Available Online at www.ijcsmc.com

International Journal of Computer Science and Mobile Computing

A Monthly Journal of Computer Science and Information Technology

ISSN 2320–088X

IJCSMC, Vol. 4, Issue. 8, August 2015, pg.209 – 215

RESEARCH ARTICLE

Improved Knowledge Discovery And Vocabulary Gap Filling Using Genetic Algorithm

K.Bhuvaneswari[1], M.Anusha[2], Dr. J.G.R.Sathiaseelen[3]

[1]Department of Computer Science, Bishop Heber College, Trichy, India
[2]Department of Computer Science, Bishop Heber College, Trichy, India
[3]Department of Computer Science, Bishop Heber College, Trichy, India
[1] bhuvanakrishnan.174@gmail.com; [2] anusha260505@gmail.com; [3] jgrsathiaseelan@gmail.com

Abstract— A major problem of classification learning is the lack of ground-truth labeled data. It is usually expensive to label new data instances for training a model. To solve this problem, domain adaptation in transferring learning has been proposed to classify target domain data by using some other source domain data, even when the data may have different distributions. However, domain adaptation may not work well when the differences between the source and target domains are large. In this paper, we design a novel transfer learning approach, called BIG (Bridging Information Gap), to effectively extract useful knowledge in a worldwide knowledge base, which is then used to link the source and target domains for improving the classification performance. BIG works when the source and target domains share the same feature space, but different underlying data distributions. A major contribution of this work with BIG, a large amount of worldwide knowledge can be easily adapted and used for learning in the target domain.

Keywords— domain adaptation, BIG algorithm, Local Mining.

I. INTRODUCTION

The incredible increase in the amount of information on the World Wide Web has caused the beginning of topic specific crawling of the Web. During a focused crawling process, an automatic Web page classification mechanism is needed to determine whether the page being considered is on the topic or not. Text classification, which aims to assign a document to one or more categories based on its content. It is a fundamental task for Web and document data mining applications, ranging from information retrieval, spam detection, to online advertisement and Web search. Traditional supervised learning approaches for text classification require sufficient labeled instances in a problem domain in order to train a high quality model. However, it is not always easy or feasible to obtain new labeled data in a domain of interest. The lack of labeled data problem can seriously damage classification performance in many real world applications [1]. To solve this problem, transfer learning techniques, in particular domain adaptation techniques in transferring learning is introduced by capturing the shared knowledge from some related domains where labeled data are available, and use the knowledge to improve the performance of data mining tasks in a target domain. In transfer learning terminologies, one or more auxiliary domain is identified as the source of knowledge transfer, and the domain of interest is known as the target domain. Much effort has been dedicated to this problem in recent years in machine learning; data mining, and information retrieval [2]. However, transfer learning may not work well when the difference between the source and target domains is large. In particular, when the distribution gap between the source and target domains is large, transfer learning can hardly be used to benefit learning in the target domain.

Domain adaptation addresses a common situation that arises when applying machine learning to diverse data. Sample data drawn from a source domain to train a model, but little or no training data from the target domain where we wish to use the model .Domain adaptation has attracted more and more attention in the recent years. In general, previous domain adaptation approaches can be classified into two categories: instance based approaches or feature-based approaches [3]. Instance-based methods try to seek some rewiring strategies on the source data, such that the source distribution can match the target distribution. Feature-based methods try to discover a shared feature space on which the distributions of different domains are pulled closer. Both types are trying to discover the relation between source and target domains within the scope of two domains. For the feature-based domain adaptation models, they assume that different domains may share some features, for instance, a subset of explicit features or implicit features. Here, consider some well-known instance-based domain adaptation methods.

The instance weighting method for natural language processing, where the method used is a type of importance sampling method for solving sample selection bias problems. Propose a boosting-style reweighing method and provide different weighting schemes for data in different domains. Other feature-based methods have been developed and are compared for instance-based methods. A major component of this approach is to use online knowledge repositories as auxiliary information sources to help bridge the gap between the source domain and the target domain. Therefore, we review some latest approaches of data mining with online knowledge repositories. In recent years, understanding and using online knowledge repositories to aid real world data mining tasks has become a hot research topic. There are more and more works trying to use the Wikipedia for further enrichment. The Open Directory Paper (ODP) for further enrichment in the text, they also show that using Wikipedia as the external Web knowledge resource for further enrichment performs better than using ODP. Their semantic analysis is explicit in the sense that they manipulate manifest concepts grounded in human cognition, rather than latent concepts used by LSA. A general framework for building classifiers with hidden topics discovered from large-scale data collections is proposed [4]. The framework is mainly based on latent topic analysis models like PLSA and LDA and machine learning methods like maximum entropy and SVMs. The underlying idea of such a framework is that for each classification task, a very large external data collection is collected and called a "universal data set," then a classification model on both a small set of labeled training data and a rich set of hidden topics discovered from that data collection is built. Currently, few previous approaches are using auxiliary knowledge, such as an online knowledge database for transfer learning or domain adaptation. Kannuri Lahr et al.[5] make an extension to the feature-based transfer learning models via incorporating a semantic kernel learned from Wikipedia. However, building a semantic kernel from the whole knowledge base is costly.

