# AN EFFECTIVE PRE-PROCESSING ALGORITHM FOR INFORMATION RETRIEVAL SYSTEMS

Vikram Singh and Balwinder Saini

Department of Computer Engineering,
National Institute of Technology, Kurukshetra, Haryana, India

## *ABSTRACT*

*The Internet is probably the most successful distributed computing system ever. However, our capabilities for data querying and manipulation on the internet are primordial at best. The user expectations are enhancing over the period of time along with increased amount of operational data past few decades. The data-user expects more deep, exact, and detailed results. Result retrieval for the user query is always relative o the pattern of data storage and index. In Information retrieval systems, tokenization is an integrals part whose prime objective is to identifying the token and their count. In this paper, we have proposed an effective tokenization approach which is based on training vector and result shows that efficiency/ effectiveness of proposed algorithm. Tokenization on documents helps to satisfy user's information need more precisely and reduced search sharply, is believed to be a part of information retrieval. Pre-processing of input document is an integral part of Tokenization, which involves pre-processing of documents and generates its respective tokens which is the basis of these tokens probabilistic IR generate its scoring and gives reduced search space. The comparative analysis is based on the two parameters; Number of Token generated, Pre-processing time.*

## *KEYWORDS*

*Information Retrieval (IR), Tokenization, Indexing/Ranking, Pre-processing, Stemming*

## 1. INTRODUCTION

Information Retrieval (IR) is the science of searching for documents, information within Documents, metadata about documents, relational databases and the World Wide Web. Summarization is a branch which deals with information retrieval [18]. Text summarization is the process of creating a summary of one or more text documents. For instance, we may summarize a large amount of news from different sources. Information retrieval system mainly consists of two phases, storing indexed documents and retrieval of relevant results, as shown in figure 1. Phase 1, mainly focus on the identification of tokens, and index the tokens based on some parameters [4]. It is clear, that identification of token is important and critical aspect of IR model.

Tokenization enables text/ token categorization/classification and text summarization, which is an active research topic in the area of data mining. The task of text categorization/classification is to classify a document under a predefined category. A document refers to piece of text. Categories may be derived from a sparse classification scheme or from a large collection of very specific text

DOI : 10.5121/ijdms.2014.6602

documents [20]. Categories may be represented numerically or using single word or phrase or words with senses, etc. In traditional approach, categorization task was carried out manually using domain experts [21]. Each incoming text documents was read and comprehended by the experts and assigned to one or more number of categories chosen from the set of predefined categories. It is inevitable that enormous human efforts was required.

Text summarization is the most challenging task in information retrieval tasks [19]. Text summarization is the process of creating a summary of one or more text documents. For instance, we may summarize a large amount of news from different sources [17]. Many summarization techniques and their evaluation methods have been developed for this purpose. Many summarization techniques and their evaluation methods have been developed for this purpose. Such techniques are RANDOM [20], LEAD [20], MEAD [21] and PYTHY [22] etc. which are used to generate the summary. MEAD is the recent toolkit for summarization. We developed a multidocument, topic-driven summarizer. The input documents were newswire articles from AQUAINT-2 Information-Retrieval Text Research Collections and they were guaranteed to be related to their given topic. The topics themselves represent “real-world questions” that the summaries should answer. In data mining, text is analogues to the token/ concept/ word. Text classification/ categorization and summarization both involved pre-processing of input document.

Tokenization is a process of identification of token/topics within input documents and it helps to reduced search with significant extent [5]. The secondary advantage of tokenization in effective use of storage space, as it reduces the storage spaces required to store tokens identified from input documents [14]. In modern age of data/information, when data/information is expanding manifold on every day from its origin, in form of documents, web pages etc, so importance of effective and efficient tokenization algorithm become critical for an IR system. There are various traditional techniques for tokenization is designed, Porter’s algorithm is one of the most prominent tokenization among all such techniques, but this algorithm suffers from accuracies during the identification and efficiency [15]. The enhanced algorithm is also designed to overcome the inaccuracy in token identification, but problem still persists. In this paper, an approach is proposed for of tokenization, in which is token identification is completely based on the documents vectors.

