ISSN (Print) : 2319-5940
ISSN (Online) : 2278-1021

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 2, Issue 12, December 2013

Effective Pre-Processing Activities in Text Mining using Improved Porter's Stemming Algorithm

C.Ramasubramanian[1], R.Ramya[2]

PG Student, ANNA UNIVERSITY, Nodal Center- Kamaraj College Of Engineering & Technology,

Virudhunagar, Tamilnadu, India[1]

Assistant Professor, Depatrment of IT, Kamaraj College Of Engineering & Technology,

Virudhunagar, Tamilnadu, India[2]

Abstract— Text Databases are rapidly growing due to the increasing amount of information available in various electronic forms. User need to access relevant information across multiple documents. Initial process in Text Mining system is Pre-Processing steps. Our approach to make an effective Pre-Processing steps to save both space and time requirements by using improved Stemming Algorithm. Stemming algorithms are used to transform the words in texts into their grammatical root form. Several algorithms exist with different techniques. The most widely used stemming algorithm is "M.F Porter stemming algorithm. However, it still has certain drawbacks of handling Named Entities. Our paper is to improve its structure by refining with certain constraints, so that improve the Information Retrieval System's Efficiency. Thus our paper is demonstrate how we can effectively overcome the problem of Named Entity during stemming process.

Keywords— Extraction, NamedEntity, Stemming, StopWordRemoval.

I. INTRODUCTION

Traditional Information retrieval techniques become inadequate for the increasingly vast amount of text data. A typical text mining problem is to locate relevant documents from a huge document collection. User need tools to compare different documents rank the importance and find patterns and trends across multiple documents. Hence Text mining plays a vital role in the Information retrieval systems. The main objective of pre-processing is to obtain the key features or key terms from stored text documents and to enhance the relevancy between word and document and the relevancy between word and category.

Pre-Processing step is crucial in determining the quality of the next stage, that is, the classification stage. It is important to select the significant keywords that carry the meaning and discard the words that do not contribute to distinguishing between the documents. The pre-processing phase of the study converts the original textual data in a data-mining-ready structure.

II. THE KNOWLEDGE DISCOVERY PROCESS

Data mining is one of the tasks in the process of knowledge discovery from the database. The data stored in the database is used to discover the patterns of data, which then interpreted by applying the domain knowledge. Following figure shows the process of Knowledge Discovery from Database.

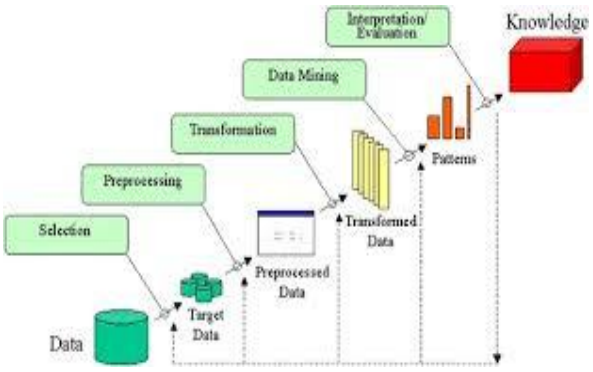

Fig.1 KDD process

III . PREPROCESSING STEPS

In this paper, we can discuss the two crucial step of preprocessing namely Stemming and Stop word Removal. Additional module "Spell Check" is added so as to overcome the problems of Named Entity faced during the stemming process. The overview of our system is depicted by the following figure.

ISSN (Print) : 2319-5940
ISSN (Online) : 2278-1021

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 2, Issue 12, December 2013

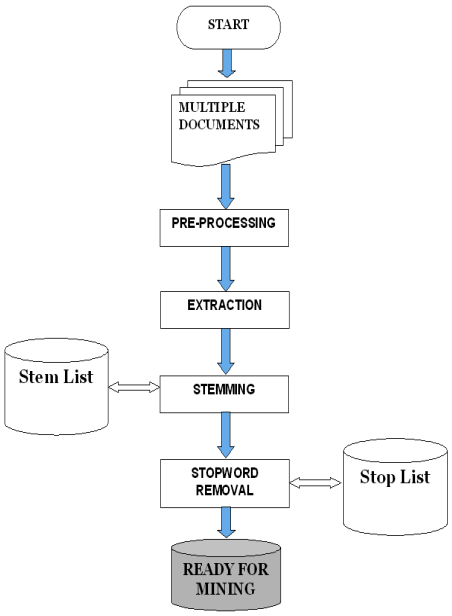

Fig.2 preprocessing task

3.1 Extraction

This method is used to tokenize the file content into individual word.

