Refined Porter Stemmer Algorithm for Enhanced Stemming in Information Retrieval Systems

Juan Pablo Martínez García[1], Sofía Delgado Álvarez[2]

[1]Departamento de Ingeniería Civil, Universidad Nacional Autónoma de México, Mexico City, Mexico
[2]Centro de Estudios de Ingeniería Ambiental, Universidad de Guadalajara, Jalisco, Mexico.

ABSTRACT

In the era of digitalization, information retrieval (IR) are retrieves and ranks documents from large collections according to users search queries, has been usually applied in the several domains. Building records using electronic and searching literature for topics of interest are some IR use cases. For the moment, Natural Language Processing (NLP), such as tokenization, stop word removal and stemming or Part-Of-Speech (POS) tagging, has been developed for processing documents or literature. This study offer that NLP can be incorporated into IR to strengthen the conventional IR models. In this paper proposed Enhanced Porter Stemmer algorithm for improving the efficiency of pre-processing in text mining. The Enhanced Porter Stemmer algorithm is extension version of new porter stemmer. The Enhanced Porter Stemmer algorithm performance is compared with several algorithms such as porter, new porter and etc. The performance of the Enhanced Porter Stemmer is better than others.

Keywords: Text Mining, Pre-processing, Stemming Techniques, Enhanced Porter Stemmer Algorithm, Porter Stemming.

I. INTRODUCTION

Information Retrieval (IR) requires several preprocessing steps for structuring the text and extracting features, including tokenization, word segmentation, Part of Speech (POS) tagging, parsing [1]. Now we give a brief overview of these techniques. Tokenization is a fundamental technique for most NLP tasks, which is splits a sentence or document into tokens which are words or phrases. For English, it is trivial to split words by the spaces, but some additional knowledge should be taken into consideration, such as opinion phrases, named entities. In tokenization, some stop words, such as “the”, “a”, will be removed as these words provide little useful information. Information Retrieval is essentially a matter of deciding which documents in a collection should be retrieved to satisfy user’s need of information. Conflation is the process of merging or lumping together non identical words which refer to the same principal concept. 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. Data mining techniques are very useful to manipulating and analyzing data from database. They are several techniques are available in data mining for analyzing data such as clustering, classification, decision tree, neural network and genetic algorithm. Among all these types of data [2], particularly data mining supports text data for representing the document. A document consists of collection of words which includes stop words. Many words used in the text are morphological variants which based from the root form e.g. connection /connect, combining /combine, preferences /preferred/prefer. Text mining is a new and exciting research area that tries to solve the information overload problem by using techniques from data mining, machine learning, natural language processing (NLP), information retrieval (IR), and knowledge management.

Text classification [3] plays vital role in text mining. Currently, handling textual documents is a great challenge. Text classification uses several key classification algorithms, e.g., decision trees, pattern (rule)-based classifiers, support vector machines, naïve Bayesian classifier and artificial neural networks. B.Ramesh et al. [4] elaborately discussed several key pre-processing techniques. Text classification applied in many fields. Biological genetic algorithm for instance selection of text classification in medical field [5]. Brajendra et al. [6] analyzed several recent stemming techniques and provided several directions. Ruba et al. [7] discussed several stemming methods and their constraints. Ramesh et al. [8] discussed several instance selection methods in pre-processing. Instance selection is an another solution of text pre-processing.

II. LITERATURE REVIEW

The enormous amount of information stored in unstructured texts cannot simply be used for further processing by computers, which typically handle text as simple sequences of character strings. Therefore, specific preprocessing methods and algorithms are required in order to extract useful patterns. Text mining refers generally to the process of extracting interesting information and knowledge from unstructured text. Text

mining is the process of discovering information in text documents. Stefano et al. [9] discussed automatic learning methods of linguistic resources for stop words removal. Wahiba et al. [10] proposed new stemmer for rectifying the limitations of porter stemmer algorithm. The new stemmer contains four classes and each class contains several morphological conditions. Ruban et al.[11] discussed various methods of affix removal stemmer. They are analyzed merits and demerits of affix removal stemmers.