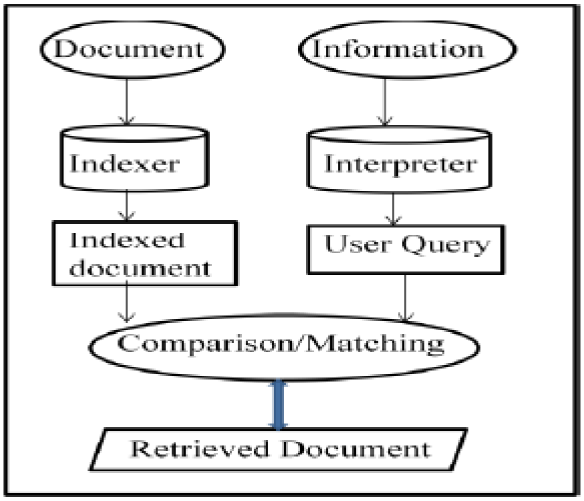


Fig. 1 Formal IR Model System [3]

Tokenization process is an integral part of IR systems, involves pre-processing of given documents and generates respective tokens. In some, tokenization techniques count of token were used to establish a value “Word Count or Token Count” which is used in indexing/ranking process of document[20] [21]. A typical structure of tokenization process is explained in figure 2. Information retrieval models historically many years back to the beginning of written language as information retrieval is related to knowledge stored in textual form [4]. Ranking algorithm/ Indexing algorithm uses the input from tokenization, which is either word count or token count? The affectivity of indexing algorithm is heavily depends upon the quality of token generated by tokenization process. Proposed tokenization model for tokenization is shown in figure 2, primary objective of the tokenization is to identify the words/token/concept and their frequency within input document.

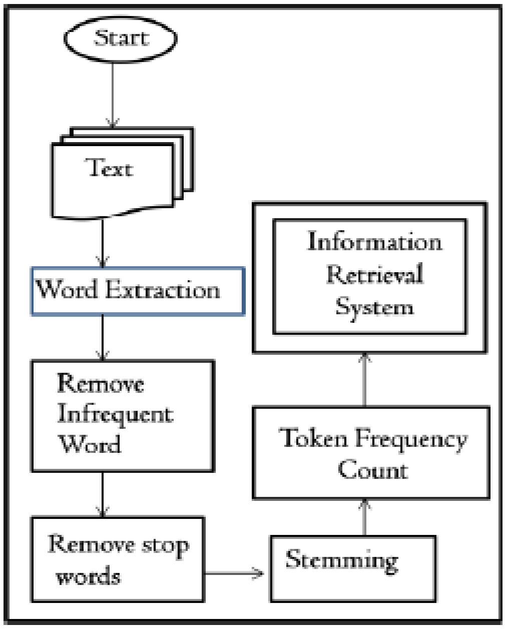


Fig. 2 Tokenization Process

Clearly, the crucial focus of an IR model is to find the relevant document to issue of finding the relevant document to user’s query. Such a decision is usually dependent on an indexing/ranking algorithm which attempts to establish a simple ordering of the documents retrieved [6] [7]. Documents appearing at the top of this ordering are considered to be more likely to be relevant and useful for future patterns. Thus, ranking algorithms are at the core of information retrieval systems. A ranking algorithm operates according to basic premises (regarding document relevance) yield distinct information retrieval models. The IR model adopted determines the predictions of what is relevant and what is not (i.e. the notion of relevance implemented by the system).

**Related Work-** Traditional document tokenization techniques are being used in various unsupervised learning approaches for solving problems [7]. Traditional approaches often fail to obtain good tokenization solution when users want to group documents according to their need [9]. An approach to make an effective pre-Processing steps to save both space and time requirements by using improved Stemming Algorithm [11]. Stemming algorithms are used to transform the words in texts into their grammatical root form [11]. Several algorithms exist with different techniques. This is most widely used porter's stemming algorithm [11]. The other enhanced working model is also proposed, in which inaccuracies encountered during the stemming process has been removed by proposing a solutions [9]. The tokenization involves multiple activities to be performed during the life cycle [13]. There are still a lot of scope of improvement on the accuracy of token identification capability of algorithm & efficiency of approach [11][12].

## 2. INFORMATION RETRIEVAL

Classical retrieval modeling considers documents as bags of words. This stands for the outlook of the model as an entity without structure where only the numbers of occurrences of terms are important for determining relevance. Whenever a query is posed to a retrieval system every document is scored with respect to the query [3]. The scores are sorted and then complete and final ranked list is presented to the user. A retrieval model is in charge of producing these scores. In general models for retrieval do not care about efficiency: they solely focus on understanding a user's information need and the ranking process.

