Identification of Legitimate Words for Text pre-processing Using K-Means Clustering Algorithm

B. Sarumathi

Research Scholar, Department of Computer Science, Quaid-e-Millath Govt. College for women (Autonomous),Chennai-600 002, Tamilnadu, India

R. Jayanthi

Assistant Professor, Department of Computer Science, Quaid-e-Millath Govt. College for women(Autonomous),Chennai-600 002, Tamilnadu, India

Abstract

World Wide Web brought us to a stage of enormous and ever growing amounts of data and information. It impacts almost all aspects of people's lives. In addition, with the ample data provided by the web, it has become a vital resource for research. The search engines are becoming more and more sophisticated trying to cover user's demands to access specific information. Stemming plays a dominant role in the contributions to the field of Information Retrieval. Porter algorithm is extensively used for stemming process here we suggest an approach that effectively improve the performance of porter algorithm by identifying the stemmed output of it is a real word or not.

***Keywords**: Stemming, Information Retrieval, Web Mining*

1. Introduction

Based on several research studies we can typically organize web mining into three domains: 1.Content 2.Structure 3.Usage. This paper mainly concentrates on Web content mining and is the process of extracting knowledge from the content of the actual web documents (text content, multimedia etc.). Web structure mining is targeting useful knowledge from the web structure, hyperlink references and so on. Web usage mining attempts to discover useful knowledge from the secondary data obtained from the interactions of the users with the web.

The term stemming is a pre-processing step in Text Mining applications and Information Retrieval systems. The goal of stemming is to reduce inflectional forms and sometimes derivationally related forms of a word to a common base form [1]. It is a common pre-processing function in text analysis. It's the process of stripping a word of inflection and reducing it to a stem. Stemming algorithm is used as an added value to enhance the search results [2].

2. Literature Survey

Stemming is a fundamental step in processing textual data preceding the tasks of information retrieval, text mining, and natural language processing. The common goal of stemming is to standardize words by reducing a word to its base. However, simply removing the suffix of the word can cause stemming errors such as under-stemming or over-stemming. Sophisticated stemmers tend to weakly stem documents with very computationally expensive approaches such as dictionary lookup [3].

Stemming algorithms relate morphologically similar indexing and search terms. It is used to improve retrieval effectiveness and to reduce the size of indexing files. Several approaches to stemming are described--table lookup, affix removal, successor variety, and n-gram [4]. The use of stemming and morphological analysis in Information Retrieval has been broadly found to have two different motivations

i. Improving precision and Recall rates

ii. Understanding the structure of vocabulary of a language and building knowledge bases [7].

The Porter algorithm is massively used for stemming. It consists of a set of condition/action rules. The conditions fall into three classes: conditions on the stem, conditions on the suffix, and conditions on the rules [4]. A dictionary look-up can help in reducing the errors and converting stems to words [1]. This Stemmer has been used in various studies. It is most widely used in Information Retrieval research. Implementations of this stemmer are also available at a website established by Porter himself. Although the extraction of some words can be resolved with this algorithm, there are some words that give a different meaning after a stemming process such as the word opening that stemmed to open [2].

3. Existing Algorithm

Porters stemming algorithm is as of now one of the most popular stemming methods proposed in 1980. Many modifications and enhancements have been done and suggested on the basic algorithm. It is based on the idea that the suffixes in the English language (approximately 1200) are mostly made up of a combination of smaller and simpler suffixes.

3.1 Advantages

1) Produces the best output as compared to other stemmers.

2) Less error rate.

3) Compared to Lovins it’s a light stemmer.

4) The Snowball stemmer framework designed by Porter is language independent approach to stemming.

3.2 Limitations

1) The stems produced are not always real words.

2) It has at least five steps and sixty rules and hence time consuming.

The results of stemming usage in information retrieval are inconsistent [14].Among the stemming algorithms porter stemmer is widely accepted and used but the major limitation of porter is the stem produced are not always real[1] to find the key solution for this problem the proposed method is designed.

4. Proposed Method

This framework is designed to suggest an approach that effectively improves the performance of porter algorithm by identifying the stemmed output of it is a real word or not. This is done by comparing the stemmed output with the dictionary lookup which holds the real words with added advantage the dictionary used here is a self-learning system this eliminates the manual workload for updating the dictionary.

In order to improve the reliability, this methodology incorporates an idea by appending the user with a list of most relevant words this helps the user for getting the legitimate words as a result. Getting feedback from user is updated with the classification table this is used for re-clustering which in turn sets the priority of relevant words displayed for next time.

4.1 Algorithm

Step 1: Input the Query String as Text Document.

Step 2: Get the Classification table as pre-trained Textual database.

Step 3: Apply stemming process using porter’s algorithm and store the result.

Step 4: Apply parts of the speech identification Tagger and identify Noun, Verb, Adjective and store the result in Dictionary.

