**International Journal of Engineering Research and Development**
*e-ISSN: 2278-067X, p-ISSN: 2278-800X, www.ijerd.com*
*Volume 8, Issue 4 (August 2013), PP.37-46*

# A Review on Semi Supervised Text Classification

Mahak Motwani[1], Hemlata Tekwani[2]
[1]*Truba Institute of Engineering & InformationTechnology Dept. of Computer Science & Engineering RGPV University, Bhopal*
[2]*Truba Institute of Engineering & Information Technology Dept. of Computer Science & Engineering RGPV University, Bhopal*

**Abstract:-** With the ever-increasing quantity of text data from a variety of online sources, it is a significant task to categorize or classify these text documents into categories that are manageable and easy to understand. In our old world, learning has been studied either in the unsupervised paradigm which include clustering where all the data is unlabeled, or in the supervised paradigm which include classification where all the data is labeled. A supervised classification of text demands labeled instances which are often arduous, formidable, expensive, or time consuming to obtain. During the intervening time, unlabeled data may be relatively easy to collect, but there are a couple of ways to use them and this method often clusters blindly. Semi-supervised learning figure out this problem by using labeled data together with large amount of unlabeled data to build better classifiers. We also make contribution towards this goal along several dimensions. This paper presents a survey on semi supervised methods of text classification using several Methods.

**Keywords:-** Semi-supervised learning; Text Classification; Labeled data; unlabeled data

## I. INTRODUCTION

Text classification is noteworthy and remarkable due to the large volume of text documents in many real-world applications. Text categorization or classification aims to assign categories or classes to unseen text documents.Categories may be represented numerically or using single word or phrase or words with senses, etc. In conventional approach, categorization of text was carried out manually using domain professionals. The human expert was requisite to read and sort the input text document to predefined category or set of categories. Thus, this approach necessitates wide-ranging human efforts and error prone also. This leads to the scheme of automated text classification scenario. Automated Text document categorization automatically assigns categories and facilitates simplicity of storage, searching, retrieval of appropriate text documents or its contents for the needy applications. [1]

There are three distinct paradigms which exist under text classification which are single label (Binary), multiclass and multi label. In single label classification a new text document belongs to exactly one of two specified classes, in multi-class case a new text document belongs to just one class of a set of m classes and in multi label text classification method each document may belong to several classes simultaneously. [2]

There are many approaches to implement multi label text classifier. More popular amongst these are supervised methods from machine learning. But majority of existing approaches are lacking in considering relationship between class labels, input documents and also relying on labeled data all the time for classification.[2] In real life unlabeled data is readily available whereas generation of labeled data is expensive and error prone as it needs human intervention. In many situations the available class labels are related to each other and consideration of this relationship can lead to better accuracy. Also, the abundantly available unlabeled data contains the joint distribution over features of an input dataset which may improve accuracy of overall classification process when used in conjunction with labeled data.

In supervised learning, the training sample consists of pairs, each containing an instance **x** and a label y: {(**x**i, yi)} for i=1 to n One can think of *y* as the label on **x** provided by a teacher, hence the name supervised learning. Such (instance, label) pairs are called labeled data. Classification is the supervised learning problem with discrete classes *Y*. The function f is called a classifier.Since the performance of supervised statistical classifiers often depends on the availability of labeled examples, one of the major bottlenecks toward automated text categorization is to collect sufficient numbers of labeled documents because of the high cost in manually labelling documents.

Unsupervised learning algorithms work on training Sample with n instances {xi} for i=1 to n.There is no teacher providing supervision as to how individual Instances should be handled this is the defining property of unsupervised learning. Unsupervised learning tasks include clustering, where the goal is to separate the *n* instances into groups, Novelty detection, which identifies the few instances that are very different from the majority, Dimensionality reduction, which aims to represent each instance with a lower dimensional feature

vector while maintaining key characteristics of the training sample. Unsupervised text classification does not need training data but is often criticized to cluster blindly.

There are two most significant strategies of Text Categorisation which include Active learning and semi-supervised learning. Active learning selects most informative unlabeled examples for manually labeling so that a good classifier can learn with significantly fewer labeled examples. Active learning has been extensively studied in machine learning for many years and has already been employed for text classification in the past. Semi-supervised learning attempts to learn a classification model from the mixture of labeled and unlabeled instances, which can be employed for text classification [3].