The remainder structures are followed in below sections. Section II briefly reviews the related work. The proposed work is respectively introduced in Section III. Section IV details the experimental results and analysis, followed by our concluding remarks in Section V.

II. RELATED WORK

Vivekanandan et al.[6] proposed a genetic algorithm which tries to gain maximum knowledge in between the generations and store them in the form of knowledge chromosomes. The gained knowledge is used to make predictions about the search space and to guide the search process in an area with potential solutions in the subsequent generations. This makes the genetic algorithm to converge quickly, which in turn reduces the learning cost. The experiments show that the run time is reduced considerably when compared with the state-of-the-art evolutionary algorithm. Liqiang et al. [7] presents a novel scheme to code the medical records by jointly utilizing local mining and global learning approaches, which are tightly linked and mutually reinforced. Local mining attempts to code the individual medical record by independently extracting the medical concepts from the medical record itself and then mapping them to authenticate terminologies. Global learning, on the other hand, works towards enhancing the local medical coding via collaboratively discovering missing key terminologies and keeping off the irrelevant terminologies by analyzing the social neighbors. More importantly, the whole process of this approach is unsupervised and holds potential to handle large-scale data.

Patrick et al. [8] implemented the system as a web-service system which consists of 3 modules such as an Augmented Lexicon, term compositor and negation detector. The algorithm uses an augmented lexicon to index concept describe in SNOMED CT, which allow a much faster mapping of longest concepts in system than naïve searching approach. A qualification identifier, and negation identifier have been implemented for recognizing composite terms and negative concepts, which can create a more effective information retrieval and information extraction. It is currently used in a hospital environment to capture patient data response with SNOMED-CT codes in real time at the point of care. The system is yet to be fully evaluated, nevertheless the test on sample data shows it is already meeting expectations.

Edwin et al. [9] suggested a new approach for solving the searching space to find the optimal distribution of object in the clusters represents a hard combinatorial problem. Genetic algorithms increased the intra and inter clusters entropy simultaneously and yield the best possible combination of elements for a present number of clusters.Xinmei et al.[10] explicitly formulates the problem in the Bayesian framework, i.e., maximizing the ranking score consistency between visually similar video shots while minimizing the ranking distance, which represents the disagreement between the objective ranking list and the initial text based. Two new methods are proposed in this paper to measure the ranking distance based on the disagreement in terms of pair-wise orders. By incorporating the proposed distances into the optimization objective, two reranking methods are developed which are solved using quadratic programming and matrix computation respectively. Daniel et al. [11] attempted to improve the coding performance by combing the advantages of rule-based and machine learning approaches. It described Auto coder, an automatic encoding system implemented at Mayo clinic. Autocoder combines example based rules and a machine learning module using Naıve Bayes. However, this integration is loosely coupled and the learning model cannot incorporate heterogeneous cues, which is not a good choice for the community-based health services.

Saman et al. [12] introduced a cascade of two classifiers to assign diagnostic terminologies to radiology reports. In their model, when the first classifier made a known error, the output of the second classifier was used instead to give the final prediction. Bhuvaneswari et al.[13] presented an extensive survey on the performance of the clustering techniques based on the genetic algorithm. Yan et al. [14] proposed a multi-level large-margin formulation that explicitly incorporated the inter-terminology structure and prior domain knowledge simultaneously. This approach is feasible for a small terminology, set but is questionable in real-life settings where thousands of terminologies need to be considered.

