Effect of Stop Word Removal on Document Similarity for Hindi Text

Urvashi Garg
HCTM, Kaithal
urvashi.mittal80@gmail.com

Vishal Goyal
Punjabi University, Patiala
vishal.pup@gmail.com

ABSTRACT-*Stop word removal is one of the important NLP techniques. Stop words are very common in any document. In this paper, we have created a list of stop words for Hindi text on the basis of frequency of words in documents. Hindi documents from EMILLE corpus have been used for finding out the stop words. UTF-8 encoding is used. The percentage of stop words in any document has been find out and experimentally analyzed. The paper discusses the effect of stop word removal on the similarity of two documents containing Hindi text. Hoad & Zobel approach is used for finding the similarity of documents containing Hindi text.*

KEYWORDS: *Stop words, removal, text, Hindi, list, frequency.*

1 INTRODUCTION

Stop words are high frequency words which have very little semantic weight. These words play an important grammatical role in any language such as in formation of sentences but do not contribute to the semantic content of a document. Stop words are commonly used in documents regardless of topic, thus have no significance.

2 ALGORITHM USED

Formation of list of stop words for any language is an intricate task. A lot of work has been done for stop words for English text. Fox [1] have used domain independent approach for creating a list of stop words for English language. The list was used later in Okapi retrieval system [2]. They used word categories like adverbs, prepositions, pronouns etc. for formation of stop word list. But [3] shows that including these categories without any modifications does not make a genuine stopword list. Proper handling of stop words is very necessary for developing any integrated solutions. Hindi word कफ which is a अव्यय is not a stop word in airline or train domain. Because it is necessary to specify time or date in airline or train domain. Similarly, कहाँ can not be included in the stop word list if we want to know the place of any event. So, instead of including अव्यय, सर्वनाम, क्रियाविशेषणas such, we have formed the list on the basis of frequency of words in a corpus. We have used Emille Corpus for finding the list of stop words.

Firstly, all the HTML tags, digits and symbols like |,", /, etc are removed from the text. A (key, value) pair is saved for each word. For each word in the document, if a key exists then its frequency is increased by one, otherwise that word is added to (key, value) pair. Hence, frequency of each word in the corpus is calculated. In the end, the list is displayed in order of decreasing frequency. The corpus consists of approx 60.4 million words out of which there are 1.24 million unique words. Many content bearing words also appeared with high frequency. So, we analyzed the words manually too. After analyzing the 3000 words having highest frequency, a list of 205 Hindi stop words has been created.

Manual additions (तभी,रहो,भइे) and deletions (रोगों, देश,दो) have been done as the results were not appropriate. A list of 165 stop words is available at [4] but according to us it is not complete. The modification in the list of stop words continues.

2.1 LIST OF STOP WORDS

का, के, की, को, क्रक, था, थे,थी, थीीं,है,हैं,ही, हो, हं, रहे,रहो, मह, मे,भें,से,ने,तो, और, बी, रु, तथा, एीं,मा,

इस, इसे, इसभे, इसने, इसकी, इसी, इसका, इसके, इन, इनसे,इन्होने,इनकी, इन्हें,इनका, इनके, उस, उसे, उसभे, उसने, उसकी, उसी, उसका, उसके, उन, उनसे,उन्होने, उनकी, उन्हें,उनका, उनके, जजस, जजसे, जजसभे, जजससे, जजसने, जजसकी, जजसका, जजसके, जजसने, जजन, जजनसे, जजन्होने, जजनकी, जजन्हें,जजनका, जजनके, ्ारा, ्ारेर्ारीी ं,्ारी,महॉ,ी ंहौं,ी ंजहॉ,ी ंहीी ं,महा ी ं,जहाी ं, मही, महीी ं,हीं, हींी ं,ऐसा, ऐसे,ऐसी, ्ैसा,्ैसे, ्ैसी,जैसा,जैसे,जैसी,रग, रगा, रगे,रगी, रगामा, रगाना, रगता, रगाता, रगाए, भै, हभ, हभें,हभाये, हभाया, हभायी, हभने, आऩ, भझे, आऩको, भैने,त,ू तम्हे, तभु, तम्हाया,ु तम्हायी,ु तम्हाये, तभनेु,आऩ, आऩके, आऩकी, आऩने, जफ, तफ, अफ, अबी, तबी, सकना, सका, सके, सकता, सकते, सकती, अऩना, अऩने, अऩनी, अऩनाऩन, इधय, उधय, जजधय, कई, इतना, उतना, जजतना, ऩड, ऩडा, ऩडे,ऩडी, ऩडीी ं,ऩडना, ऩडने, हारी ंक्रक,रेक्रकन, क्मोंक्रक, इसलरए, फजकक, ताक्रक, ऩय, ऩयींतु, मदद, तक, कुछ, फीच, चादहए, द्र्ाया, अनसायु, फाये,अन्म, तयह, लरए, सबी, तयप, क्रपय, फाय, कापी, सफसे, जफक्रक, अरार्ा, फहुत, प्रतत, अगय, के र्, तौय, लसपव, हय, कबी, जजन्हें,जजन्हों, ततन्हें,ततन्हों, इत्मादद, इन्हें,इन्हों, उन्हों, इत्मादद, इन्हीी ं,उन्हीी ं,ग़ैयह,लरमे

3 EFFECT OF STOP WORD REMOVAL ON THE SIZE OF CORPUS

Zipf gave a vital observation on the distribution of words in natural languages.

