Attribute Overlap Minimization and Outlier Elimination as Dimensionality Reduction Techniques for Text Classification Algorithms

Simon Fong
Department of Computer and Information Science, University of Macau, Macau SAR
Email: ccfong@umac.mo

Antonio Cerone
International Institute for Software Technology, United Nations University, Macau SAR
Email: antonio@iist.unu.edu

***Abstract*—Text classification is the task of assigning free text documents to some predefined groups. Many algorithms have been proposed; in particular, dimensionality reduction (DR) which is an important data pre-processing step has been extensively studied. DR can effectively reduce the features representation space which in turn helps improve the efficiency of text classification. Two DR methods namely Attribute Overlap Minimization (AOM) and Outlier Elimination (OE) are applied for downsizing the features representation space, on the numbers of attributes and amount of instances respectively, prior to training a decision model for text classification. AOM works by swapping the membership of the overlapped attributes (which are also known as features or keywords) to a group that has a higher occurrence frequency. Dimensionality is lowered when only significant and unique attributes are describing unique groups. OE eliminates instances that describe infrequent attributes. These two DR techniques can function with conventional feature selection together to further enhance their effectiveness. In this paper, two datasets on classifying languages and categorizing online news into six emotion groups are tested with a combination of AOM, OE and a wide range of classification algorithms. Significant improvements in prediction accuracy, tree size and speed are observed.**

***Index Terms*—Data stream mining, optimized very fast decision tree, incremental optimization.**

I. Introduction

Text classification is a classical text mining process that concerns automatically sorting unstructured and free text documents into predefined groups [1]. This problem receives much attention from researchers from data mining research community for its practical importance in many online applications such as automatic categorization of web pages in search engines [2], detection of public moods online [3] and information retrieval that selectively acquire online text documents into the preferred categories.

Given the online nature of the text classification applications, the algorithms often would have to deal with massive volume of online text that are stored in unstructured format, such as hypertexts, emails, electronic news archive and digital libraries. A prominent challenge or difficulty of text classification application is processing the high dimensionality of the attribute representation space manifested from the text data.

Text information is often represented by a string variable which is a single dimensional data array or linked-list in computer memory. Though the size of a string may be bounded, a string variable can potentially contain infinite number of words combinations; each string that represents an instance of text document will have a different size. The large number of values from the training dataset and the irregular length of each instance make training a classifier extremely difficult. To tackle this issue, the text strings are transformed into a fixed-sized list of attributes that represent the frequency of occurrence of each corresponding word in the dataset. The frequency list is often called Word Vector in the form of a bit-vector, which is an occurrence frequency representation of the words. The length of a word vector is bounded by the maximum number of unique words exist in the dataset. An example in WEKA, which stands for 'Waikato Environment for Knowledge Analysis' is a popular suite of machine learning software written in Java, developed at the University of Waikato, illustrates how sentences in nature language are converted to word vector of frequency counts.

doi:10.4304/jetwi.4.3.259-263

```
@relation news

@attribute snipplet string
@attribute class {good_news, bad_news}

@data
'Cancer Research Breakthrough', good_news
'Syria Crisis Worsen', bad_news
```

↓

```
@relation news-transformed

@attribute class {good_news, bad_news}
@attribute Cancer numeric
@attribute Research numeric
@attribute Breakthrough numeric
@attribute Syria numeric
@attribute Crisis numeric
@attribute Worsen numeric

@data
good_news,1,1,1,0,0,0
bad_news,0,0,0,1,1,1
```

Figure 1. An example of text string converted to word vector.

Although word vectors could be processed by most classification algorithms, the transformation approach is no scalable. For large texts that contain many words the word vector grows to prohibitory huge which slows down the model training time and it leads to the well-known data mining problem called 'curse of dimensionality'. The word vector is most sparse that occupies unnecessary run-time memory space.

Hence dimensionality reduction techniques (DR) are extensively studied by researchers. The techniques aim to reduce the number of components of a dataset such as word vectors; while at the same time, the original data are represented as accurately as possible. DR often yields fewer features and/or instances. Therefore a compact representation of the data could be achieved for improving text mining performance and reduces computational costs.

Two types of DR are usually applied, often together, for reducing the number of attributes/features and to streamlining the amount of instances. They attempt to eliminate irrelevant and redundant attributes/data from the training dataset and/or its transformed representation, making the training data compact and efficient for data-intensive task like constructing a classifier.