III. PROPOSED WORK

The proposed system presents a novel scheme that is able to code the records with corpus-aware terminologies. The proposed scheme consists of two mutually reinforced components, namely, local mining and global learning. Local mining aims to locally code the records by extracting the concepts from individual record and then mapping them to terminologies based on the external authenticated vocabularies. These systems establish a tri-stage framework to accomplish this task, which includes noun phrase extraction, domain adaptation and concept normalization. As a byproduct, a corpus-aware terminology is naturally constructed, which can be used as terminology space for further learning in the second component. However, the local mining approach may suffer from the problem of information loss and low precision due to the possible lack of some key concepts in the records and the presence of some irrelevant concepts. The global learning complements the local coding in a graph-based approach. It collaboratively learns missing key concepts and propagates precise terminologies among underlying connected records over a large collection. Besides the semantic similarity among medical records and terminology-sharing network, the inter-terminology and inter-expert relationships are seamlessly integrated in the proposed model. The inter-terminology relationships are mined by exploiting the external well-structured ontology, which are able to alleviate the granularity mismatch problems and reduce the irrelevant sibling terminologies. The inter expert relationships are inferred from the experts' historical data. It may be capable of excluding a wealth of domain-specific context information.

The figure 1 illustrates the architecture of bridge domain. To solve this problem, introduce a bridge between the two different domains by leveraging additional knowledge sources that are readily available and have wide coverage in scope. And then apply to semi supervised learning (SSL) to domain adaptation problems based on the use of the auxiliary data. The labeled data taken from the source domain and the unlabeled data taken from the target domain, as well as an auxiliary data taken from sources such as the Wikipedia. After selecting the data from the various domains apply SSL to utilize the information contained in the unlabeled data to help in the classification task on the target data. Although domain adaptation (DA) and SSL share similar problem settings, which is directly using SSL to solve DA problems nevertheless the results would produce poor performance. However, the existence of the information gap can make the assumption invalid. Using this approach, the extracted bridge for filling in the information gap, SSL-based algorithms can be applied successfully to the classification problems.

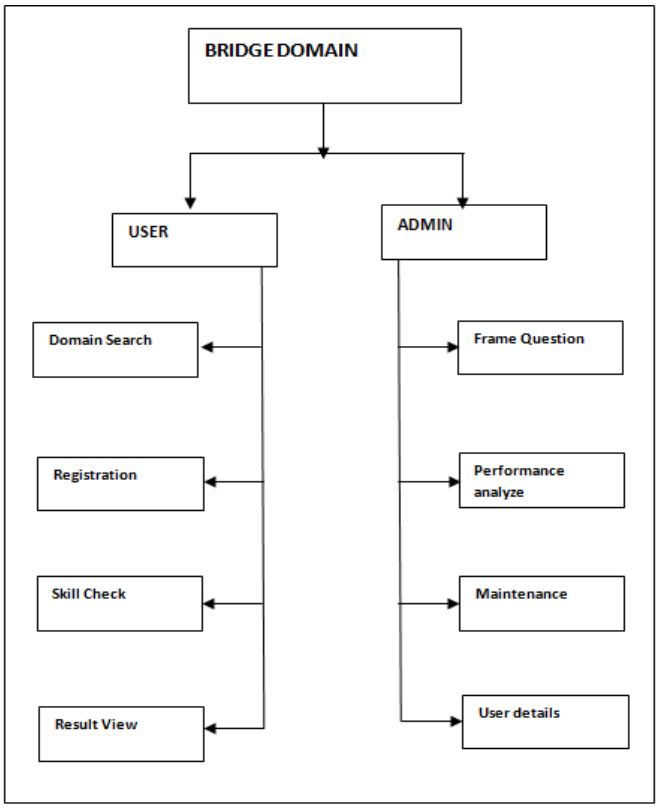

Fig. 1 Architecture of bridge domain

This paper also introduces a novel domain adaptation algorithm called BIG (Bridging Information Gap). Our BIG algorithm requires that the source domain and the target domain share the same feature space, but the distributions between domains can be highly different. A major contribution of this work is that use of a large amount of worldwide knowledge to build a bridge for linking the source and target domains, even when their distribution differences are large. It reduces the information gap between source domain and target domain. This is also worked in a large kind of domains. It shows that we can obtain useful information to bridge the source and the target domains from auxiliary data sources. Minmargin algorithm can effectively identify and reduce the information gap between two domains. An intuitive way to understand the concept of information gap is to consider separability of the source and target domains. Consider the simplest case when we want to transfer knowledge from a single source domain to a target domain intuitively, the difficulty in separating these domains shows how large the information gap is between them. If the two domains can be easily separated, then there exists a large information gap between them, which may prevent our adapting the original learned model from the source to the target domain. On the contrary, if the two domains cannot be separated from each other easily, then the information gap is small, in which case we can treat the two domains as essential data that are sampled from a single underlying distribution

In other words, the original "domain adaptation problem" is transformed into a classification problem under the supervised setting or a semi supervised (transductive) setting. A similar idea is used where a classifier is trained to distinguish the source and target domains and the classification error is used as an empirical estimation for domain distance. Although this idea is useful, it did not consider the existence of auxiliary information sources that can be used to bridge two domains.This document is a template. An electronic copy can be downloaded from the conference website. For questions on paper guidelines, please contact the conference publications committee as indicated on the conference website. Information about final paper submission is available from the conference website.

