Effect of Stop-Word Removal for Marathi Language Text Retrieval

Harshali B. Patil
Department of Computer Science
Dr. Annasaheb G. D. Bendale M. M. Jalgaon

Ajay S. Patil
School of Computer Sciences Kavayitri Bahinabai Chaudhari North Maharashtra University Jalgaon

ABSTRACT
Automatic e-document processing systems have been one of the main fields of research and development over past decades. Preprocessing techniques are found to be useful for the process of organizing unstructured text while implementation of various web and data mining techniques like information retrieval, clustering, classifications, etc. Stop-word removal is one of the important preprocessing techniques used for removal of the tokens that do not have any linguistic meaning, and affects on the performance of text mining tasks. These words serve no purpose for Information Retrieval, but they are used very frequently in composing the documents. In modern Information Retrieval process, effective indexing can be achieved by removal of stop words. Many stop word lists have been developed for the major European languages that motivated researchers to work on Asian languages. In case of Indian languages, attempts could be found for Hindi, Bengali, etc. This paper discusses the procedures of two types of stop-word list construction for Marathi text retrieval systems and their impact on reduction in index size. The experimental results reveals that the proposed stop-word list achieves maximum reduction in index size over prior published lists.

General Terms
Information Retrieval, Natural Language Processing

Keywords
Stop-word, Marathi, Fox guidelines, Zipf's law.

1. INTRODUCTION
In Information Retrieval (IR), a document is usually indexed by the words. Statistical analysis of words appearing in documents showed that some words have quite low frequency, while some others are having high frequency. These high frequency words are not carrying any significant information to the document; instead, they are used just because of the grammar. These set of words are called as stop words [1]. In case of Marathi language, "आहे, या, ते, आणि, व, असे, होते, पण" are the stop-words. Stop-word removal is one of the important prepossessing steps that improves the effectiveness of text-retrieval system. Stop-words are grouped into two categories: general and domain specific [2]. Domain specific stop-word list contains the stop-words related to some specific domain while general stop-word list can be applied to any corpus. The domain specific stop-word lists are distinct for different domains. The stop word list contains large number of pronouns, articles, determiners, prepositions, conjunctions, question words & to be verbs [3]. Stop-word list have a wide range of applications; such as information retrieval, natural language processing, digital libraries, document classification, clustering, etc. In IR the stop-word list is generally developed for two main purposes: first one is that each match between a query and a document will be based on the good indexing terms and the second reason is to reduce the size of inverted index which will positively affect the speed of retrieval process. Use of word frequency statistics is a common base for constructing several stop-word lists related to various languages. No commonly accepted stop word list has been constructed for Marathi language. Present work explores the construction of Marathi stop-word lists. The rest of the paper is organized into following sections: Section 2 contains the related work, followed by section 3 that discuss the experiment carried out for stop list development. Section 4 presents the results and section 5 concludes the paper.

2. RELATED WORK
The work of stop-word list development was started in the last century. Some work related to stop-word list development was based on the Zipf's law [4], and the next line follows guidelines given by Fox [5]. The earliest work related to stop-word removal has been carried out by Luhn [6], and suggested that words in natural language text can be divided into keyword terms and non-keyword terms (stop-words). Inspired on this work various pre-compiled stop-lists have been generated such as the Van stoplist and brown stoplist for English language and became standard. The work till 20th century has been mostly carried out for East European languages like English. In 21st century the spectacular growth in digital documents appearing on World Wide Web in native languages gave birth for the research and development of information retrieval for Non-English languages.