In paper, a DR method called Attribute Overlap Minimization (AOM) is introduced which reduces the number of dimensions by refining the membership of each group that the word vector is more likely to belong to. Furthermore the corresponding instances that do not fit well in the rearranged groups are removed. This paper reports about this DR technique and experiments are conducted to demonstrate its effectiveness over two different datasets.

II. Model Framework

A typical text mining workflow consists of data pre-processing that includes data cleaning, formatting and missing value handling, dimensionality reduction and data mining model training. Figure 1 shows such a typical text mining workflow. A classifier which is enabled by data mining algorithms needs to be trained initially by processing through a substantial amount of pre-labeled records to an acceptable accuracy, before it could be used for classifying new unseen instances to the predicted groups.

The data mining algorithms are relatively mature in their efficacy and their performance is largely depending on the quality of the training data – which is the result of the DR that tries to abstract the original dataset to a compact representation.

A type of DR methods which is well-known as Stemming [4] has been proposed and widely used in the past. Stemming algorithms or so-called stemmers are designed to reduce a single word to its stem or root form [5] by finding its morphological root. This is done by removing the suffix of the words. It helps shorten the length of most terms. The other important type of DR is Feature Selection which selects only the attributes whose values represent the words that exist in the text document, and it filters off those attributes that have less or little predictive power with respect to the classification model. So a subset of the original attributes can be retained for building an accurate model. A comparative study of different feature selection methods [6] have been evaluated pertaining to subsiding text space dimensionality. It was shown possible that between 50% and 90% of the terms from the text space can be removed by using suitable feature selection schemes without sacrificing any accuracy in the final classification model.

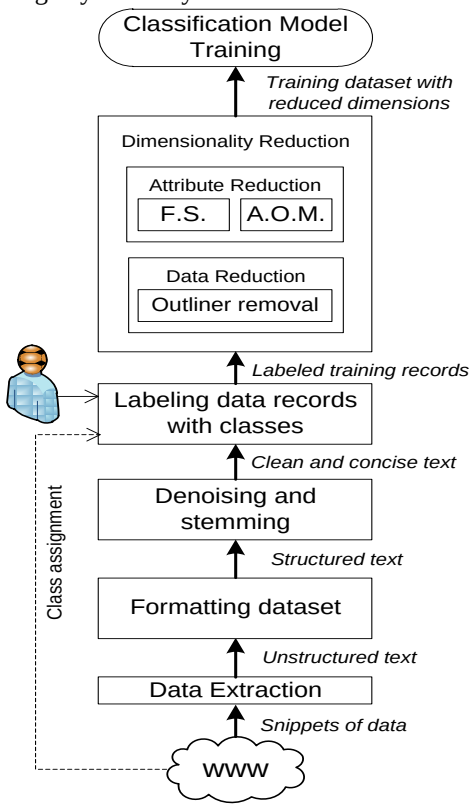

Figure 2. A typical text-mining workflow.

Both types of DR methods reduce the dimensionality of a dataset as an important element of the text data pre-processing stage. However, it is observed that feature selection heavily removes less-important attributes based

on their potential contributing power in a classifier, without regards to the context of the training text data. We identify that one of the leading factors to misclassification is the confusion of the contexts of the words in different groups. The confusion disrupts the training process of the classification model by mistakenly interpreting a word/term from an instance as an indication to one group but in fact it is more likely to belong to another. A redundant and false mapping-relation between the attributes and the target group is therefore created in the model that dampers the accuracy of the resultant classification model.

The source of this problem is originated from the common attributes which are owned by more than one group. A single term, without referring to the context of its use, can be belonged to two or more target groups of text. For the example given in Figure 1, the individual term 'Cancer' is actually in a case that belongs to 'Good news', while the same term can potentially and intuitively be deemed as an element of 'Bad news'.