- The user's internal cognitive state or information need is turned into an external expression or query based on a query model.
- Each document is assigned a representation that indicates what the document is about and what topics it covers based on a document model and it is the responsibility of similarity function to estimate the relevance between query & stored word and retrieve.

Therefore the three classic models in information retrieval are called Boolean, Vector and Probabilistic. In the Boolean model documents and queries are represented as set of index terms. Thus we say that the model is set theoretic. In the vector model documents and queries are represented as vectors in the dimensional space. Thus we say that the model is algebraic. In the probabilistic model the framework for modeling document and query representations is based on probability theory [3] [4]. Thus as the name indicates we say that the model is probabilistic.

## 3. TOKENIZATION

Tokenization is a critical activity in any information retrieval model, which simply segregates all the words, numbers, and their characters etc. from given document and these identified words, numbers, and other characters are called tokens [7] [8]. Along with token generation this process also evaluates the frequency value of all these tokens present in the input documents. All the phases of tokenization process are shown in figure 2. Pre-processing involves the set of all documents are gathered and passed to the word extraction phases in which all words are extracted [12]. In next phase all the infrequent words are queued and removed, like removal of word having frequency less than two. Intermediate results are input to the next step, which is stop word removal phase. Primary objective of this phase is to remove words, which are useless in information retrieval these english words are known as stop words. The generic definition of useless is indication for the words type, e.g. adjective, conjunctive etc words are not been generally consider under this category.

For example stop words [2] include "the, as, of, and, or, to etc. This phase is very essential in the tokenization because it' advantages, some are like: It reduces the size of indexing file and it also improves the overall efficiency and makes effectiveness in result retrieval. Next phase in tokenization is stemming [2]. Stemming phase is used to extract the sub-part i.e. called as stem/root of a given word [8][9][11]. For example, the words continue, continuously, continued all can be rooted to the word continue. The main role of stemming is to remove various suffixes as result in the reduction of number of words, to have exactly matching stems, to minimize storage requirement and maximize the efficiency of IR Model. The typical stemming process is illustrated in figure 3. On the completion of stemming process, next step is to count the frequency of each word. Information retrieval works on the output of this tokenization process for achieving or producing most relevant results to the given users [7] [14].

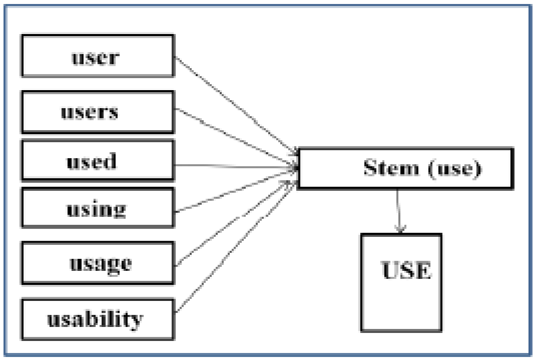
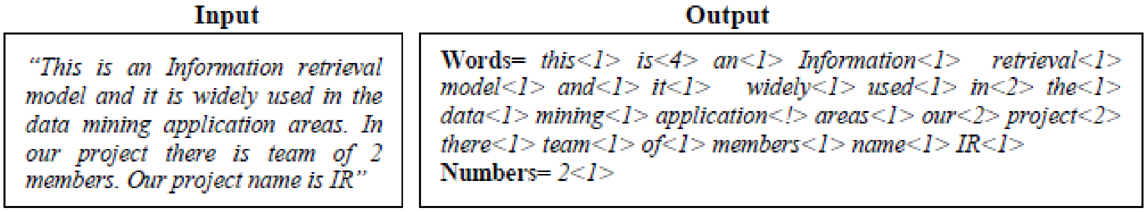


Fig.3 Stemming process on "USE" related stem

For example, there is a document in which the information likes "This is an Information retrieval model and it is widely used in the data mining application areas. In our project there is team of 2 members. Our project name is IR". If this document is passing to the tokenization technique or process then the output of the process is like it separate the words this, is, an, information etc. if there is a number it can also separate from other words or numbers and finally give the tokens with their occurrences count in the given document. This is shown by following example:



**Input**

*"This is an Information retrieval model and it is widely used in the data mining application areas. In our project there is team of 2 members. Our project name is IR"*