An efficient stop-word removal algorithm for Arabic language based on finite state machine (FSM) was presented by Al-Shalabi (2004) and their results approximately reached to 98% [7]. Lo et al. (2005) presented a procedure for automatic building stop-word list for an IR system based on Kullback Leibler divergence measure [8]. In 2006 Zou et al. proposed an automatic aggregated methodology based on statistical and information model for Chinese stop-word list construction and found that their proposed algorithm is a promising technique which saves the time for manual generation [1]. Lazarinis (2007) engineered and utilized a stop-word list in Greek web retrieval and concluded that removal of stop-words is beneficial in traditional IR experiments; however, the importance of stop-word removal in terms of relevance has not been extensively studied in non-English web retrieval [9]. Makrehchi et. al (2008) reported their work related to automatic extraction of domain-specific stop-words from labeled documents. They extracted 150 terms from 6 different domains based on the notion of backward filter level performance and sparsity measure of training data [10]. Dolamic el al. (2009) evaluated the use of two stop-word lists for English language and showed that the performance level may be decrease when using a short stop-word list or no stop-word list [3]. Yuang et al. (2012) did an empirical evaluation of stop-word removal in statistical machine translation [11]. An Arabic stop-word list has been

generated by Alajmi (2012) using statistical approach and their proposed list got improvements as compared to general list [12]. Hassan et al (2014) presented their work based on stopwords, filtering and data sparsity for sentiment analysis of twitter [13]. In 2016 Raulji & Saini creates the Stopword List of 75 words using hybrid approach for Sanskrit language and obtained better results [14]. Kaur & Buttar (2018) carried out a systematic review on stopword removal algorithms and conclude that many stopword removal methods have been developed by the researchers predominantly for the English language and there is a requirement of efficient stopword removal techniques to be developed for other languages [15]. Rani & Lobiyal in 2018 automatically constructed generic stop-word list for Hindi Text and concludes that the proposed technique saves time and create standard list that can be applied on other languages [16]. Sahu & Pal in 2022 presented effect of stop-words in Indian Language IR and conclude that removal of stop-words improved MAP significantly compared with without stop-word removal in general [17].

Though the objectives for developing stopwords list seems to be comprehensible there is not any clear theoretical foundation upon which we can define a methodology for the development of a stop list, thus a certain arbitrariness is required [5]. Table 1 presents some of the available stop-word lists for several languages.

Table.1. Existing Stop-word lists

Language	Stop-word list/Authors/ Reference	Approach	Size
English	Rijsbergen	Luhn's approach	250
	SMART	-	571
	Fox [5]	Fox method & FSM	421
Chinese	Zou et.al [1]	Statistical & information model	NA
	Zhou et al[18]	Merging of different stop-word lists	1289
Arabic	Alajmi et. Al [12]	Statistical approach	200
	Al-shalabi et. Al [7]	FSM	>1000
French	Savoy [19]	Fox method	215
Greek	Lazarinis [9]	Fox method	99
Hindi	Pande et. al [20]	Fox method	350
Bengali	Dolamic et.al.[3]	Fox method	165
	Dolamic et.al.[3]	Fox method	114
Sanskrit	Raulji & Saini [14]	Hybrid approach	75
Hindi	Rani & Lobiyal [16]	Statistical and information-based methods	1475
Marathi	Almedia et.al [21]	Words occur in more than 80% doc.	10
	Dolamic et.al. [3]	Fox method	99

Table 1 shows that the stop-word lists are varying in size. For preparing Marathi stopword list some of the notable efforts were reported by Dolmaic et.al. [3]. where they presented the list of 99 terms as stopwords for Marathi language. Another attempt is carried out by Almedia et. al [21] where they had used 10 terms as stop words related to Marathi time and space domain. To the date, there is no standard stopword list exists for Marathi language information retrieval.

3. PROPOSED STOP-WORD LIST DEVELOPMENT

The process of stop-word list development requires a corpus. The corpus used for this experiment is a standard Marathi corpus acquired from FIRE[1]. FIRE is the Forum for Information Retrieval for Indian languages. The corpus consists of 99,275 documents of Marathi newspapers, Maharashtra Times and Daily Sakal spanning from April 2004 to September 2007. All the documents and topics are encoded in UTF-8.

3.1 Development Process

Stop-word lists are constructed by performing the statistical analysis on the corpus of target language. Term frequency, inverse document frequency, document frequency is the most commonly used measure for the task. Document cleaning is performed to remove unnecessary information like tags, punctuation marks, etc. Tokenization is then performed to separate words. Based on the statistical analysis the words are termed as stop-words and indexing terms. The logical view of the process for development of Marathi stop-word list is as given in figure 1.