To rectify this problem a data pre-processing method called Attribute Overlap Minimization (AOM) is proposed. In principle, it works by relocating the terms to a group in which the term has the highest occurrence frequency. The relocation can be absolute, that is based on the Winner-takes-all approach. The group that has the highest frequency count of the overlapped word recruits it all. In the dataset, the instances that contain the overlapped words would have to delete them off if the labeled class group is not the winner group. The instances that belong to the winner group continue to own the words for describing the characteristic of the group. Another milder approach is to assign 'weights' according to the relative occurrence frequencies across the groups. The strict approach may have a disadvantage of over-relocation that leads to a situation where the winner group monopolizes the ownership of the frequently occurring terms, leaving the other groups lack of key terms for training up their mapping relations. However, when the instances have a sufficient number of instances and the overlapped terms are not too many, AOM works well and fast. Comparing to FS, AOM is having the advantage of preserving most of the attributes and yet it can prevent potential confusion in the classification training. Another benefit is the speed due to the fact that it is not necessary to refer to some ontological information during the processing. An example is shown in Figure 3, where in linguistic languages common words that have the same spellings are overlapped across different languages. AOM is a competitive scheme that allows a language group in which the words appear most frequently acquires away the overlapped words.

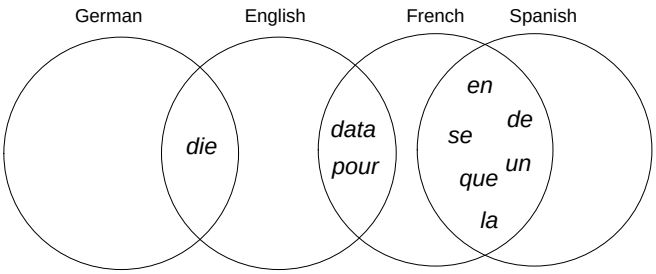

Figure 3. An illustration of overlapped words among different languages

III. EXPERIMENT

In order to validate the feasibility our proposed model, a text mining program is built in WEKA over two representative datasets, and by using a wide range of classification algorithms. We aim to study the performance of the classifiers together with the use of different dimensionality reduction methods. The training data which are obtained from online websites are unstructured in nature. After the conversion the word vector grows to a size of 8135 attributes for maintaining frequency counts for each word in the documents. A combination of DR techniques is applied in our experiment. An outliner removal algorithm is used for trimming off data rows that have exceptionally different values from the norm. For reducing the number of attributes, a standard Feature Selection algorithms (FS) called Chi-Square is used because of its popularity and generality, together with our novel approach called Attribute Overlap Minimization (AOM) are applied.

Two training data are used in the experiment: one is a collection of sample sentences on the related topics of data mining, retrieved from Wikipedia websites of four different languages – Spanish, French, English and German. The other one is excerpted from CNN news website, of the news articles that were released for ten days across the New Year 2012. The news collection has a good mix of political happenings, important world events and lifestyles. One hundred of sample news was obtained in total, and they were rated manually according to six basic human psychological emotions, namely, Anger, Fear, Joy, Love, Sadness and Surprise. The data are formatted into ARFF format, having one news/instance per row in the following structure: <*emotion*>, <"*text of the news*"> where the second field has a variable length. Similarly for the language sample dataset, the structure is <*language*>, <" *wiki page text*">. The HTML tags, punctuation mark and symbols are filtered off. The training datasets are then subject to the above-mentioned dimensionality reduction methods for transformation to a concise dataset in which the attributes have substantial predictive powers contributing to. Accuracy which is a key performance indicator is defined by the percentage of the number of correctly classified instances over the total number of instances in the training dataset. Others are decision tree size or the amount of generated rules which implies the demand of the runtime memory requirement, and the time taken for

training up the model. By applying attribute reduction and data reduction, we can observe that the initial number of attributes have reduced greatly from 8135 to 11. Having a concise and elite amount of attributes is crucial in real-time application, and in text mining online news, the number of attributes is proportional to the coverage of news articles – the more unique words (vocabularies) that are being covered, the greater the number of attributes there are.

TABLE I.

PERFORMANCE OF DECISION TREE MODEL TESTED UNDER DIFFERENT TYPES OF DR METHODS APPLIED, LANGUAGE DATASET.

DR Methods	Accuracy %	Tree size	Time (s)	# attributes	# instances
None (Original)	86.7725	23	2.47	1902	189
FS	86.7725	23	0.22	187	189
AOM	89.418	19	2.19	1891	189
OE	92.6966	19	1.69	1902	178
FS+AOM	89.418	19	0.2	182	189
FS+AOM+OE	98.8764	19	0.16	173	178

TABLE II.

PERFORMANCE OF DECISION TREE MODEL TESTED UNDER DIFFERENT TYPES OF DR METHODS APPLIED, EMOTION DATASET.