**Output**

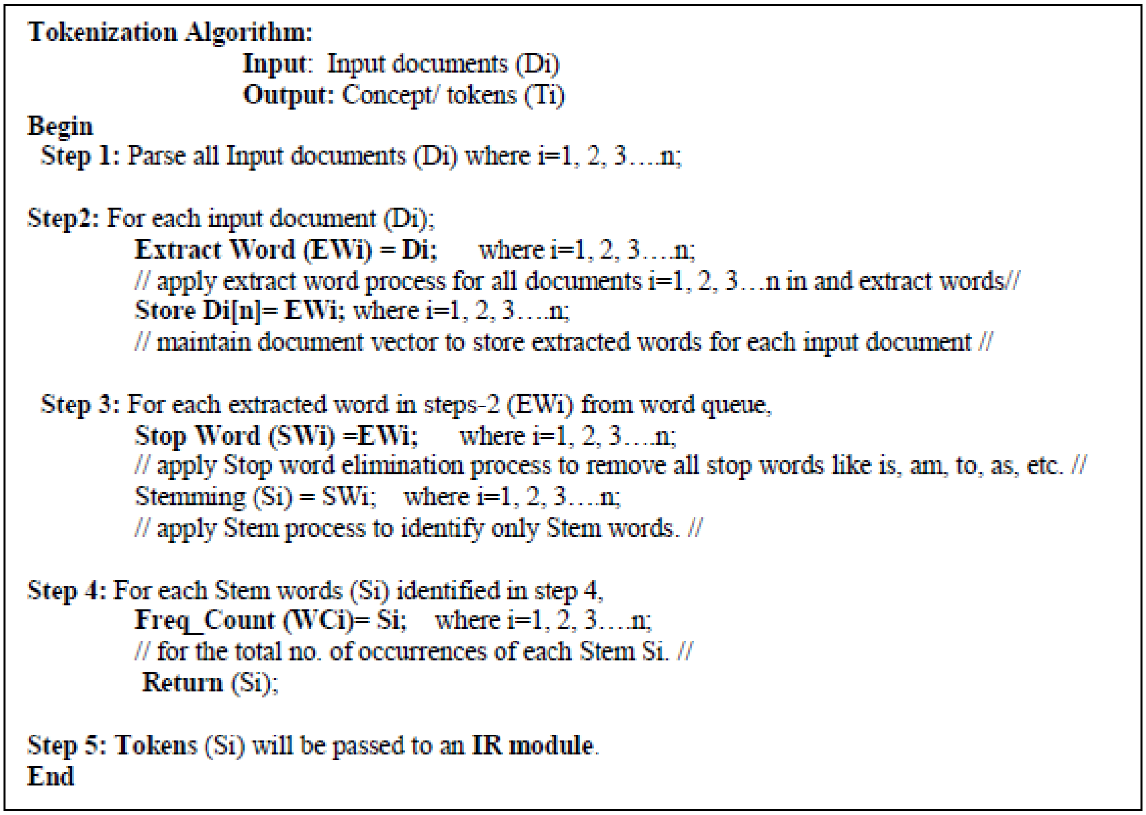
**Words=** *this<1> is<4> an<1> Information<1> retrieval<1> model<1> and<1> it<1> widely<1> used<1> in<2> the<1> data<1> mining<1> application<!> areas<1> our<2> project<2> there<1> team<1> of<1> members<1> name<1> IR<1>*
**Numbers=** *2<1>*

Fig. 4: Input Document (a) & respective output tokens (b)

After applying tokenization process to the figure 4 (a) then the output is formed like 4(b). In angular braces, the value shows the frequency of a word in the given document for example word "our" and "project" occurs two times in the document so their frequency is 2. It also provides the facility to separate the stop words and only gives the distinct words form the given document. In this paper, tokenization process plays a crucial part of finding distinct keywords and their respective frequency values present in the document. The tokenization technique, which tokenize all the documents and then applying the working principle of probabilistic information retrieval model on the output of this tokenization technique for finding their probability scores it extends the overall ranking process for obtaining better results.