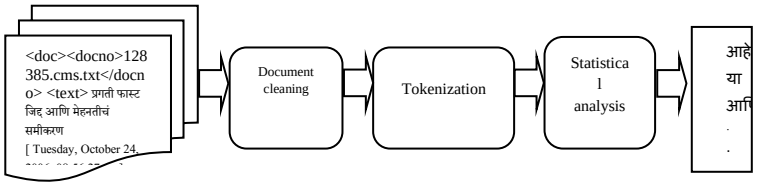

Fig 1: Logical view of proposed stop-word list construction

3.2 ZIPF'S LAW BASED APPROACH

The approach used for development of proposed SL#1 is inspired by the work of Zipf [4]. In the book entitled Human Behavior and the Principle of Least Effort: An Introduction to Human Ecology, the author demonstrated that there exists a correlation between the number of different words and their usage frequency. He states that the occurrence frequency of a word is inversely proportional to its rank i.e. r * f= c where r refers to the word's rank and f to its frequency [4]. The initial steps for development of SL#1 are same as discussed in stop-word list development process. The proposed SL#1 list is developed by adopting the algorithm based on Zipf's law[4,8]. The algorithm is as follows:

- Generate a list of term frequency based on corpus.
- Sort the term frequency list in descending order and assign the rank to the term.
- Draw the graph of term frequency vs rank. This should obey Zipf's law.
- Choose a threshold and any word that appears above threshold is treated as stop-words.

The graph of term frequency vs their rank for the first 500 words in FIRE corpus is as follows.

[1] http://www.isical.ac.in/~fire/

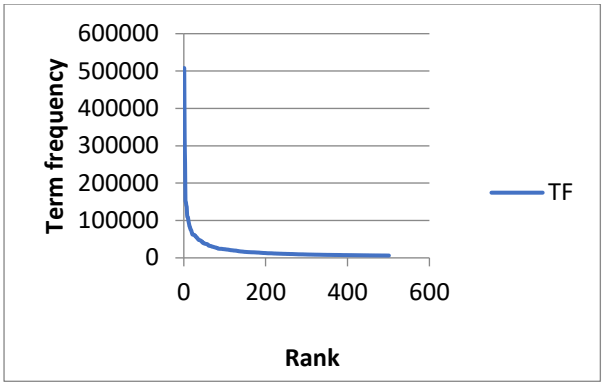

Fig. 2 Term Frequency vs Rank

For finalizing the stop-word list the multiplication of frequency and rank is observed, based on this value a threshold has been set. All the terms those lies above threshold are treated as stop-words while all terms below threshold are not considered as stop-words. Table 2 presents the proposed SL#1 stop-word list for Marathi language developed using this technique.

Table.2. Proposed SL#1 Stop-word list

आहे	झाली	त्यांच्या	कोटी	ही	तरी	यांना	होईल
या	होता	एका	भारतीय	पण	अशा	म्हणाले	असते
आणि	दोन	येथील	निर्णय	करण्यात	म्हणून	असलेल्या	टक्के
व	झाले	सर्व	येणार	काही	त्यांना	म	जात
नाही	त्या	डॉ	महाराष्ट्र	केले	अनेक	गेल्या	अधिक
आहेत	आता	तीन	नाहीत	एक	पोलिस	टा	त्यामुळे
यांनी	मुंबई	न	दिला	केला	माहिती	लाख	तसेच
हे	असा	पाटील	झालेल्या	मात्र	हजार	कमी	त्याला
तर	ता	येथे	पाच	अशी	सांगितले	होणार	आले
ते	आली	असल्याचे	देण्यात	त्यांनी	होत	किंवा	ती
असे	यांच्या	काय	व्यक्त	सुरू	काम	रुपये	येत
होते	की	आपल्या	चार	करून	दिली	का	
केली	तो	म्हणजे	असल्याने	होती	आला	घेऊन	
हा	झाला	मी	दिले	असून	आज	प्रयत्न	

After construction of this list, it is observed and noticed that many Marathi nouns like मुंबई, पाटील, महाराष्ट्र, पोलिस, etc. are included in the stop-word list. These terms may appear many times in this corpus but they can be the index terms and might be used in search queries; hence there is need to perform manual evaluation of terms to be added in the stop-word list.

