovery* 424–31

[16]Joachims T 1999 Transductive inference for text classification using support vector machines *Proceedings of the Sixteenth International Conference on Machine Learning* 200–9

[17]Dempster A P, Laird N M, Rubin D B 1977 Maximum likelihood from incomplete data via the EM algorithm *Journal of the Royal Statistical Society* 39(B) 1–38

Authors	
	<p>Hoan Manh Dau, born in June, 1976, Vietnam</p> <p>Current position, grades: Ph.D student at computer field at School of Computer Science from 2011, Wuhan University of Technology, China. University studies: M.A degree at Informatics at Hue University of Sciences in Vietnam in 2004. Experience: lecturer of informatics for 12 years at University.</p>
	<p>Ning Xu, China</p> <p>Current position, grades: Professor at the Computer Science Department of Wuhan University of Technology, senior member of China Computer Federation and the Chinese Institute of Electronics. University studies: Ph. D. degree in electronic science and technology at the University of Electronic Science and Technology of China in 2003. Scientific interest: Computer-aided design of VLSI circuits and systems, computer architectures, data mining, highly combinatorial optimization algorithms. Publications: Over 50 research papers. Experience: 5 research projects.</p>