DR Methods	Accuracy %	Tree size	Time (s)	# attributes	# instances
None (Original)	24	53	10.42	8135	100
FS	33	15	0.03	11	100
AOM	51	39	5.24	8124	100
OE	30.1587	33	3.4	8135	63
FS+AOM	69	21	0.08	52	100
FS+AOM+OE	75.8242	21	0.06	50	91

In general, it can be seen that the results from the above tables have the smallest tree size, highest accuracy and a very short training time when the three DR methods are used together. The language dataset represents a scenario where the number of attributes is approximately 10 times larger than the number of instances which is usual in text mining when vector space is used. The emotion dataset represents an extremely imbalanced case where the ratio of attributes to instances is greater than 80:1. It should be highlighted that by applying a series of FS+AOM+OE in the extreme case of emotion dataset, the number of attributes was not cut to an extremely small number (50 instead of 11) that are sufficient to characterize an emotional group, the instances amount are not overly eliminated (91 over 63) for sufficiently training the model; yet the accuracy achieved is the highest possible.

The experiment is then extended to evaluate the use of machine learning algorithms, with the benchmarking objective of achieving the highest accuracy. The selection list of the machine learning algorithm used in our experiment here is by no means exhaustive, but will form the basis of a performance comparison which should supposedly cover most of the popular algorithms. The machine learning algorithms are grouped by four main categories, Decision Tree, Rules, Bayes, Meta and Miscellaneous; all of them are known to be effective for data classification in data mining to certain extents.

Three versions of inflected datasets were text-mined by different classification algorithms in this experiment. They are the dataset with FS only, transformed dataset with reduced attributes and overlapped attributes rearranged (by both FS and AOM), and transformed dataset with both attributes reduced and outliners removed (FS+AOM+OE). The full performance results in terms of accuracy, tree/rule size and time taken are shown in Tables 3, 4 and 5.

TABLE III.

PERFORMANCE COMPARISON USING DIFFERENT CLASSIFIERS FOR LANGUAGE DATASET WITH FS TECHNIQUE ONLY.

	Accuracy %	Tree size or No. of rules	Time taken (Sec)
J48-ChiSq	86.7725	23	0.2
BestFirstTree-ChiSq	85.1852	21	0.58
FTree-ChiSq	91.0053	n/a	1.13
NBTree-ChiSq	88.8889	21	28.07
LogisticModelTrees-ChiSq	91.0053	n/a	11.77
RandomForest-ChiSq	95.2381	n/a	0.22
RandomTree-ChiSq	86.7725	93	0.02
REPTree-ChiSq	79.3651	19	0.17
ConjunctiveRule-ChiSq	51.8519	n/a	0.12
DecisionTable-ChiSq	82.5397	28	0.82
FuzzyUnorderedRuleInduction-ChiSq	85.7143	16	0.86
Ripper-ChiSq	86.7725	10	0.46
PART-ChiSq	82.5397	9	0.21
BayesNet-ChiSq	94.709	n/a	0.1
ComplementNaiveBayes-ChiSq	96.2963	n/a	0
NaiveBayes-ChiSq	96.8254	n/a	0.08
Bagging-ChiSq	84.6561	25	0.26
EnsembleSelection-ChiSq	84.127	n/a	0.8
RotationForest-ChiSq		out of memory	
SVM-ChiSq	87.8307	n/a	0.17
NeuralNetwork-ChiSq	93.1217	n/a	39.8

TABLE IV.

PERFORMANCE COMPARISON USING DIFFERENT CLASSIFIERS FOR LANGUAGE DATASET WITH FS AND AOM TECHNIQUES.

	Accuracy %	Tree size or No. of rules	Time taken (Sec)
J48-ChiSq	89.418	19	0.26
BestFirstTree-ChiSq	87.3016	25	0.5
FTree-ChiSq	92.0635	n/a	1.1
NBTree-ChiSq	90.4762	23	29.96
LogisticModelTrees-ChiSq	92.0635	n/a	8.31
RandomForest-ChiSq	95.7672	n/a	0.16
RandomTree-ChiSq	89.9471	83	0.03
REPTree-ChiSq	83.0688	17	0.15
ConjunctiveRule-ChiSq	51.8519	n/a	0.07
DecisionTable-ChiSq	87.3016	22	0.79
FuzzyUnorderedRuleInduction-ChiSq	87.8307	15	0.78
Ripper-ChiSq	86.7725	9	0.35
PART-ChiSq	87.8307	7	0.29
BayesNet-ChiSq	95.2381	n/a	0.09
ComplementNaiveBayes-ChiSq	96.8254	n/a	0.02
NaiveBayes-ChiSq	98.4127	n/a	0.07
Bagging-ChiSq	91.0053	17	0.22
EnsembleSelection-ChiSq	88.8889	n/a	0.63
RotationForest-ChiSq	out of memory	out of memory	out of memory
SVM-ChiSq	93.6508	n/a	0.12
NeuralNetwork-ChiSq	95.2381	n/a	29.24

TABLE V.