## 4. PROPOSED ALGORITHM AND EXAMPLE

The tokenization on documents is performed with respect to the vectors, use of vectors in pre processing helps to make whole tokenization process more precise and successful, in the proposed algorithm. The effect on tokenization by the use vectors in IR system is shown in results section, where the number of token generated and over-all time consumed for the process significantly differ. The steps of proposed algorithm for the tokenization of given documents in information retrieval systems is below discussed.



```
Tokenization Algorithm:
                    Input:  Input documents (Di)
                    Output: Concept/ tokens (Ti)
Begin
 Step 1: Parse all Input documents (Di) where i=1, 2, 3….n;

Step2: For each input document (Di);
         Extract Word (EWi) = Di;      where i=1, 2, 3….n;
         // apply extract word process for all documents i=1, 2, 3…n in and extract words//
         Store Di[n]= EWi; where i=1, 2, 3….n;
         // maintain document vector to store extracted words for each input document //

 Step 3: For each extracted word in steps-2 (EWi) from word queue,
         Stop Word (SWi) =EWi;      where i=1, 2, 3….n;
         // apply Stop word elimination process to remove all stop words like is, am, to, as, etc. //
         Stemming (Si) = SWi;    where i=1, 2, 3….n;
         // apply Stem process to identify only Stem words. //

Step 4: For each Stem words (Si) identified in step 4,
         Freq_Count (WCi)= Si;    where i=1, 2, 3….n;
         // for the total no. of occurrences of each Stem Si. //
          Return (Si);

Step 5: Tokens (Si) will be passed to an IR module.
End
```

| Document No. | Documents Contents |
|---|---|
| **D1** | *Military is a good option for a career builder for youngsters. Military is not covering only defense it also includes IT sector and its various forms are Army, Navy, and Air force. It satisfies the sacrifice need of youth for their country.* |
| **D2** | *Cricket is the most popular game in India. In crocket a player uses a bat to hit the ball and scoring runs. It is played between two teams; the team scoring maximum runs will win the game.* |
| **D3** | *Science is the essentiality of education, what we are watching with our eyes happening non-happening all include science. Various scientists working on different topics help us to understand the science in our lives. Science is continuous evolutionary study, each day something new is determined.* |

| | |
|---|---|
| **D4** | *Engineering makes the development of any country, engineers are manufacturing beneficial things day by day, as the professional engineers of software develops programs which reduces man work, civil engineers gives their knowledge to construction to form buildings, hospitals etc. Everything can be controlled by computer systems nowadays.* |
| **D5** | *Land rover to built discovery sport in brazil and also opens new plant in America. Maruti launches the new version of alto k10 with or without automatic facility in the Indian market.* |
| **D6** | *Religion is a set in which collection of beliefs, different-different cultures and the views that are related with civilization. At present so many religions are there and all are explaining the origin of life and origin of universe* |
| **D7** | *Economy is also called as economic system and it consists of several things like production, distribution, using up of important goods and services by different agents in a given ecological scene.* |
| **D8** | *Entertainment includes so many things that are responsible for holding the attention and interest of an audience. It includes various things like entertainment news, celebrities, fashion, movies, music and television shows etc.* |

The input documents are written in the English literature and well versed from the specific to the area/domain. Above mentioned document are used as input for the proposed algorithm and preprocessing is applied on the document.

**Phase 2:**

Input documents are preprocessed and further; all the words/ tokens are extracted from each of the input document. The primary objective of this step is to separate out each individual word/token from given input paragraph of document. The word extraction process does use the semantic meaning of document or sentence or paragraph for the extrication of a word. However process performs word extraction based assuming a plain string .The extracted words from each of the document is shown following:

Document: D1
[Military, is, a, good, option, for, a, career, builder, for, youngsters, Military, is, not, covering, only, defense, it, also, includes, IT, sector, and, its, various, forms, are, Army,, Navy,, and, Air, force., It, satisfies, the, sacrifice, need, of, youth, for, their, country.]

Document: D2
[Cricket, is, the, most, popular, game, in, India. In, crocket, a, player, uses, a, bat, to, hit, the, ball, and, scoring, runs. It, is, played, between, two, teams;, the, team, scoring, maximum, runs, will, win, the, game.]

Document: D3
[Science, is, the, essentiality, of, education,, what, we, are, watching, with, our, eyes, happening, non-happening, all, include, science., Various, scientists, working, on, different, topics, help, us, to, understand, the, science, in, our, lives., Science, is, continuous, evolutionary, study,, each, day, something, new, is, determined.]