Semi-supervised learning can be applied when limited amount of training data is accessible. Because semi-supervised learning requires less human effort and Gives higher accuracy, it is of great advantage both in theory and in practice. In many Classification Applications labeled Training data are scarce but unlabeled data are plenteous. It is very usable if we can use unlabeled data to aid labeled data in learning a classifier. Semi Supervised learning deals with such problems. Some representative semi-supervised learning methods include Mixture model, EM, Transductive SVM, Cotraining, Graph Methods.

Earlier work in semi-supervised learning assumes that there are two classes, and in each class there is a Gaussian distribution. Hence we assume that the complete data comes from a mixture model. With a huge amount of unlabeled data, the mixture components are identified with the Expectation-Maximization (EM) algorithm. Only a single labeled example per component is required to fully determine the Mixture model. This particular model has been successfully applied to Text Categorization. A variant of this model is self-training: Firstly, A classifier is trained with the labeled data then it is used to classify the unlabeled data. The most confident unlabeled points and their predicted labels are added to the training set. The classifier is re-trained and this procedure is repeated. The classifier uses its own predictions to teach itself. This is basically a 'hard' version of the mixture model and EM algorithm. The method is also called self-teaching or bootstrapping1 in some research communities. But, any classification mistake can reinforce itself. Both methods have been used since long time ago. They remain popular because of their conceptual and algorithmic simplicity [4].

Co-training degrades the mistake reinforcing danger of self-training. This method is based on the assumption that the features of an item can be split into two subsets. To train a good classifier, each sub feature set is sufficient and the two sets are given an independent class. On each sub-feature set, two classifiers are trained with the labeled data, firstly. Each classifier then classifies the Unlabeled data iteratively and also teaches the other classifier with its own predictions [4].

With the popularity of support vector machines (SVMs), transductive SVMs have emerged which are extension to standard SVMs for semi-supervised classification. Transductive SVMs find labels for the unlabeled data, and a separate hyper plane, and hence maximum margin can be achieved on both the labeled data and the unlabeled data. [4]. When we deal with the gene expression datasets, many effortful challenges such as curse of dimensionality and insufficient labeled data is inevitable. A new method came

Known as Iterative Transductive Support Vector Machine (ITSVM).This proposed algorithm when applied on gene expression datasets showed that it can exploit unlabeled data distribution. Also this method improves the accuracy as compared to other related methods**.** The experimental results demonstrate that ITSVM is not sensitive to datasets and informal decision in labeling samples can lead to better generalization. This proposed method calculates the quantitative value for unlabeled samples and also chooses best action at every step. This feature could also help us in building a novel transductive multi-class classifier [5].

Lately, graph-based semi-supervised learning methods have also attracted great attention. These methods starts with a graph where the nodes represent the labeled and unlabeled data points and edges which are weighted reflect the similarity of nodes. Here the assumption made is that the nodes connected by a large-weight edge tend to have the same label, and labels can propagate throughout the graph. Graph-based methods enjoy nice properties from spectral graph theory also [4].

## 1. BACKGROUND

The high quantity of electronic information obtainable on the Internet increases the difficulty of dealing with it in modern years. The less complicated methods for web page segmentation rely on structure of wrappers for a specific type of web pages. The obligation that the new material not be in the text overtly means that the system must have access to external information of some category, such as a knowledge base or an ontology, and be able to perform combinatory deduction. Since no large-scale resources of this kind yet exist. Due to the poor performance of the initially learned hypothesis based on the very few training data, it is inescapable to restrain much noise in the self-labeled instances. Extremely few labeled training data in sparsely labeled text classification aggravate such situation. A variety of algorithms are proposed in this area and it is a widely chosen area for recent development.

## II. LITERATURE SURVEY

In the Year 2001 Rayid Ghani proposed a Method for "Combining labeled and unlabeled data for text classification with a large number of categories [6]. They developed a framework to incorporate unlabeled data in the Error-Correcting Output Coding (ECOC) setup by decomposing Multiclass problems into multiple binary problems and then CO-Training is used to learn the individual binary classification problems. They used a dataset obtained from WhizBang Labs consisting of Job titles and Descriptions organized in a two Level hierarchy with 15 first level categories and 65 leaf categories. All the codes used were BCH codes. They have shown that the framework presented in this paper results in text classification systems that are both computationally efficient and need very few labeled examples to learn accurately. Their approach is more efficient since use of ECOC reduces the number of models that classifier constructs and hence this approach scales up sublinearly with the number of classes.