BIG ALGORITHM INTEGRATE A GREEDY ALGORITHM FOR MIN-MARGIN DOMAIN ADAPTATION

```
Input
        Source Domain Dataset D^s,
        Target Domain Dataset D^t,
        Knowledge base K,
        Terminating threshold t.
Output
        A subset D from K
Initialize
        Set D=0
Preprocess
        Set D0=Select(D^s,D^t,K,δ)
        Where, xi Є D^s U D^t, Yi = 1
        If xi Є D^s and yi= -1
While D0≠ 0 do
        For  j=1 to k do
                Select xi from c satisfying |w^t xi - b| < 1 by:
                        Xi= arg min xi |w^t xi-b|
                Let D = D U{ xi }, D0=D0/{xi}.
        End for
        W^old =w
        Train a TVSM model using(xi,yi) and D.
        Calculate G
        If G≤t then
                Output D and Exit
        End if
End while.
```

IV. EXPERIMENT RESULT AND CONCLUSION

In this experimental result, first demonstrate that our algorithm using .NET to reduce the information gap between domains during the process of including the unlabeled data from the related domains. Randomly selects sample for these tasks and for each of the data sets display the performance, together with their corresponding margin sizes. For each iteration, include the top level unlabeled data that are closest to the decision boundary for information gap. The x-axis is the iteration count. The BIG algorithm is able to reduce the information gap and converge quickly. To compare the performance of the classification methods that use classification accuracy, which is defined as the percentage of correct predictions among all test examples. In order to validate the robustness of this method, for each time randomly sampled ninety percent training data for training the model. Each experiment was repeated ten times, and both the mean and variance of the accuracies is reported. All the experiments in different tasks are consistent and validate the effectiveness of this proposed algorithm. After the execution, it will investigate how our proposed method can find where the bridging documents are located in the knowledge base, and verify that our BIG algorithm could converge and its performance is stable. Figure 2 demonstrate the performance analysis between the BIG, local mining, and global learning

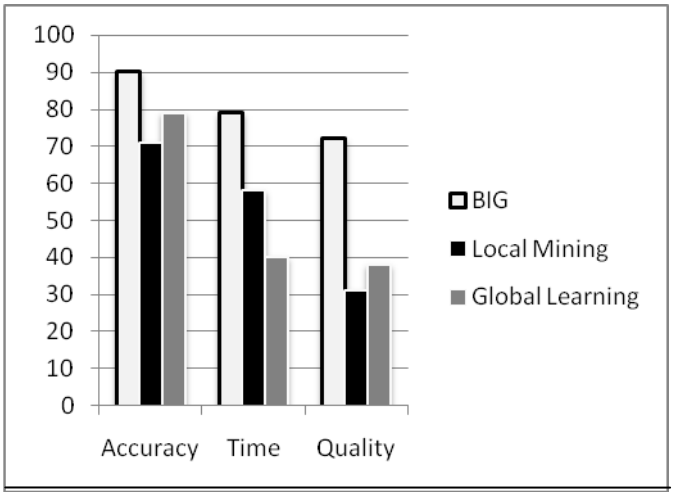

Fig. 2 Performance analysis

V. CONCLUSION

A major contribution of this work is that use of a large amount of worldwide knowledge to build a bridge for linking the source and target domains, even the distribution differences are large. It reduces the information gap between source domain and target domain. This is also worked in a large kind of domains. It shows that easily obtain useful information to bridge the source and the target domains from auxiliary data sources. Minmargin algorithm can effectively identify and reduce the information gap between two domains. Extensive evaluations on a real-world data set demonstrate that our scheme is able to produce a promising performance as compared to the prevailing coding methods. More importantly, the whole process is an unsupervised and holds potential to handle large-scale data. Every application has its own merits and demerits. This system has covered almost all the requirements. Further requirements and improvements can easily be done since the coding is mainly structured or modular in nature. First, to validate the effectiveness of this approach through other semi supervised learning algorithms and other relational knowledge bases to more extensively demonstrate the effectiveness of this approach. Second, investigate the source, target and auxiliary data sources share the same feature space. In future plan to extend this approach to be able to consider heterogeneous transfer learning.