Document: D4

[Engineering, makes, the, development, of, any, country,, engineers, are, manufacturing, beneficial, things, day, by, day,, as, the, professional, engineers, of, software, develops, programs, which, reduces, man, work,, civil, engineers, gives, their, knowledge, to, construction, to, form, buildings,, hospitals, etc., Everything, can, be, controlled, by, computer, systems, nowadays.]

Document: D5

[land, rover, to, built, discovery, sport, in, brazil, also, open, new, plant, in, America, maruti, launch, new, version, alto, k10, or, without, automat, facility, in, indian, market]

Document: D6

[religion, set, in, which, collect, belief, different, different, culture, view, that, are, related, civilization , at, present, so, many, religion, are, there, all, are, explain, origin, life, origin, universe]

Document: D7

[Economy, also, call, economic, system, it, consist, sever, thing, like, product, distribution, us, up, import, good, services, by, different, agent, in, given, ecology, scene]

Document: D8

[Entertainment, include, so, many, thing, that, are, responsible, for, hold, attention, interest, and, audience, it, include, various, thing, like, entertain, new, celebrities, fashion, movie, music, television, show, etc]

**Phase 3 and Phase 4:**

The text generated post extraction is going through a process of removal & stemming. The Removal process simply removes stop words from the set of extracted words. The resultant of the removal & stemming process is shown the following:

Document: D1

[Military, good, option, for, career, builder, for, youngster, military, not, cover, online, defense, it, also, include, it, sector, it, various, form, are, army, navi, air, force, it, satisfy, sacrifice, need, youth, for, their, country]

Document: D2

[Cricket, most, popular, game, in, India, in, crocket, player, us, bat, to, hit, ball, score, run, it, plain, between, two, team, team, score, maximum, run, win, game]

Document: D3

[science, essential, educate, what, we, are, watch, our, eye, happen, non, happen, all, include, science, various, scientist, work, on, differ, topic, help, to, understand, science, in, our, live, science, continue, evolutionary, studies, each, dial, something, new, determine]

Document: D4

[engine, make, develop, countries, engine, are, manufacture, beneficiary, thing, by, profession, engine, software, develop, program, which, reduce, man, work, civil, engine, give, their, knowledge, to, construct, to, form, build, hospital, etc, everything, can, be, control, by, computer, system, now]

Document: D5

[land, rover, to, built, discovery, sport, brazil, open, new, plant, America, maruti, launch, new, version, alto, k10, without, automat, face, Indian, market]

Document: D6

[religion, set, collect, belief, differ, differ, culture, view, are, relate, civil, present, man, religion, all, explain, origin, life, origin, universe]

Document: D7

[economy, call, economy, system, consist, sever, thing, like, product, distribute, up, import, good, service, differ, agent, given, ecology, scene]

Document: D8

[Entertain, include, man, thing, are, response, hold, attend, interest, audience, it, include, various, thing, like, entertain, new, celebrate, fashion, movie, music, television, show]

Now, as above mentioned, the documents are ready to process by information retrieval model. All the comparative and significant improvements on the performance in algorithm are discussed in next subsequent section.

## 5. Results And Experiments

In this section, the results are shown, the comparison on both cases tokenization with vectors (with pre-processing) and tokenization without vectors (without pre-processing) on given input documents are shown. The results shown in the paper are based on the experimentation over more than 100 input documents and more than 50 input document vectors. Further, for the comparative analysis below mentioned parameters are used:

1) Number of Concept Generated: Total no of tokens/topic generated distinctly in one input documents after processing are one of the parameter for result analysis. This number varies in both scenario's, as tokenization with pre-processing generate more accurate and effective token with respect to input document, which results less storage space required and more accurate results to the user. Tokenization without processing leads to large number of tokens, which is difficult to store and affects user results adversely.

2) Approach : There are two alternatives of strategy, tokenization with pre-processing and tokenization without pre-processing. Pre-processing involves creation of document vectors based on training documents and then identifying token on input documents based with respect to vectors. The tokenization with pre-processing generates more accurate and effective tokens with more efficient manner, while in without pre-processing strategy simply parses input documents and generates tokens.

3) Overall-Time Value: Time consumed in entire tokenization process is directly proportional to performance measure of an IR system, as it deeply affects the Indexing & storage aspects.

The performance analysis shown in figure 5 is between strategy (tokenization with pre-processing or without pre-processing) and number of tokens generated. As mentioned previously also, the tokenization with pre-processing generates less no of tokens but the tokens are accurate with in context of result retrieval. In tokenization with pre-processing 275 numbers of tokens generated while for same set of input documents another strategy (without pre-processing) generates more than 400 tokens. The more is the number of token generated, bigger is the challenge to manage them into storage space & effect in accuracy of results retrieval.