In the year 2005 Steven M Beitzel at all proposed "Improving Automatic Query Classification via Semi-supervised Learning" [7]. An application of computational linguistics to generate an approach for mining the vast amount of unlabeled data in web query logs to improve automatic Topical web query classification is proposed in this method. They showed that their approach in combination with manual matching and supervised learning allows classifying a substantially larger proportion of queries than any single technique. They examined the performance of each approach on a real web query stream and showed that their combined method accurately classifies 46% of queries, outperforming the recall of best single approach by nearly 20%, with a 7% improvement in overall effectiveness.. This large increase in recall proves that the combined approach may in fact an interesting solution to the recall problem that has hindered past efforts at query classification.

In the year 2005 Kamal Nigam, Andrew McCallum, Tom Mitchell proposed "Semi-Supervised Text Classification Using EM" [8]. This approach when applied to the domain of text classification proved very effective and remarkable. Text files are characterized here with a bag-of-words model, which advances to a generative classification model based on a mixture of multinomials. This model is an extremely simplistic representation of the complexities of written text. They also demonstrated that deterministic annealing, a variant of EM, can help overcome the problem of local maxima and increase classification accuracy further when the generative Model is appropriate. Here, Expectation-Maximization finds more likely models and improved classification accuracy. Likelihood and better accuracy are not well correlated with the naive Bayes model in other domains. Here, they have used a more expressive generative model that allows for multiple mixture components per class. This helps restore a moderate correlation between model likelihood and classification accuracy, and again EM finds more accurate models. Here it is proved that even with a well-correlated generative model; local maxima are a convincing hindrance with EM.

In the year 2006 Rong Liu, Jianzhong Zhou, Ming Liu, proposed "A Graph-based Semi-supervised Learning Algorithm for Web Page Classification*" [4]. This paper proposes a graph-based semi-supervised learning algorithm which applied to the web page classification. A *K*-Nearest Neighbor graph is constructed using this algorithm which uses a similarity measure between web pages. Labeled and unlabeled web pages are represented as nodes in the weighted graph and edge weights encode the similarity between the various web pages. Combining weighting schemes and link information of web pages helps in computing edge weights of the graph .The learning problem is then formulated in terms of label propagation in the graph. The labeled nodes push out labels through unlabeled nodes by using probabilistic matrix methods and belief propagation. Graph-based semi-supervised learning method performs better than Harmonic Gaussian model and TSVM. The graph-based semi-supervised learning method could also be used for enhancing web search.

In the year 2008 Zenglin Xu et al proposed "Semi-supervised Text Categorization by Active Search" [1] .For agglomeration of the unlabeled documents with the help of web search engines and utilizing them to improve the accuracy of supervised text classification, they characterized a general framework for semi-supervised text categorization. Experimental Results have established that the projected semi-supervised text categorization framework can incomparably improve the classification accuracy.

In the year 2008 shiliang sun proposed "Semantic Features for Multi-view Semi-supervised and Active Learning of Text Classification"[9].For pattern representation. Semantic features assimilating information from multiple views are gathered. For learning the representation of semantic spaces where semantic features are projections of original features on the basis vectors of the spaces, Canonical correlation analysis is used .They cross-examined the practicability of semantic features on two learning paradigms which include semi - supervised learning and active learning. This use of semantic features can bulge to a convincing improvement of attainment and accuracy.

In the year 2008 HuanLing Tang, ZhengKui Lin, Mingyu Lu, Na Liu proposed "A Novel Features Partition Algorithm for Semi-Supervised Categorization" [10].They have given formulas to evaluate mutual independence between two features, feature and sub-view, sub-view and sub-view and these features with weaker mutual independence are categorized in the same sub-view, those with stronger mutual independence are categorized in separate Sub-views. This method can effectively split features set into two sub-views with higher independence. A new semi-supervised categorization algorithm which is known as SC-PMID is promoted based

on Partition-MID algorithm. When labeled data is sparse, this strategy utilizes both unlabeled data and labeled data and incomparably improves classification precision.

