Preprocessing Techniques for Text Mining

Dr.S.Kannan,

Associate Professor,

Department of Computer Applications,

Madurai Kamaraj University.

skannanmku@gmail.com

Vairaprakash Gurusamy,

Research Scholar,

Department of Computer Applications,

Madurai Kamaraj University.

vairaprakashmca@gmail.com

Abstract

Preprocessing is an important task and critical step in Text mining, Natural Language Processing (NLP) and information retrieval (IR). In the area of Text Mining, data preprocessing used for extracting interesting and non-trivial and knowledge from unstructured text data. Information Retrieval (IR) is essentially a matter of deciding which documents in a collection should be retrieved to satisfy a user's need for information. The user's need for information is represented by a query or profile, and contains one or more search terms, plus some additional information such as weight of the words. Hence, the retrieval decision is made by comparing the terms of the query with the index terms (important words or phrases) appearing in the document itself. The decision may be binary (retrieve/reject), or it may involve estimating the degree of relevance that the document has to query. Unfortunately, the words that appear in documents and in queries often have many structural variants. So before the information retrieval from the documents, the data preprocessing techniques are applied on the target data set to reduce the size of the data set which will increase the effectiveness of IR System The objective of this study is to analyze the issues of preprocessing methods such as Tokenization, Stop word removal and Stemming for the text documents

Keywords: Text Mining, NLP, IR, Stemming

I. Introduction

Text pre-processing is an essential part of any NLP system, since the characters, words, and sentences identified at this stage are the fundamental units passed to all further processing stages, from analysis and tagging components, such as morphological analyzers and part-of-speech taggers, through applications, such as information retrieval and machine translation systems. It is a Collection of activities in which Text Documents are pre-processed. Because the text data often contains some special formats like number formats, date formats and the most common words that unlikely to help Text mining such as prepositions, articles, and pro-nouns can be eliminated

Need of Text Preprocessing in NLP System

1. To reduce indexing(or data) file size of the Text documents
   i) Stop words accounts 20-30% of total word counts in a particular text documents
   ii) Stemming may reduce indexing size as much as 40-50%
2. To improve the efficiency and effectiveness of the IR system
   i) Stop words are not useful for searching or Text mining and they may confuse the retrieval system
   ii) Stemming used for matching the similar words in a text document

II. Tokenization

Tokenization is the process of breaking a stream of text into words, phrases, symbols, or other meaningful elements called tokens .The aim of the tokenization is the exploration of the words in a sentence. The list of tokens becomes input for further processing such as parsing or text mining. Tokenization is useful both in linguistics (where it is a form of text segmentation), and in computer science, where it forms part of lexical analysis. Textual data is only a block of characters at the beginning. All processes in information retrieval require the words of the data set. Hence, the requirement for a parser is a tokenization of documents. This may sound trivial as the text is already stored in machine-readable formats. Nevertheless, some problems are still left, like the removal of punctuation marks. Other characters like brackets, hyphens, etc require processing as

well. Furthermore, tokenizer can cater for consistency in the documents. The main use of tokenization is identifying the meaningful keywords. The inconsistency can be different number and time formats. Another problem are abbreviations and acronyms which have to be transformed into a standard form.

Challenges in Tokenization

Challenges in tokenization depend on the type of language. Languages such as English and French are referred to as space-delimited as most of the words are separated from each other by white spaces. Languages such as Chinese and Thai are referred to as unsegmented as words do not have clear boundaries. Tokenizing unsegmented language sentences requires additional lexical and morphological information. Tokenization is also affected by writing system and the typographical structure of the words. Structure of languages can be grouped into three categories:

Isolating: Words do not divide into smaller units. Example: Mandarin Chinese

Agglutinative: Words divide into smaller units. Example: Japanese, Tamil

Inflectional: Boundaries between morphemes are not clear and ambiguous in terms of grammatical meaning. Example: Latin.

III. Stop Word Removal

Many words in documents recur very frequently but are essentially meaningless as they are used to join words together in a sentence. It is commonly understood that stop words do not contribute to the context or content of textual documents. Due to their high frequency of occurrence, their presence in text mining presents an obstacle in understanding the content of the documents.

Stop words are very frequently used common words like ‘and’, ‘are’, ‘this’ etc. They are not useful in classification of documents. So they must be removed. However, the development of such stop words list is difficult and inconsistent between textual sources. This process also reduces the text data and improves the system performance. Every text document deals with these words which are not necessary for text mining applications.