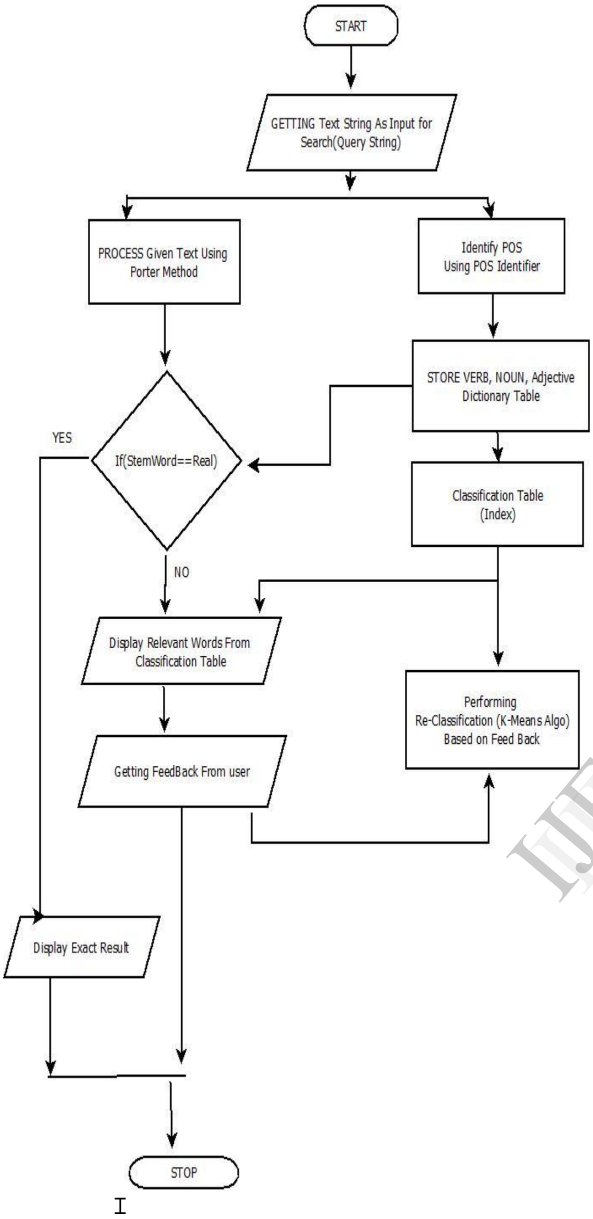
Step 5: Compare the Stemming result with the Dictionary.

Step 6: If the Condition is true then display the Real word otherwise display the Relevant word.

Step 7: Get the feedback from the user and store the result.

Step 8: Include the updated result with the Classification table.

4.2 Flow chart


START
GETTING Text String As Input for Search(Query String)
PROCESS Given Text Using Porter Method
Identify POS Using POS Identifier
STORE VERB, NOUN, Adjective Dictionary Table
YES
If(StemWord==Real)
Classification Table (Index)
NO
Display Relevant Words From Classification Table
Performing Re-Classification (K-Means Algo) Based on Feed Back
Getting FeedBack From user
Display Exact Result
STOP
I

4.3 Working Methodology

The Text document is given as input to both porter's method and POS tagger. The text is stemmed by porter algorithm and POS tagger splits the text into parts of the speech which makes sense that it is real word. The identified words are taken to the dictionary which stores the Verb, Noun and adjective. The output of the porter algorithm is compared with the dictionary to check whether the stemmed word is a real word or not. With the reference of comparison, if the stemmed word is in Dictionary then it is proved that it is a real word and then the text is displayed. If the stemmed word is not in the dictionary then the list of most relevant words will be displayed. The dictionary used here parses to the classification table which already holds some pre-trained data. The words in the classification table are classified for indexing.

The user can give feedback by selecting the appropriate word from the list. The feedback of the user is taken into account for re-clustering the data thus the priority of words of most relevant changes accordingly thus the methodology works out.

4.4 Working Example

INPUT: The Indian early medieval age is defined by regional kingdoms and cultural diversity. Earliest civilizations that arose in the world developed in the late fourth millennium had communism. The effects of global warming in community are the ecological and social changes caused by the rise in global temperatures aged people are helpers. In interview communication is important.

OUTPUT OF PORTER: The Indian earli mediev ag is defin by region kingdom and cultur divers Earliest civil that aros in the world develop in the late fourth millennium it had commun The effect of global warn in commun ar the ecolog and social chang caus by the rise in global temperatur ag peopl ar helper. In interview commun is import.

OUTPUT OF POS TAGGER: The_DT Indian_JJ early_JJ medieval_JJ age_NN is_VBZ defined_VBN by_IN regional_JJ kingdoms_NNS and_CC cultural_JJ diversity_NN Earliest_JJS civilizations_NNS that_WDT arose_VBD in_IN the_DT world_NN developed_VBD in_IN the_DT late_JJ fourth_JJ millennium_JJ communism NN. The_DT effects_NNS of_IN global_JJ warning_NN are_VBP the_DT ecological_JJ and_CC social_JJ changes_NNS caused_VBN by_IN the_DT rise_NN in_IN global_JJ warming NN community NN temperatures_NNS aged_JJ people_NNS are_VBP helpers_NNS In NN interview NN communication NN is VBZ important JJ