REFERENCES

[1] Selma Ayse Ozel, "A Web page classification system based on a genetic algorithm using tagged-terms as features," Elsevier, pp. 3407–3415, 2011.
[2] Ali Ahmadia, Mehran Fotouhib, Mahmoud Khaleghi, "Intelligent classification of web pages using contextual and visual features," Elsevier, pp. 1638–1647, 2011.
[3] Selma Ayse Ozel, "A Web page classification system based on a genetic algorithm using tagged-terms as features." Elsevier, pp. 176-185, 2010.
[4] Pekka Malo, Pyry Siitari , Ankur Sinha, "Automated query learning with Wikipedia and genetic programming," Elsevier, pp. 87-95, 2012.
[5] Erdinç Uzuna, Hayri Volkan Agunb, Tarık Yerlikayab, "A hybrid approach for extracting informative content from web pages." Elsevier, pp. 928-944, 2013.
[6] P.Vivekanandan, M.Rajalakshmi, R.Nedunchezhian, "An intelligent Genetic algorithm for mining classification rules in large datasets," computing and informatics, pp. 1-22, 2013.
[7] Liqiang Nie, Yi-Liang Zhao, Mohammad Akbari, Jialie Shen, Tat-Seng Chua," Bridging the Vocabulary Gap between HealthSeekers and Healthcare Knowledge," IEEE transactions on knowledge and data engineering, pp.396-409, 2015.
[8] Kannuri Lahari, M. Ramakrishna Murty, and Suresh C. Satapathy, "Prediction based clustering using genetic algorithm and Learning Based Optimization Performance Analysis," Advances in Intelligent Systems and Computing," pp. 338, 2015.
[9] Edvin Aldana-Bobadhilla, Angel Kuri-Morales, "A Clustering based method on the maximum entropy principle," Entropy Article, pp. 151-180, 2015.
[10] Danial Gomes Ferrari, Leandro Numes de Castro, " Clustering algorithm selection by meta-learning systems: A new distance based problems characterization and ranking combination methods," Elsevier, pp.181-194, 2015.
[11] Jon Patrick, Yefeng Wang and Peter Budd, " An Automated System for Conversion of Clinical Notes into SNOMED Clinical Terminology," ACSW Frontiers conferrence, pp. 219-226, 2010.

[12] Xinmei Tian, Linjun Yang, Jingdong Wang, Yichen Yang, Xiuqing Wu, Xian-Sheng Hua, "Bayesian Video Search Reranking ", IEEE, pp. 131-140, 2011
[13] Koby Crammer and Mark Dredze and Kuzman Ganchev and Partha Pratim Talukdar, " Automatic Code Assignment to Medical Text", BioNLP proceedings of the workshop on BioNLP: Boiological, Translational,and clinical language processing, pp. 129-136, 2009.
[14] Rohit J. Kati, "Towards converting clinical phrases into SNOMED CT expressions" , Biomed inform insights, pp. 29-37, 2013.
[15] Saman Hina, Eric Atwell, Owen Johnson, "Semantic Tagging of Medical Narratives with Top Level Concepts from SNOMED CT Healthcare Data Standard ",International Journal of Intelligent Computing Research, pp. 204-210, 2011.
[16] K.Bhuvaneswari, M.Anusha, Dr.J.G.R Sathiaseelan, " A comaparative analysis of clustering techniques using Genetic Algorithm," ., International Journal of Computer Science and Mobile Computing , pp. 1-4, 2014.
[17] P. Bansal, S. Bertels, T. Ewart, P. MacConnachie, and O. James, "Bridging the research–practice gap," Acad. Manag. Perspectives, pp. 73–91, 2012.
[18] Gunjan Verma, Vineeta Verma, "Role and Application of Genetic Algorithm in DataMining," International Journal of Computer Application, pp. 5-8, 2012.
[19] R. L. Cilibrasi and P. M. B. Vitanyi, "The google similarity distance," IEEE , pp. 370–383, 2012.
[20] Raj Kumar, Dr. Rajesh Verma, "Classification Algorithms for Data Mining: A Survey," International Journal of Innovations in engineering and Technology, pp. 7-14, 2012.
[21]] E. J. M. Laur_ıa and A. D. March, "Combining Bayesian text classification and shrinkage to automate healthcare coding: A data quality analysis," International Journal of Innovations in engineering and Technology, pp. 20-31, 2013
[22] N.Chu, Y. Choi, J. Wei, and A. Cheok, "Games bridging cultural communications," IEEE Global Conferrence Consumer Electron., pp. 329–332, 2013.