PERFORMANCE COMPARISON USING DIFFERENT CLASSIFIERS FOR LANGUAGE DATASET WITH FS+AOM+OE TECHNIQUES.

	Accuracy %	Tree size or No. of rules	Time taken (Sec)
J48-ChiSq	98.8764	19	0.17
BestFirstTree-ChiSq	97.191	19	0.43
FTree-ChiSq	100	n/a	0.89
NBTree-ChiSq	97.7528	29	30.39
LogisticModelTrees-ChiSq	98.3146	n/a	5.38
RandomForest-ChiSq	100	n/a	0.14
RandomTree-ChiSq	96.0674	51	0.01
REPTree-ChiSq	93.2584	15	0.12
ConjunctiveRule-ChiSq	55.0562	n/a	0.08
DecisionTable-ChiSq	95.5056	19	0.64
FuzzyUnorderedRuleInduction-ChiSq	91.573	15	0.55
Ripper-ChiSq	97.191	9	0.25
PART-ChiSq	98.8764	5	0.14
BayesNet-ChiSq	99.4382	n/a	0.1
ComplementNaiveBayes-ChiSq	100	n/a	0
NaiveBayes-ChiSq	100	n/a	0.07
Bagging-ChiSq	94.9438	17	0.18
EnsembleSelection-ChiSq	93.8202	n/a	0.55
RotationForest-ChiSq	out of memory	out of memory	out of memory
SVM-ChiSq	91.573	n/a	0.11
NeuralNetwork-ChiSq	100	n/a	22.67

The experiments are repeated with respect to accuracy only, but graphically showing the effects of applying no

technique at all, techniques that are responsible for reducing the attributes, and techniques that reduce both attributes and instances. The results are visually displayed at scattered plots in Figure 4 and Figure 5 for language dataset and emotion dataset respectively.

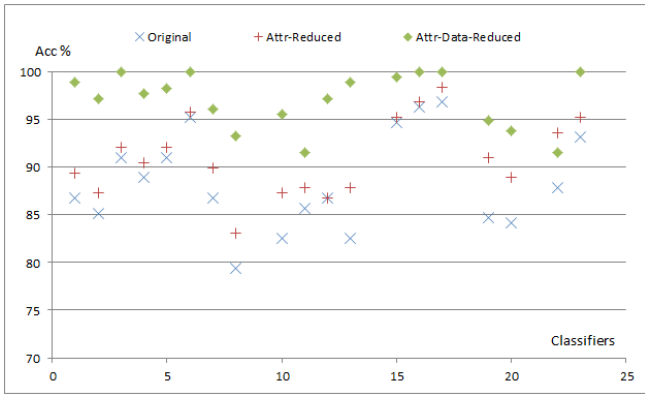

Figure 4. Accuracy graph of classifiers over the language dataset

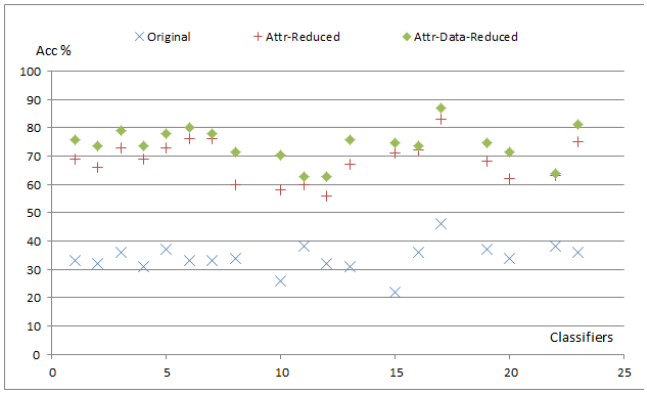

Figure 5. Accuracy graph of classifiers over the emotion dataset

From the charts when the accuracy value of various classifiers lay across, it can be observed that in general DR methods indeed yield certain improvement. The improvement gain between the original dataset without any technique applied and the inflected datasets with DR techniques is very apparent in the emotion dataset which represents a very large vector space. It means for text mining applications that deal with a wide coverage of vocabularies like online news it is very essential to apply DR techniques for maintaining the accuracy. In fact the gain ratio results from Table 6 shows a big leap of improvement gain between the no-DR-applied and DR-applied, for language dataset and emotion dataset - 3.584684% vs 101.3853% increases.