DICTIONARY		
VERB (VB)	**NOUN(NN)**	**ADJECTIVE (JJ)**
Developing	Evolution	Recent
Has	Community	Up
Strengthened	Organizations	Effective
Are	Countries	Local
Addressing	View	Larger
Is	Organizations	Charitable
Defined	Communism	Indian
Arose	Age	Rarely
Developed	Kingdoms	Medieval
Caused	Diversity	Regional
Develop	Civilizations	Cultural
	World	Earliest
	Effects	Late
	Warming	Fourth
	Changes	Millennium
	Rise	Ecological
	Temperature	Social
	People	Aged
	Effect	Helpers
	Ill	Serious
	Aids	
	Program	
	interview	
	Communication	

Table 1: Dictionary table

Table 1 shows the stored data of the dictionary which holds verb, noun and adjectives**.** The output of porter's algorithm has some real words (Indian,world and etc) these words will exist in the dictionary and so it will be displayed as such and the error prone output holds some unreal words such as (earli,ag,commun.) these words will get the suggestion list.

CLUSTERING		
Age	Earliest	Community
Aged	Early	Communism
		Communication

1

The table 2 shows the suggestion list given to the user for user's feedback by selecting most relevant word.

4.5. Comparison

Output of Porter Algorithm	**Output of Suggested Approach**
Indian, **earli**, mediev, ag , defin , region kingdom, culture, divers, Earliest, civil aros, world develop late ,fourth millennium it had **commun**, effect , global,warn , commun, ar t ecology, social chang, caus, rise global temperature, **ag** ,people, ar helper. interview commun, import.	Age,aged,earliest,early,community,communism,communication

Table 3: Comparision table

Though the output of porter algorithm give stemmed result it also has some error words (earli,ag,commun)whereas the new approach has a output with a suggestion list of real words to overcome the drawback of porter algorithm.

5. Conclusion

With the onset of advancement in Information Technology, information is almost doubling in every year. In order to retrieve information from this information explosion, Stemming is an effective system with many stemming algorithms. The approach presented proposes a new framework for effective retrieval performance with the help of Porter algorithm is improved in terms of recall and falls back in precision in terms overcoming its major drawback non-word errors. Moreover, here the framework uses k – means clustering for information retrieval process. There are some possible extensions to this work with artificial intelligence which makes more sense to its efficiency.

6. Future work

Future research can be done on other languages. This paper concentrates on English language alone. This can be implemented for various languages depending upon their own grammar. It can be further enhanced with neural network technology since its performance can be highly automated and minimizes human involvement.

7. References

1. Anjali Ganesh Jivani "A Comparative Study of Stemming Algorithms", Int J. Comp. Tech. Appl., Vol 2 (6), 1930-1938.

2. Noraida Haji Ali " Porter Stemming Algorithm for Semantic Checking"

3. Eiman Tamah Al-Shammari "Towards an error- free stemming" ISBN: 978-972-8924-63-8 © 2008 IADIS.

4. W.B.Frakes- Stemming Algorithms Software Engineering Guild, Sterling, VA 22170.

5. http://en.wikipedia.org/wiki/Stemming.

6. Edie Rasmussen "Information retrieval" chapter 16- clustering Algorithms".

7. Rashmi S, Kratika Singh, Ajay Dhamelia- "Stemming and Morphological Analysis "

8. Giridhar N S, Prema K.V, Professor, N .V Subba Reddy, "A Prospective Study of Stemming Algorithms for Web Text Mining".

9. J. B. Lovins, "Development of a stemming algorithm," Mechanical Translation and Compute Linguistic., vol.11, no.1/2, pp. 22-31, 1968.

10. Llia Smirnov DePaul University "overview of stemming algorithms"

11.http://stackoverflow.com/questions/190775/stemming-algorithm-that- Produces-real-words.

12. Adapting a WSJ trained Part-of-Speech tagger to Noisy Phani Gadde Language Technologies Research Centre IIIT-Hyderabad, India

13. Raymond J. Mooney and Razvan Bunescu University of Texas at -"Mining Knowledge from Text Using Information Extraction"

14. Fadillah Z Tala "A Study of Stemming Effects on Information Retrieval in Bahasa Indonesia"

15. Christopher D. Manning " Part-of-Speech Tagging from 97% to 100%:Is It Time for Some Linguistics?", Departments of Linguistics and Computer Science, Stanford University.

16. Hull, D. (1996). Stemming algorithms: A case study for detailed evaluation. Journal of the American Society for Information Science, 47(1), 70-84.

17. The Porter Stemming Algorithm: Then and Now - White Rose,eprints.whiterose.ac.uk,1434/01willettp9_PorterStemmingReview.pdf

18. Hybrid Approach for Stemming in Punjabi - International Journal of Computer Science and Computer Network,

19. Popovic,ˇ M. and Willett, P. (1992). The effectiveness of stemming for natural-language access to slovene textual data. Journal of the American Society for Information Science, 43(5):384–390.

20. Porter, M. F. (1980). An algorithm for suffix stripping. Program, 14(3):130–137.