3.2 Stemming

This method is used to find out the root/stem of a word. for example, the words user, users, used, using all can be stemmed to the word "**USE**". The purpose of this method is to remove various suffixes, to reduce number of words, to have exactly matching stems, to save memory space and time. The stemming process is done using various algorithms. Most popularly used algorithm is "M.F. Porters Algorithm.

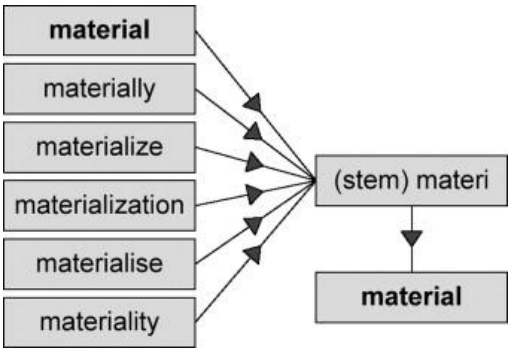

Fig.3 stemming process

After Stemming, it was observed that

Total Vocabulary = 10,000 words
Number of words not reduced = 3650
Resulting stems = 6370 distinct entries.

i) Drawback of existing stemming approach

1) Leads to Large Degree
2) Context dependent
3) Computer Storage space for endings
4) It is not possible to match something with nothing.
5) Objects are pigeon holed.
6) Not in all circumstances a suffix should be removed.
7) Evaluation of the worth of a suffix stripping system is very difficult.

ii) Process to overcome drawbacks of earlier stemming approach

To avoid the problems faced in the earlier stemming process, we have to prefer Spell-check utility. This module is used to avoid problems of named entity. Named entity problem is nothing but the name of the person, place and organization is unnecessarily stemmed and leave the data in meaningless state. Eg. Manipaul → Manip (Earlier M.F porters Algorithm)[4]
To avoid such mismatch, we have to prefer Spell-Check utility which is very efficient process. This spell-check utility is constructed using Boyer Moore's Algorithm which is an efficient string pattern checking algorithm. Using this algorithm, if any named entities found, skip that particular word from stemming process. Hence our stemmed data will not involve any wrong named entities.

Also, to avoid over stemming errors, we refine the existing stemming algorithm with certain constraints implemented from Improved Porters Algorithm [7]

iii) Advantage of spell check utility

The benefits of spell check utility is to overcome the drawbacks of existing stemming system such as

1) To avoid wrong matches than recoding
2) To overcome the lack of accuracy
3) To save time in matching of misspelled words

3.3) Stop word removal

Most frequently used words in English are useless in Text mining. Such words are called Stop words. Stop words are language specific functional words which carry no information. It may be of the following types such as pronouns, prepositions, conjunctions. Our system uses the SMART stop word list.[4]

ISSN (Print) : 2319-5940
ISSN (Online) : 2278-1021

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 2, Issue 12, December 2013

IV. WORK AND RESULT

Our system involves two crucial steps namely stemming and stop word removal. This process has been developed effectively to handle Named Entities problems. The observed techniques in our system are depicted by following figure.
Thus our experiments conducted had the following setups.

(i) STEMMING WITH SPELLCHECK + STOPWORD REMOVAL
(ii) IMPROVED PORTER'S STEMMING WITH SPELLCHECK + STOPWORD REMOVAL

From the Figure given below, it could be seen that the application of all the pre-processing techniques have a positive impact on the number of terms selected.