In the year 2009 HAN Hong qil, ZHU Dong-hua, WANG Xue-feng proposed "Semi-supervised Text Classification from Unlabeled Documents Using Class Associated Words" [11]. A semi-supervised classification algorithm is proposed which requires use of the prior knowledge of class associated words. Class associated words are defined as the words which represent the subject of classes and provide prior knowledge of classification for training a classifier. The combination of Expectation-Maximization and a Naive Bayes classifier is introduced as a new algorithm to categorize documents from fully unlabeled documents using class associated words. The algorithm first iterates to build the probabilistically-weighted Association between documents and class associated Words, and then assigns class labels for documents According to the relations between classes and class Associated words. Such kind of class associated words are basically used to set classification constraints during learning process to restrict to classify documents into corresponding class labels and help in advancing the classification accuracy and competence . Experimental results demonstrate that it has better classification competence and accuracy for those categories which have small quantities of samples.

In the year 2009 Mohammad Salim Ahmed, Latifur Khan proposed "A Text Classification Approach Using Semi Supervised Subspace Clustering known as SISC" [12]. Semi-supervised Impurity based Subspace Clustering known as SISC in used with κ-Nearest Neighbor approach and it is based on semi-supervised subspace clustering that considers sparse nature in text data and the high dimensionality in text data. This method catches clusters in the text document subspaces which have high dimensional text data and fuzzy cluster membership. There are two important factors of this method which include chi square statistic of the dimensions and the impurity measure within each cluster. Experiments on real world data sets reveal the significance of this remarkable approach as it incomparably outperforms other state-of-the-art text classification and subspace clustering algorithms. This algorithm achieves an Area under The ROC Curve value of 0.813 whereas the closest any other method can achieve is 0.77.

In the year 2009 Frank Lin and William W. Cohen proposed "Semi-Supervised Classification of Network Data Using Very Few Labels" [13] They proposed MultiRankWalk, a semi-supervised learning method as a simple yet intuitive representative of a class of semi-supervised learning methods based on Random graph walks, and show it to significantly out Perform other semi-supervised and supervised learning Methods when only a few labeled instances are given on Five network datasets. They also showed that using high authority labeled in-stances dramatically reduce the amount of labels required to achieve high classification performance, which sheds light on why random graph walk-based methods have an advantage over methods such as Gaussian fields classifier
when the size of training data is small.

In the year 2010 Fangming Gu, Oayou Liu and Xinying Wang Proposed "Semi-Supervised Weighted Distance Metric Learning for kNN Classification" [14]. To increase the classification Information which is provided by user this method uses a graph-based semi-supervised Label Propagation algorithm and then adopts an approach of improved weighted Relevant Component Analysis to learn a Mahalanobis distance function. After such processes, Mahalanobis Distance metric function is used to replace the Euclidean distance of original kNN classifier. Experiments and attempts on UCI datasets show that this method can significantly enhance the accuracy of kNN classification.

In the year 2010 Fang Lu and Qingyuan Bai proposed

"Semi-supervised Text Categorization with Only a few Positive and Unlabeled Documents [15]. In this paper a refined method to do the PU-Learning with the known technique combining Rocchio and K-means algorithm is proposed. They have described that a text classifier can be built with a set of labeled positive documents from one class which is known as Positive class and a set of large number of unlabeled documents from both positive class and other diverse classes .This kind of semi-supervised text classification is called positive and unlabeled learning (PU-Learning).The experimental results show that the technique has better performance in PU-Learning when P is very small.

In the year 2010 Lei Shi, Rada Mihalcea and Mingjun Tian proposed "Cross Language Text Classification by Model Translation and Semi-Supervised Learning" [16]. In this paper, a method that automatically builds text classifiers in a new language by training on already labeled data in another language is proposed. This method transfers the classification knowledge across languages by translating the model features and by using an Expectation Maximization (EM) algorithm. Moreover, the model is tuned to fit the distribution in the target language with the assistance of semi-supervised learning. Also remarkable improvement is shown over previous methods that rely on machine translation by Experiments on different datasets covering different languages and different domains.