IV. Stemming

Stemming is the process of conflating the variant forms of a word into a common representation, the stem. For example, the words: "presentation", "presented", "presenting" could all be reduced to a common representation "present". This is a widely used procedure in text processing for information retrieval (IR) based on the assumption that posing a query with the term presenting implies an interest in documents containing the words presentation and presented.

Errors in Stemming

There are mainly two errors in stemming.

1. over stemming
2. under stemming

Over-stemming is when two words with different stems are stemmed to the same root. This is also known as a false positive.

Under-stemming is when two words that should be stemmed to the same root are not. This is also known as a false negative.

TYPES OF STEMMING ALGORITHMS

i) Table Look Up Approach

One method to do stemming is to store a table of all index terms and their stems. Terms from the queries and indexes could then be stemmed via lookup table, using b-trees or hash tables. Such lookups are very fast, but there are problems with this approach. First there is no such data for English, even if there were they may not be represented because they are domain specific and require some other stemming methods. Second issue is storage overhead.

ii) Successor Variety

Successor variety stemmers are based on the structural linguistics which determines the word and morpheme boundaries based on distribution of phonemes. Successor variety of a string is the number of characters that follow it in words in some body of text. For example consider a body of text consisting of following words.

Able, ape, beatable, finable, read, readable, reading, reads, red, rope, ripe.

Let's determine the successor variety for the word read. First letter in read is R. R is followed in the text body by 3 characters E, I, O thus the successor variety of R is 3. The next successor variety for read is 2 since A, D follows RE in the text body and so on. Following table shows the complete successor variety for the word read.

Prefix	Successor Variety	Letters
R	3	E,I,O
RE	2	A,D
REA	1	D
READ	3	A,I,S

Table 1.1 Successor variety for word read

Once the successor variety for a given word is determined then this information is used to segment the word. Hafer and Weiss discussed the ways of doing this.

1. Cut Off Method: Some cutoff value is selected and a boundary is identified whenever the cut off value is reached.
2. Peak and Plateau method: In this method a segment break is made after a character whose successor variety exceeds that of the characters immediately preceding and following it.
3. Complete word method: Break is made after a segment if a segment is a complete word in the corpus.

iii) N-Gram stemmers

This method has been designed by Adamson and Boreham. It is called as shared digram method. Digram is a pair of consecutive letters. This method is called n-gram method since trigram or n-grams could be used. In this method association measures are calculated between the pairs of terms based on shared unique digram.

For example: consider two words Stemming and Stemmer

Stemming → st te em mm mi in ng

Stemmer → st te em mm me er

In this example the word stemming has 7 unique digrams, stemmer has 6 unique digrams, these two words share 5 unique digrams st, te, em, mm ,me. Once the number of unique digrams is found then a similarity measure based on the unique digrams is calculated using dice coefficient. Dice coefficient is defined as

$$S=2C/(A+B)$$

Where C is the common unique digrams, A is the number of unique digrams in first word; B is the number of unique digrams in second word. Similarity measures are determined for all pairs of terms in the database, forming a similarity matrix. Once such a similarity matrix is available, the terms are clustered using a single link clustering method.

iv) Affix Removal Stemmers

Affix removal stemmers removes the suffixes or prefixes from the terms leaving the stem. One of the example of the affix removal stemmer is one which removes the plurals form of the terms. Some set of rules for such a stemmer are as follows (Harman)

a) If a word ends in “ies” but not “eies” or “aies ”

Then “ies” -> “y”

b) If a word ends in “es” but not “aes”, or “ees” or “oes”

Then “es” -> “e”

c) If a word ends in “s” but not “us” or “ss ”

Then “s” -> “NULL”

V. Conclusion

In this work we have presented efficient preprocessing techniques. These pre-processing techniques eliminates noisy from text data, later identifies the root word for actual words and reduces the size of the text data. This improves performance of the IR system.

References

1.Vishal Gupta , Gurpreet S. Lehal “A Survey of Text Mining Techniques and Applications” Journal of Emerging technologies in web intelligence, vol,1 no1 August 2009.

2. Durmaz,O.Bilge, H.S “Effect of dimensionality reduction and feature selection in text classification ” in IEEE conference ,2011, Page 21-24 ,2011.

3. G.Salton. The SMART Retrieval System: Experiments in Automatic Document Processing. Prentice-Hall, Inc.

4. Paice Chris D. “An evaluation method for stemming algorithms”. Proceedings of the 17th annual international ACM SIGIR conference on Research and development in information retrieval. 1994, 42- 50.

5. J. Cowie and Y. Wilks, *Information extraction*, New York, 2000.

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