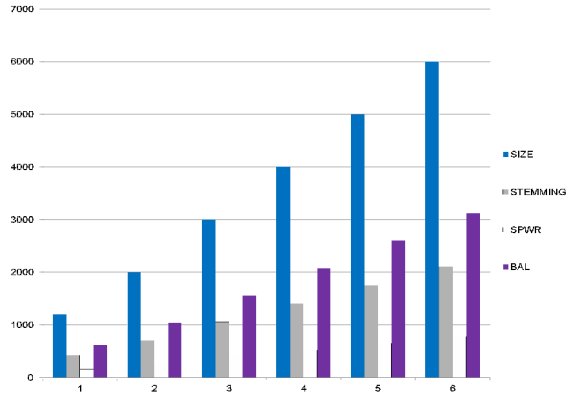

Fig.4 Effect of Preprocessing

The results further reveal an important fact that Improved Porter's stemming with the Spell-check utility increase the accuracy level of output content. Using Improved Porter's Algorithm [7], we refine certain constraints in the existing Porters Stemming Algorithm and improve the information retrieval process.

Using Existing Porters				Using Improved Porters			
SIZE	STMD	SPWR	BAL	UNIQ	STMD	SPWR	BAL
500	143	143	214	275	92	73	110
1200	343	343	514	660	220	176	264
2000	571	572	857	1100	367	293	440
3000	857	857	1286	1650	550	440	660
4000	1143	1143	1714	2200	733	587	880
5000	1429	1428	2143	2750	917	733	1100
6000	1714	1714	2572	3300	1100	880	1320

Fig.5 Performance Comparison

V. CONCLUSION

Thus pre-processing activities plays a vital role in the various applications. Therefore it is concluded that the domain specific applications are more proper for text mining. The present work uses three important pre-processing techniques namely stop word removal, stemming and spell check. From our approach, we observed that, it was an efficient approach to design and develop a Stemming algorithm which works dynamically for any domain. Designing a domain specific system is also a challenging task. Thus our approach is designed effectively to overcome the problem of Named Entity. Also, it is necessary to cover various subject disciplines.

REFERENCES

[1] Ning Zhong, Yuefeng Li, and Sheng-Tang Wu, "Effective Pattern Discovery for Text Mining", IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, VOL. 24, NO. 1, JANUARY 2012.
[2] V. Srividhya, R. Anitha, "Evaluating Preprocessing Techniques in Text Categorization - International Journal of Computer Science and Application" Issue 2010.
[3] Xue, X. and Zhou, Z. (2009) "Distributional Features for Text Categorization", IEEE Transactions on Knowledge and Data Engineering, Vol.21, No. 3, Pp. 428-442.
[4] M.F. Porter, "An Algorithm for Suffix Stripping " Program, vol. 14, no. 3, pp. 130-137, 1980.
[5] Salton, G. and Buckley, C. (1988) "Term weighting approaches in Automatic text retrieval, Information Processing and Management", Vol. 24, No.5, Pp. 513-523.
[6] Karbasi, S. and Boughanem, M. (2006) "Document length normalization using effective level of term frequency in large collections, Advances in Information Retrieval, Lecture Notes in Computer Science", Springer Berlin / Heidelberg, Vol. 3936/2006,Pp.72-83.
[7] Fadi Yamout, "Further Enhancement to the Porter's Stemming Algorithm", Issue 2006
[8] Diao, Q. and Diao, H. (2000) "Three Term Weighting and Classification Algorithms in Text Automatic Classification", The Fourth International Conference on High-Performance Computing in theAsia-Pacific Region, Vol. 2, P.629.
9. Website: "http://www-igm.univ-mlv.fr/~lecroq/string".

BIOGRAPHIES

C.RAMASUBRAMANIAN - Received the Master Degree in Computer Applications fromAlagappa University of India in 2008 and the M.L.I.Sc.,. degree in Madurai Kamaraj University of India in 2007. Currently he is doing Master of Engineering in Computer Science at Anna University of India. His research interests include Data mining and its Web Applications.

R. RAMYA - Received the Bachelor of Engineering in Computer Science from Anna University of India in 2002 and the Master of Engineering in Computer Science from Anna University of India in She is working as an Assistant Professor in Kamaraj College of Engineering and Technology, Tamilnadu, India. Her research interests include Digital Image Processing and Mining Applications.