Sandeep et al. [12] analyzed strength of affix removal stemmers. Also, they are discussed comparative analysis of affix removal stemming algorithm accuracies. Giridhar et al. [13] conducted a prospective study of stemming techniques in web documents. Tomas et al. [14] is explained High Precision Stemmer(HPS) methods. HPS is difficult to decide a threshold for creating clusters and requires significant computing power. Venkat sudhakarareddy et al. [15] discussed stemming techniques applied to information extraction using RDBMS.

III. PROPOSED WORK

Enhanced Porter Stemming algorithm reduce different morphological modifications to their fundamental rules. Stemming is used to enable matching of queries and documents in keyword-based IR systems. This assumes that morphological variants of words have similar semantic interpretations and can be considered as equivalent for the purpose of IR applications. Fig.1. shows the overview of proposed system. The enhanced porter stemmer rectifies the drawbacks of porter algorithm.

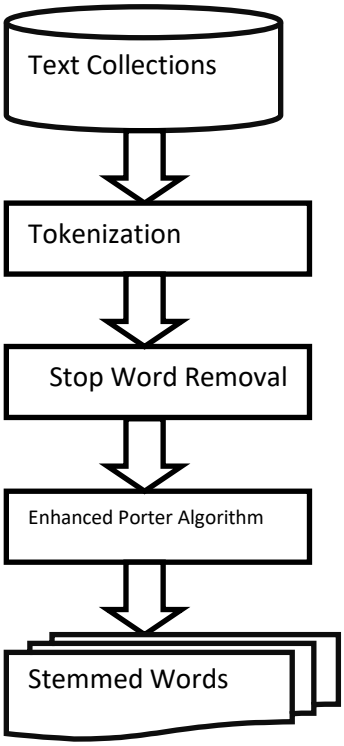

Fig.1. overview of proposed system.

Enhanced Porter Algorithm Description

In order to improve Porter stemmer, we studied the English morphology, and used its characteristics for building the enhanced stemmer. The resultant stemmer to which we will refer as Enhanced Porter Stemmer algorithm includes four steps. Fig.2 presents Enhanced Porter Stemmer algorithm.

Recode plurals and simple present
Recode double suffix and simple suffix and remaining suffices
Recode stem with e & I and Irregular verbs
Stem Retrieved from Table (SRT)

Fig.2. Rule engine to Enhanced Porter Stemmer

In this section, thesis describes step by step method of our Enhanced Porter Stemming algorithm as follows:

Step 1 Initialization:
Input the text collection.
Step 2 Select relevant text ending:
Examine the final letters of the text collection;
Consider the first rule in the relevant ending for the input query term, and to indicate the first query term among stemming candidates.
Step 3 Check applicability of rule:
If the final letters of the query term do not match the ending rule, output stem, then terminate;
if the final letters of the query term and the ending rule matches, then goto 4;
if the final letters of the query term and the rule matches, and matching ending acceptability conditions are not satisfied, then goto 5;
Step 4 Apply rule:
Delete from the right end of the token the number of characters specified by the remove rules;
if there is an add string, then add it to the end of the query term specified by the rule;
if there is a replace string, then replace the number specified to the end of the query term;
if the condition specified is "no applicable rule" output the stem, then terminate;
if the condition specified is "match ending found" then take output to the next rule to access;
Otherwise goto 2.
Step 5 Search for another rule:
Go to the next rule in the rule engine database;
if the endings of the query term has changed, output stem, then terminate;
Otherwise goto 3.
Step 6 SRT Condition:
If matching endings acceptability conditions are not satisfied, or stem word is meaningless, then correct stemmed word retrieved from table.
Step 7 Termination Condition:
If matching endings acceptability conditions are satisfied, and then terminate the stemming process.