3.3 FOX GUIDELINES BASED APPROACH

For development of proposed SL#2 stop-word list the guidelines given by Fox are used [5]. The detail procedure used for development is as follows:

- Initially all the word forms appearing in the corpora are sorted according to their frequency of occurrence and then 1000 topmost terms are extracted whose occurrence frequency is greater than 3000. The partial list of terms along with their frequency is given in table 3.

Table.3. Some High frequency terms

Term	Frequency	Term	Frequency
आहे	507755	व	151946
या	320366	आहेत	140193
आणि	254946	यांनी	136380
नाही	149366		

- In second step, the list is inspected to remove all numbers, all nouns & adjectives more or less directly related to with the main subjects of the underlying collections.
- Then some non-information bearing words are introduced, even if they didn't appear in the topmost frequent words. Ex- addition of some pronoun, preposition, conjunction. While selecting the words to be added in the resulting stop-word list, we have hand-picked some high frequency words which are apparently useful and can be expected to appear in IR queries. Table 4 presents list of words that are omitted during the construction of SL#2 stop-word list. Despite being highly frequent, they cannot be considered as unimportant as they can form an integral part of relevant queries.

Table.4. Word omitted during SL#2 development

Term	Frequency	Term	Frequency
सुरू	61719	हजार	37751
मुंबई	45000	सांगितले	37478
पोलिस	38846	काम	37215
माहिती	37852	पाटील	28762

All the words that are omitted in the manual inspection are the meaningful words. These words include some common and proper nouns which might be used while searching, for instance if someone wants the information about Mumbai Police, he/she may use the search query as "मुंबई पोलिस", where both the terms are meaningful terms and not stop-words hence for development of a general stop-word list omission of these terms makes a sense. Proposed SL#2 stop-word list is given in table 5.

Table.5. Proposed SL#2 stop-word list

न	तो	स्वत	अगदी	कारण	तसंच	येथे	यांचे	त्यावर	त्यातून
म	ना	होता	अथवा	काही	तसेच	असतात	यांना	नव्हता	दरम्यान
व	ने	आले	अनेक	किती	तिथे	असेही	यांनी	होत्या	यांच्या
वा	पण	आहे	अन्य	कुठे	तिने	शिवाय	येथील	माझ्या	असलेल्या
ऑफ	मग	इतर	असतं	केलं	तिला	आम्ही	असताना	म्हणून	असल्याचा
का	मी	इथे	असता	केला	होते	आल्या	असलेली	याबाबत	असल्याची
की	या	कधी	असते	होती	तेथे	करणार	असलेले	यामुळे	असल्याचे
के	रु	करत	असतो	केली	त्या	करावा	आपल्या	यावेळी	असल्याने
हा	ला	करू	असली	केले	होतो	काहीच	आमच्या	यासाठी	आपल्याला
चा	वर	कसे	असले	केवळ	नंतर	किंवा	करताना	आम्हाला	करणाऱ्या
ची	अशा	काय	असून	कोणी	नाही	होतील	केलेले	आलेल्या	करण्याचा
चे	अशी	कोण	असेल	गेला	होईल	तरीही	केल्या	करण्यात	करण्याची
जर	असं	खूप	आणखी	गेली	फक्त	त्याच	झालेले	कोणताही	करण्याचे
जी	होत	तरी	आपला	च्या	माझा	त्यात	तिच्या	कोणतीही	त्यांच्यावर
ही	असा	तसे	आपली	ज्या	माझी	नव्हे	तुम्ही	तुम्हाला	यांच्याकडे
जे	असे	हेच	आपले	झाला	होऊन	नाहीत	त्यांचा	त्याच्या	करण्यासाठी
टा	आणि	नको	होतं	झाली	परंतु	तेव्हा	त्यांची	त्यानंतर	असल्यामुळे
तर	आता	नये	आहेत	याचा	मध्ये	त्याचा	त्यांचे	त्यामुळे	त्यामुळेच
ता	आदी	मला	करणे	याची	मात्र	त्याची	त्यांना	त्यावेळी	होणाऱ्या
ती	आपण	याच	करता	याचे	यांचा	त्याचे	त्यांनी	त्यासाठी	कोणत्याही
हे	आला	यात	करीत	याला	यांची	त्याने	त्यातील	नसल्याचे	
ते	आली	झाले	करून	यावर	त्यांच्या	त्याला	कोणत्या	झालेल्या	