In the year 2011 Yawei Chang and Houquan Liu proposed

"Semi- Supervised Classification Algorithm based on the KNN" [17]. An approach based on the EM-KNN semi-supervised classification is described in which firstly the center of each category is calculated, to cluster the training set then center of each category is combined and finally text is clustered to form new training

set. The new training set obtained is trained with classical KNN algorithm. Experimental results demonstrate that the overall computational complexity can be reduced and the accomplishment and implementation of the classifier can also be improved by this algorithm.

In the year 2012 Nagesh Bhattu and D.V.L.N. Somayajulu "Semi-supervised Learning of Naive Bayes Classifier with feature constraints" [18]. In this Method, An objective function is used which learns both from labeled data and feature constraints over unlabeled data and results in a single point solution. Posterior regularization (PR) is a Framework recently proposed for incorporating bias in the form prior knowledge into posterior for the label. The main focus is on incorporating labeled features into a naïve bayes classifier in a semi-supervised setting using PR framework. Generative learning approaches utilize the unlabeled data more effectively compared to discriminative approaches in a semi-supervised setup. In this paper they have formulated a classification method which uses the labeled features as constraints for the posterior in a semi-supervised generative learning setting. Their experimental results show how very few feature constraints can also help to improve the classifier by a significant margin over the base-line.

In the year 2012 Wang- xin Xiao, Xue Zhang proposed "Active Transductive KNN for Sparsely Labeled Text Classification"[19]. An active transductive KNN framework (AcTrKRF) is proposed in this paper, which is designed for very sparsely labeled classification problem. This algorithm works by combining active learning and transductive learning together, and borrowing the thinking of self-training and multi-view learning. It integrates the supremacy of semi-supervised learning and active learning and also employs several techniques to cope with the training data bias and sparsity. The fusion of active learning with rechecking strategy, and the employment of common feature extraction technique, makes this framework robust to the training data bias and sparsity. Experimental results show that this algorithm is effective and efficient for sparsely labeled classification problem and that it significantly outperforms the baseline model KNN and several state-of-the-art algorithms.

**Semi-Supervised Methods, Their Advantages and Drawbacks:**

| METHOD | ADVANTAGE | DRAWBACK |
|---|---|---|
| 1. Combining labeled and unlabeled data for text classification with a large number Of categories [6]. (Rayid Ghani,2001) | Method is especially useful for classification tasks involving a large number of categories where CO-training doesn't perform very well by itself and when combined with ECOC, outperforms several other algorithms that combine labeled and Unlabeled data for text classification in terms of accuracy, trade-off and efficiency. | Since the two classes in each bit are created artificially by ECOC and consist of many "Real" classes, there is no guarantee that CO-Training Can learn these arbitrary binary functions. |
| 2. Improving Automatic Query Classification via Semi-supervised Learning" [7]. (Steven M. Beitzel, Eric C. Jensen, Ophir Frieder, | Using this approach in combination with Manual matching and supervised learning allows | selectional preference classifiers cannot Make classification decisions on single-term |

| | | |
|---|---|---|
| David D. Lewis, Abdur Chowdhury, Aleksander Kolcz ,2005) | Us to classify a substantially larger proportion of queries than any single technique. | queries. If evaluated over multi-term queries alone, higher Recall is observed. |
| 3 .Semi-Supervised Text Classification Using EM(Kamal Nigam,Andrew McCallum, Tom Mitchell, 2005)[8] | Expectation-Maximization finds more likely models and improved classification accuracy | the approach of deterministic annealing does provide much higher likelihood models, but often loses the correspondence With the class labels. When class label correspondence is easily corrected, high accuracy Models result. |
| 4.A Graph-based Semi-supervised Learning Algorithm for Web Page Classification*[4] (Rong Liu, Jianzhong Zhou Ming Liu,2006) | Graph-based semi-supervised learning method performs better than Harmonic Gaussian model and TSVM. | The weight learning algorithm taking account of link information tends to be more computationally expensive, and It is also apparent from the result that the benefit of semi - supervised Learning diminishes as the labeled set size grows. |
| 5.Semi-supervised Text Categorization by Active Search [1](Zenglin Xu et al, 2008) | A general framework for semi-supervised text categorization that collects the unlabeled documents via web search engines and utilizes them to improve the accuracy of supervised text categorization | Less efficient as compared to the existing work |