IV. EXPERIMENTAL RESULTS

To evaluate the performance of the Enhanced Porter stemmer described in this study, we have applied these algorithms to the sample vocabulary downloaded from the web site http://snowball.tartarus.org/algorithms/english /voc.txt. It contains distinct words, arranged into “conflation groups”. Some of them are incorrect words. The new porter stemmer is developed using HTML and PHP with java script. For example, there are 155 incorrect words in the sample of 500 words which begin with alphabet ‘b’. For the comparison, we apply the information retrieval method, first without using a stemmer, second using the original Porter stemmer, new porter stemmer and finally using the enhanced Porter Stemmer. We used the two well-known metrics, namely recall and precision, calculated as follows:

$$Precision = \frac{\text{correct} \cap \text{retrieved}}{\text{retrieved}}$$

$$Recall = \frac{\text{correct} \cap \text{retrieved}}{\text{correct}}$$

A good IR system should retrieve several relevant documents, and it should retrieve few non-relevant documents. Since the stem of a term represents a larger concept than the original term, the stemming process increases the number of retrieved documents. When the document retrieval system uses the new porter stemmer we perceive an improvement in retrieval effectiveness compared to the original Porter stemmer. Thus the precision and recall variant words of same group to correct stem is better in enhanced stemmer algorithm than the earlier stemmers.

Table 1. Comparison of Algorithm Results

Analysis of Stemmers	Precision	Recall
Without Stemmer	0.661	0.671
With Porter Stemmer	0.732	0.775
With new Porter Stemmer	0.852	0.884
With Enhanced Stemmer	0.891	0.901

Precision

Precision obtains 0.891 by with enhanced stemmer, this precision is better than comparing with others. Fig.2. shows comparison of precision. . The enhanced stemmer performance is better than another existing stemming techniques.

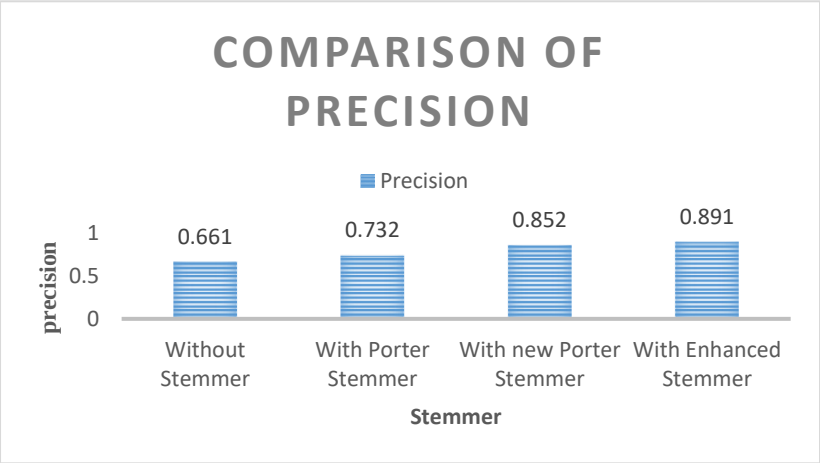

Fig.2. Comparison of Precision.

Recall

Recall obtains 0.901 by with enhanced stemmer, this recall is better than comparing with others. Fig.2. shows comparison of recall. . The enhanced stemmer performance is better than another existing stemming techniques.