4. RESULT

Here two stop-word lists for Marathi language text retrieval are proposed and evaluated for the indexing process. For evaluating these stop-word lists, three different test datasets are used. The first test dataset is prepared from online archives of *Esakal* Marathi newspaper while the rest two test datasets are prepared from FIRE corpus. Each test data set contains 500 documents. The proposed stop-word lists are compared with existing stop-word lists for indexing process. Table 6 provides the reduction obtained due to the use of proposed stop-word lists for indexing task.

Table.6. Reduction obtained

Approach	TDS1 reduction (%)	TDS 2 reduction (%)	TDS 3 reduction (%)
Almeida et.al.[21]	3.50	1.53	1.46
Dolamic et.al. [3]	13.09	11.56	10.58
Proposed list SL#1	24.21	24.13	21.83
Proposed list SL#2	23.64	23.19	23.92

From table 6 it is observed that with proposed SL#1 average reduction obtained is 23.39 while 23.58% reduction is obtained with proposed SL#2 list which is much greater as compared to reduction obtained by already existing stop-word list.

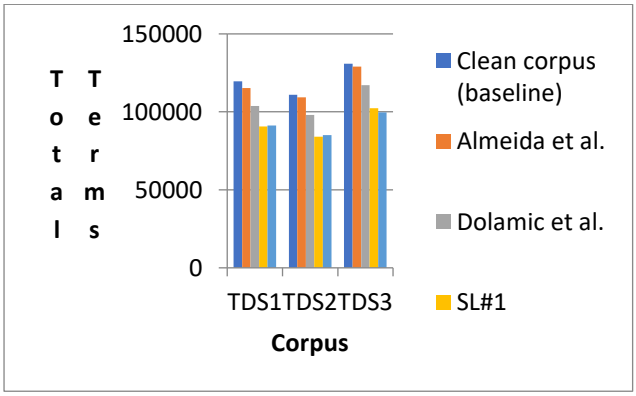

Fig. 3 Term Frequency vs Rank

From fig. 3 it is clearly observed that proposed SL#2 list outperforms for indexing process of Marathi documents as compare to previously available lists.

5. CONCLUSION

In this paper, the major emphasis is on the development process of stop-word lists for Marathi language information retrieval task. Stop-word list is one of the significant resources that are used in many natural language processing applications like: information retrieval, text mining, clustering, classification systems etc. Being a resource-poor language, very few standard NLP resources are available for Marathi language. This paper detailed the development & evaluation of two types of stop-words lists for Marathi language. The lists are tested in terms of reduction in index size and concluded that the proposed list SL#2 got near about 24% reduction in number of words when tested on 500 documents. We found that Fox guidelines-based approach for stop-word list development outperforms amongst all other approaches.

In future, the impact of these lists will be checked for Marathi text retrieval process along with stemming and other pre-processing tools. The applicability of these lists for other related applications will also be checked.

6. ACKNOWLEDGMENTS

The authors are very much thankful to Forum for Information Retrieval for providing the Marathi corpus which had been used as a corpus for this experiment.

7. REFERENCES

[1] F. Zou, F. L. Wang, X. Deng, S. Han, and L. S. Wang, "Automatic Construction of Chinese Stop Word List", *in Proceedings of the 5th WSEAS International Conference on Applied Computer Science*, pp. 1010-1015, 2006.

[2] A. S. Patil and B. V. Pawar, "Analysis of Traditional Information Retrieval Techniques Applied to the World Wide Web", *International Journal in Computer Science and Information Technology (IJCSIT)*, Vol. 1, No. 2, pp-63-73,2008.