| | | |
|---|---|---|
| 6. Semantic Features for Multi-view Semi-supervised and Active Learning of Text Classification[9] (Shiliang Sun,2008) | Experiments on text classification with two state-of the-art multi-view learning algorithms co-training & cotesting Indicate that this use of semantic features can lead to a significant improvement of performance. | The feasibility of semantic features on other applications, Must be taken into account. |
| 7A Novel Features Partition Algorithm for Semi-Supervised Categorization"[10] .(HuanLing Tang , ZhengKui Lin, Mingyu Lu, Na Liu 2008) | Based on Partition-MID algorithm, a new semi-supervised categorization algorithm named SC-PMID is also proposed.SC-PMID algorithm can significantly improve classification, especially When labeled data is sparse. | Not very efficient. |
| 8. " Semi-supervised Text Classification from Unlabeled Documents Using Class Associated Words" [11]( HAN Hong qil, ZHU Dong-hua, WANG Xue-feng ,2009) | Training set does not need to be provided for classification And consistency ratio of 92.66% is achieved. | The algorithm is based on strict assumptions. |
| 9. " SISC: A Text Classification Approach Using Semi Supervised Subspace Clustering" [12] (Mohammad Salim Ahmed, Latifur Khan,2009) | SISC performs well while considering both labeled and Unlabeled data and minimizes the effect of high dimensionality | Chi Square Statistic component included in the Objective function and Impurity component used to modify the dispersion |

| | | |
|---|---|---|
| | and its sparse Nature during training. | measure requires more calculations. |
| 10. Semi-Supervised Classification of Network Data Using Very Few Labels [13].(Frank Lin and William W. Cohen ,2009) | This method shows that using high authority labeled in-stances dramatically reduce the amount of labels required to achieve high classification performance, which sheds light on why random graph walk-based methods have an advantage over methods such as Gaussian fields Classifier when the size of training data is small. | It is always important for the algorithm to propagate the labels further by not "damping" the walk too much, especially when the Number of labeled instances is small. |
| 11. Semi-Supervised Weighted Distance Metric Learning for kNN Classification [14] ( Fangming Gu, Oayou Liu, Xinying Wang,2010) | This method uses a graph-based semi-supervised Label Propagation algorithm to increase the classification Information and significantly improve the accuracy of kNN classification. | The classification result depends on the distribution of labeled data points, and if the distribution is uneven the classification accuracy will be reduced greatly |
| 12. "Semi-supervised Classification Algorithm Based on the KNN"[17].( In the year 2011 Yawei Chang and Houquan Liu,2011) | EM-KNN algorithm is better than the traditional KNN algorithms in accuracy, Reducing computational complexity greatly & combination | The complexity of the early training process relative to the training process complexity 0 of the KNN algorithm have Many defects. |

| | | |
|---|---|---|
| | property. | |
| 13. "Active Transductive KNN for Sparsely Labeled Text Classification" [19] ( Wang- xin Xiao, Xue Zhang, 2012) | This algorithm is effective and efficient for sparsely labeled classification problem and that it significantly Outperforms the baseline model KNN. | Powerful data editing techniques such as CCA and kernel functions are not used. |

## III. CONCLUSION

In this paper we present numerous techniques that are used to classify text using various semi supervised classification.

We render an approach for mining the vast amount of unlabeled data from web. Since the performance of supervised statistical classifiers often depends on the availability of labeled examples and unsupervised text classification does not need training data but is often criticized to cluster blindly, using and implementing semi-supervised learning methods to text classification is desirable to build better classifiers. Because semi-supervised learning gives higher accuracy and requires less human effort, it's of great advantage both in theory and in practice. Further, our hope is that by leveraging unlabeled data the need for periodically labeling new training data can be minimized to keep up with changing trends in the query stream over time.

## REFERENCES


[1] Zenglin Xu, Rong Jin, Kaizhu Huang, Michael R. Lyu, Irwin King "Semi-supervised Text Categorization by Active Search", Proceedings of the 17th ACM conference on Information and knowledge management, pp. 1517-1518, 2008.

[2] Shweta C. Dharmadhikari, Maya Ingle, Parag Kulkarni "A Novel Multi label Text Classification Model using Semi supervised learning", International Journal of Data Mining & Knowledge Management Process (IJDKP) Vol.2, No.4, pp. 11 – 20, July 2012.