On a second note, the improvement gain between with and without outlier elimination is relatively higher for language dataset. 5.519077% > 3.584684%. That infers to the importance of removing outliers especially in a relatively small vector space.

Of all the classifiers being under test, Decision tree type and Bayes type outperform the rest. This phenomenon is observed consistently over different datasets and different DR techniques used. All the classification algorithms yield improvement and survive the model training with dataset of high dimensionality, except the Rotation Forest.

TABLE VI.

% PERFORMANCE GAIN – (L) LANGUAGE, (R) EMOTION

	Classifier	FS+AOM	FS+AOM+OE		Classifier	FS+AOM	FS+AOM+OE
Decision Tree	J48	3.048777	10.57774	Decision Tree	J48	109.0909	9.890145
	BFTree	2.484469	11.32786		BFTree	106.25	11.55515
	Ftree	1.162789	8.62068		Ftree	102.7778	8.384795
	NBTree	1.785712	8.042557		NBTree	122.5806	6.704928
	LMT	1.162789	6.789987		LMT	97.2973	6.879452
	RandForest	0.555555	4.419885		RandForest	130.303	5.552368
	RandTree	3.658532	6.804333		RandTree	130.303	2.660526
	REPTree	4.666661	12.26646		REPTree	76.47059	19.04767
		2.315661	*8.606187*			*109.3842*	*8.834379*
Rule	DecTable	5.769224	9.397308	Rule	DecTable	123.0769	21.2581
	FURI	2.469133	4.260811		FURI	57.89474	4.395667
	Ripper	0	12.00668		Ripper	75	11.8525
	PART	6.410249	12.57613		PART	116.129	13.17045
		3.662151	*9.560232*			*93.02517*	*12.66918*
Bayes	BayesNet	0.558659	4.410105	Bayes	BayesNet	222.7273	5.246901
	CompNB	0.54945	3.278685		CompNB	100	2.258889
	NB	1.639343	1.612902		NB	80.43478	4.594217
		0.915817	*3.100564*			*134.3874*	*4.033336*
Meta	Bagging	7.499991	4.32777	Meta	Bagging	83.78378	9.890147
	Ensemble	5.660371	5.547712		Ensemble	82.35294	15.20742
		6.580181	*4.937741*			*83.06836*	*12.54878*
Misc	SVM	6.626498	-2.21867	Misc	SVM	65.78947	1.16873
	NN	2.272725	4.999995		NN	108.3333	8.424933
		4.449612	*1.390664*			*87.0614*	*4.796832*
Average gain		**3.584684**	**5.519077**	Average gain		**101.3853**	**8.576502**

IV. CONCLUSION

Novel dimensionality reduction techniques for text mining namely Attribute Overlap Minimization and Outlier Elimination are introduced in this paper. The performance is tested in empirical experiments for verifying the advantage of the techniques. The results show that the techniques are effective especially on large vector space.

REFERENCES

[1] E. Leopold and Kindermann J. Text categorization with support vector machines: how to represent texts in input space? Machine Learning, **(2002)**, Vol.46, pp.423-444.

[2] X. Qi and B. Davison. Web Page Classification: Features and Algorithms. ACM Computing Surveys, **(2009),** Vol.41, No.2, pp.12-31.

[3] S. Fong, Measuring Emotions from Online News and Evaluating Public Models from Netizens' Comments: A Text Mining Approach. Journal of Emerging Technologies in Web Intelligence, **(2012)**, Vol.4, No.1, pp.60-66.

[4] P. Ponmuthuramalingam and T. Devi. Effective Dimension Reduction Techniques for Text Documents, International Journal of Computer Science and Network Security, **(2010)**, Vol.10, No.7, pp.101-109.

[5] Porter M.F. An Algorithm for Suffix Stripping. Program, **(1980)**, Vol.14, no.3, pp.130-137.

[6] Y. Yang and J. O. Pederson. A comparative study on feature selection in text categorization. In Proceedings of Fourteenth International Conference on Machine Learning, **(1997)**, pp.412-420.