[3] L. Dolamic and Jacques Savoy, "Comparative Study of Indexing and Search Strategies for the Hindi, Marathi and Bengali Language", *ACM Transactions on Asian Language Information Processing* Vol. 9, No. 3, Article No.: 11, pp 1–24, 2010.

[4] K. Zipf, "Human behaviour and the principle of least effort: an introduction to human ecology", Addison-Wesley Press, 1949.

[5] C. Fox, "A stop list for general text." *In ACM SIGIR Forum*, vol. 24, no. 1-2, pp. 19-21, ACM 1989.

[6] H. P. Luhn, "A statistical approach to mechanized encoding and searching of literary information." *IBM Journal of research and development* Vol. 1, no. 4, , pp. 309-317, 1957.

[7] R. Al-Shalabi, G. Kanaan, J. M. Jaam, A. Hasnah and E. Hilat, "Stop-word removal algorithm for Arabic language," *Proceedings of the International Conference on Information and Communication Technologies: From Theory to Applications,* pp. 545, *2004.*

[8] R. T.W. Lo, B. He and I. Ounis, "Automatically Building a Stopword List for an Information Retrieval System", *Journal of Digital Information Management*, Vol. 3 No. 1 pp 03- 08, 2005.

[9] F. Lazarinis, "Engineering and utilizing a stopword list in Greek Web retrieval", *Journal of the American Society for Information Science and Technology*, Vol. 58, No. 11 , pp. 1645-1652., 2007

[10] M. Makrehchi and M. S. Kamel, "Automatic Extraction of Domain-Specific Stopwords from Labeled Documents*", In European Conference on Information Retrieval* pp. 222-233, 2008.

[11] C. T. Yuang, R. E. Banchs, and C. E. Siong, "An empirical evaluation of stop-word removal in statistical machine translation", *in Proc.13th conference of the European chapter of the association for computational linguistics* 2012, pp 30-37.

[12] A. Alajmi, E. M. Saad, and R. R. Darwish. "Toward an ARABIC stop-words list generation." *International Journal of Computer Applications* Vol.46, no. 8, pp.8-13,2012.

[13] S. Hassan, F. Miriam and A. Harith, "On stopwords, filtering and data sparsity for sentiment analysis of twitter", *in Proceedings of the 9th International Language Resources and Evaluation Conference*, pp. 810-817, 2014

[14] J. K. Raulji and J. R. Saini, "Stop-Word Removal Algorithm and its Implementation for Sanskrit Language", *International Journal of Computer Applications*, Vol. 150 – No.2, pp. 15-17, 2016.

[15] J. Kaur and P. K. Buttar, "A Systematic Review on Stopword Removal Algorithms", *International Journal on Future Revolution in Computer Science & Communication Engineering*, Vol. 4, No. 4, pp. 207-210, 2018.

[16] R. Rani, and D. K. Lobiyal, "Automatic Construction of Generic Stop Words List for Hindi Text", *Procedia Computer Science, Vol.* 132 pp. 362-370, 2018.

[17] S. S. Sahu and S. Pal, "Effect of Stop-words in Indian Language IR", *Sadhana, Vol .47, No. 1,* 2022 .

[18] Y. Zhou, and C. Ze-wen. "Research on the construction and filter method of stop-word list in text preprocessing." *In Intelligent Computation Technology and Automation (ICICTA)*, vol. 1, IEEE, pp. 217-221, 2011.

[19] J. Savoy, "A Stemming Procedure and Stopword List for General French Corpora", *Journal of the American Society for Information Science*, Vol. 50, No. 10, pp. 944–952, 1999.

[20] A.K Pandey., and T.J. Siddiqui, " Evaluating Effect of Stemming and Stop-word Removal on Hindi Text Retrieval", *In Proceedings of the First International Conference on Intelligent Human Computer Interaction*, 2009.

[21] A. Almeida and P. Bhattacharya, "Using Morphology to Improve Marathi Monolingual Information Retrieval", *in proceeding of Forum for Information Retrieval Evaluation*, 2010