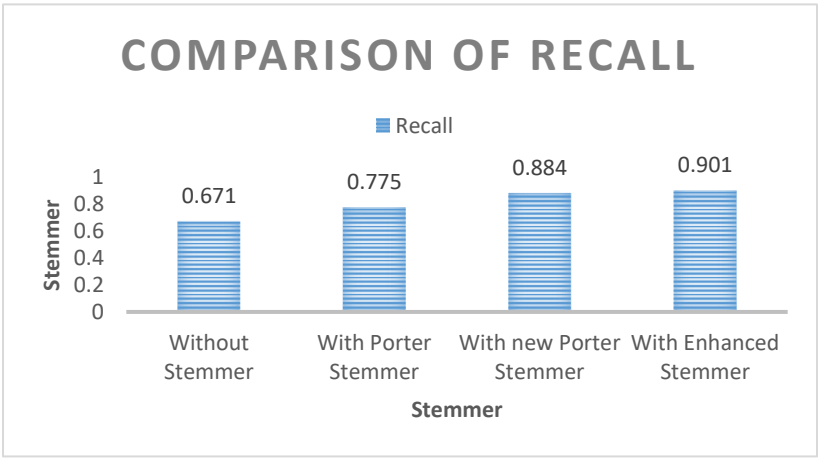

Fig.3. Comparison of Recall

V. CONCLUSION

Stemming can be effectively used in natural language processing. The benefit of stemming algorithm in mining will reduce the database size. Stemming techniques are useful for library and information science in the fields of classification and indexing, as it makes the operation less dependent on particular forms of words. The Enhanced Porter stemmer obtains 0.891 precision, 0.901 recall. The Enhanced Porter stemmer is produced less error rate and more conflated words. The performance of Enhanced Porter stemmer is better than another existing stemmers. In future, to reduce the time consumption of Enhanced Porter Stemmer and decrease the utilization of memory storage

VI. REFERENCES

[1] Shiliang Sun, Chen Luo, Junyu Chen, "A Review of Natural Language Processing Techniques for Opinion Mining Systems", Information Fusion, Elsevier, 2016.

[2] R. Sagayam, S.Srinivasan and S. Roshni., " A Survey of Text Mining: Retrieval, Extraction and Indexing Techniques", International Journal of Computational Engineering Research (ijceronline.com) Vol. 2 Issue. 5, September 2012.

[3] Francesco Colace, Massimo De Santo, Luca Greco and Paolo Napoletano, "Text classification using a few labeled examples", Computer in Human Behavior 30(2014)689-697, Elsevier, 2013.

[4] B.Ramesh, J.G.R.Sathiaseelan, "A Theoretical Stuudy on Advanced Techniques in Pre-Processing and Text Classification" International Journal of Data Mining and Emerging Technologies, Vol.5, No.1, pp.6-10, 2015.

[5] AlperKursatUysal and SerkanGunal, "The impact of preprocessing on text classification", Information Processing and Management 50(2014) 104-112, Elsevier, 2013.

[6] Brajendra Singh Rajput, NilayKhare, A survey of Stemming Algorithms for Information Retrieval,*IOSR Journal of Computer Engineering, Vol.17, Issue.3, pp. 76-80, 2015.*

[7] S.P.Ruba Rani, B.Ramesh, M.Anusha, and J.G.R.Sathiaseelan, "Evaluation of Stemming Techniques for Text Classification" International Journal of Computer Science and Mobile Computing, *Vol. 4, Issue. 3, pg.165 – 171 2015.*

[8] B.Ramesh, J.G.R.Sathiaseelan, "An Analysis of Instance Selection Algorithms Using Support Vector Machine for Text Classification" International Journal of Modern Computer Science, Vol.3, Issue.2, pp.81-84, 2015.

[9] Stefano Ferilli, Floriana Esposito and Domenico Grieco, "Automatic Learning of Lingustic Resources for Stopword Removal and Stemming from Text", Procedia Computer Science 38 (2014) 116-123, Elsevier, 2014.

[10] Wahiba Ben AbdessalemKaraa,"A new stemmer to improve information retrieval", International Journal of Network Security & Its Applications, July 2013.

[11] Rupan Gupta and Anjali Ganesh Jivani, "Empirical Analysis of Affix Removal Stemmers", IJCTA, March- April 2014.

[12] Sandeep R.Sirsat, Vinay Chavan and Hemant S.Mahalle, "Strength and Accuracy Analysis of Affix Removal Stemming Algotithms", International Journal of Computer Science and Information Technologies, Vol. 4(2), 2013, 265-269.

[13] GiridharN.S,Prema K.V and N.V SubbaReddy,"A Prospective Study of Stemming Algorithms for Web Text Mining", Ganpat University Journal of Engineering & Technology,Vol-1,Issue-1,Jan-Jun-2011.

[14] Tomáš Brychcín, Miloslav Konopík. "HPS: High precision stemmer". Information Processing and Management 51 pp.68–91, 2015.

[15] VenkatSudhakaraReddy.Ch and Hemavathi.D, "Information extraction using RDBMS and stemming algorithm", International Journal of Science and Research (IJSR), April 2014.