According Zipf's law, in a corpus, frequency of a word is inversely proportional to its rank. So, words with high frequency have low rank i.e. importance. So they can be removed without affecting the semantics of the text. Stop word elimination can be considered as an implementation of Zipf's law, where high frequency terms are dropped from a set of index terms [5]. Text from the clean document is considered as a bag of words. If a word present in the bag is also in array of stop word list then it is deleted otherwise, it is put in a buffer. Experimentally, we have found that in a corpus of 22.6 million words, the frequency count of stop words is 8.9 million which covers approx 40% of corpus. So if we remove the stop words the size of the corpus reduces significantly. Reduction in size of corpus leads to less number of n grams and less number of index terms. Hence it makes information retrieval faster. Figure 1 shows the percentage of stop words in a particular Hindi corpus. Similar kind of analysis is done in [6] for Punjabi language.

Pandey and Siddiqui [7] suggest that the stop word removal improves information retrieval significantly in terms of precision and recall. However, we have analyzed the impact of stop word removal on similarity of Hindi documents.

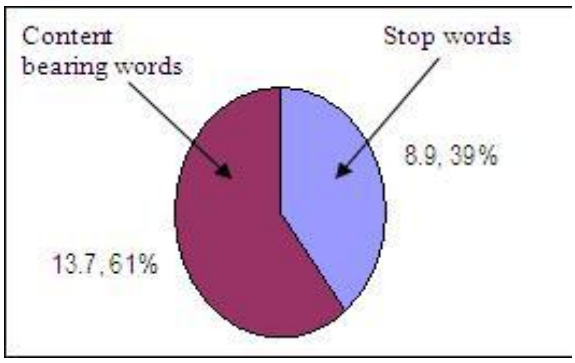

Fig. 1 Percentage of stop words in a corpus

|R|= w - (m-1)

m is the value on n in n-grams. Similarly, the value of |S| is calculated. If a document is evaluated against itself, the given score is the highest possible score that can be reached. So, the calculated score for each document is divided by the highest possible score in order to get a normalized similarity value between 0 and 1.

4 CONCLUSIONS

Where n is the number of documents. Term t denotes the features to be used for evaluation; in this case, they are represented by word n-grams. N-gram is a contiguous sequence of given items from a given text. Total number of n-grams in a document containing w words is given by:

Stop word removal reduces the number of n-grams which leads to time saving. Our experiments suggest that removal of stop words decreases the similarity of documents for Hindi text. Similarity score rises because of frequent words. It suggests that stop word removal eliminates the excess similarity. Ceska and Fox [9] show the effect of stop word removal on determining the identity of fragments of text which is significant. They have done stop word removal for English language.

[10] give improvement in performance with respect to information retrieval when stop word removal is done. Undoubtedly, stop words have a great significance in any language discourse. Stop word removal decreases the degree of comprehension of the text [11] but for more accuracy it is necessary to remove stop words.

REFERENCES

1. Fox, C. (1990) A Stoplist for General Text. SIGIR Forum Vol 24, No 1-2 ,pp 19-35

2. Abu El-Khair ,I. (2006).Effect of Stop Words Elimination for Arabic Information Retrieva A Comparative Study. International Journal of Computing & Information Sciences, Vol 4 No. 3 pp 119-133

3. Dragut, E., Fang, F. , Sistla, P. , Yu, C. & Meng, W. Stop word and related problems in web interface integration (2009). Proceedings of the VLDB Endowment, Vol 2 , Issue 1,pp 349-360.

4. http://members.unine.ch/jacques.savoy/clef/index.html

5. Manning, C.D. & Schutze, H. (1999) Foundations of Statistical Natural Language Processing. The MIT press Cambridge, England. Pp23-24

6. Gupta, V. & Lehal,G. S. (2012) Complete Pre Processing phase of Punjabi Text Extractive Summarization System. Proceedings of COLING 2012: Demonstration Papers, pp 199-206.

7. Pandey A.K & Siddiqui T.J (2009) Evaluating Effect of Stemming and Stop-word Removal on Hindi Text Retrieval. Proceedings of the First International Conference on Intelligent Human Computer Interaction pp 316-326

8. Hoad, T., Zobel, J (2007). Methods for Identifying Versioned and Plagiarized Documents. In Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Amsterdam, The Netherlands pp. 825- 826.

9. Ceska, Z., & Fox, C. (2011). The Influence of Text Pre-processing on Plagiarism Detection. International Conference on Recent Advances in Natural Language Processing 2009. Association for Computational Linguistics, pp. 55-59.

10. Ljiljana, D., & Savoy,F(2009). When Stopword Lists Make the Difference. American Society for Information Science and Technology Vol. 61, Issue 1, pp 200-203

11. Serrano, J. I., del Castillo, M. D., Oliva, J., & Iglesias, A. (2011). The influence of stop-words and stemming on human text base comprehension. Proceedings of the European Perspectives on Cognitive Science.