[3] Xue Zhang and Wangxin Xiao "Active Semi-supervised Framework with Data Editing", International Conference on Systems and Informatics (ICSAI -2012), Vol. 9, No. 4, Special Issue, pp. 46 – 50, 2012.

[4] Rong Liu, Jianzhong Zhou, Ming Liu, "A Graph-based Semi-supervised Learning Algorithm for Web Page Classification*" IEEE 2006 Proceedings of the Sixth International Conference on Intelligent Systems Design and Applications (ISDA'06) (Volume:2 ),pp. 856-860

[5] Hossein Tajari and Hamid Beigy "Gene Expression Based Classification using Iterative Transductive Support Vector Machine "International Journal of Machine Learning and Computing vol. 2, no. 1, pp. 76-81, 2012.

[6] Rayid Ghani "Combining labeled and unlabeled data for text classification with a large number of categories" Proceedings of the IEEE International Conference on Data Mining, version-3, pp. 597-598, 2001

[7] Steven M. Beitzel, Eric C. Jensen, Ophir Frieder, David D. Lewis, Abdur Chowdhury, Aleksander Kolcz "Improving Automatic Query Classification via Semi-supervised Learning" 2005 Proceedings of the Fifth IEEE International Conference on Data Mining (ICDM'05),pp.42-49.

[8] Kamal Nigam, Andrew McCallum, Tom Mitchell "Semi-Supervised Text Classification Using EM"2005 In Semi-Supervised Learning, citeulike: 5393729 pp.33-56 Key: citeulike: 5393729

[9] Shiliang Sun "Semantic Features for Multi-view Semi-supervised and Active Learning of Text Classification" 2008 IEEE International Conference on Data Mining Workshops ICDMW'08, pp.731-735.

[10] HuanLing Tang, ZhengKui Lin, Mingyu Lu, Na Liu "A Novel Features Partition Algorithm for Semi-Supervised Categorization" IEEE Intelligent Control and Automation, 2008. WCICA, 7th World Congress on, pp. 129 – 134.

[11] HAN Hong Qil, ZhU Dong-hua, WANG Xue-feng "Semi-supervised Text Classification from Unlabeled Documents Using Class Associated Words" IEEE Computers & Industrial Engineering, 2009 CIE 2009. International Conference, pp. 1255 – 1260.

[12] Mohammad Salim Ahmed, Latifur Khan "SISC: A Text Classification Approach Using Semi Supervised Subspace Clustering" 2009 IEEE International Conference on Data Mining Workshops, , pp.1-6

[13] Frank Lin and William W. Cohen "Semi-Supervised Classification of Network Data Using Very Few Labels", 2009 Language Technologies Institute; School of Computer Science; Carnegie Mellon University pp.1-6

[14] Fangming Gu, Oayou Liu and Xinying Wang "Semi-Supervised Weighted Distance Metric Learning for kNN Classification" IEEE Computer, Mechatronics, Control and Electronic Engineering (CMCE), 2010 International Conference on (Volume:6 ), pp. 406-409

[15] Fang Lu and Qingyuan Bai "Semi-supervised Text Categorization with Only a Few Positive and Unlabeled Documents" Biomedical Engineering and Informatics (BMEI), 2010 3rd International Conference on (Volume: 7), pp. 3075 – 3079

[16] Lei Shi, Rada Mihalcea and Mingjun Tian "Cross Language Text Classification by Model Translation and Semi-Supervised Learning" Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing, pp. 1057–1067.

[17] Yawei Chang and Houquan Liu"Semi-supervised Classification Algorithm based on the KNN". Communication Software and Networks (ICCSN), 2011 IEEE 3rd International Conference, pp-9-12

[18] Nagesh Bhattu and D.V.L.N. Somayajulu "Semi-supervised Learning of Naive Bayes Classifier with feature constraints", Proceedings of the First International Workshop on Optimization Techniques for Human Language Technology (COLING 2012), pp. 65–78, December 2012

[19] Wang- xin Xiao, Xue Zhang "Active Transductive KNN for Sparsely Labeled Text Classification" IEEE Soft Computing and Intelligent Systems (SCIS) and 13th International Symposium on Advanced Intelligent Systems (ISIS), 2012 Joint 6th International Conference on, pp. 2178 – 2182