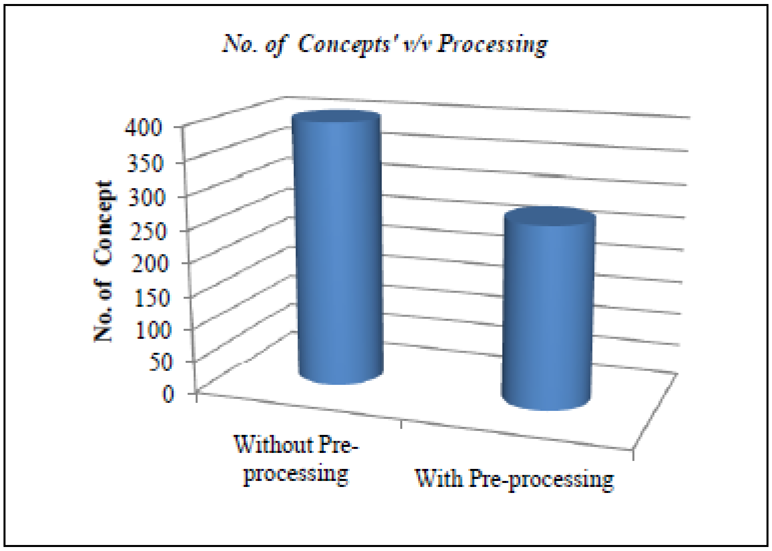


Fig. 5: Document Tokenization Graph

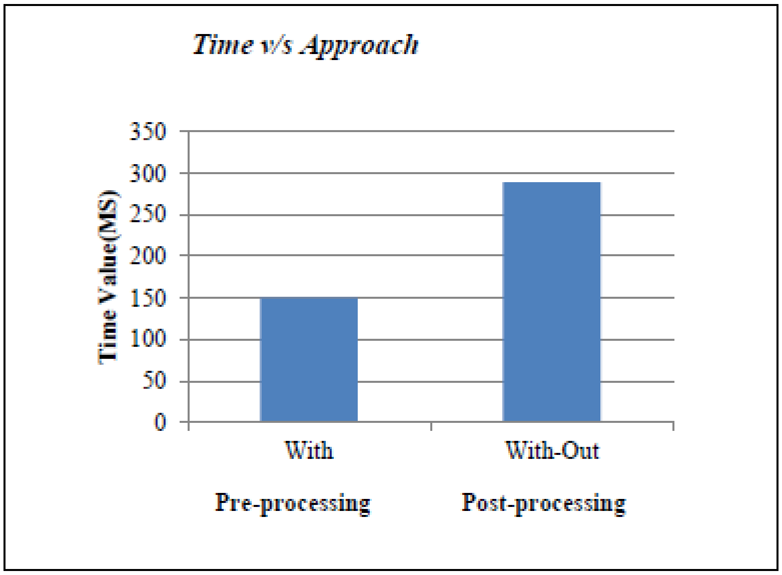


Fig.6: Overall-Time Graph

Another result graph is shown in figure 6; overall time consumed by the strategy is an important factor, which affects overall efficiency of an IR system. The Tokenization with Pre-processing leads to effective and efficient approach of processing, as shown in results strategy with pre-processing process 8 input documents and generate 275 distinct and accurate tokens in 150 (ms), while processing same set of documents in another strategy takes 290 (ms) and generates more than 400 tokens.

## 6. CONCLUSION

IR model centrally focused on providing relevant results to the user. Relevancies of retrieved results are deeply affected with the quality of indexing / ranking algorithm. Finding information is not the only activity that exists in an Information Retrieval (IR) system. Indexing process, represent into document based on some score like word count, which is generally obtained from tokenization process. There are various traditional techniques for tokenizations are designed, Porter's algorithm is one of the most prominent tokenization among all such techniques, but this algorithm suffers from accuracies during the identification and efficiency. The enhanced algorithm is also designed to overcome the inaccuracy in token identification, but problem still persists. In this paper, an approach is proposed for of tokenization, in which is token identification is completely based on the documents vectors. The documents vectors are created after the training process. The vectors plays critical role in overall token identification and make entire process effective and efficient. The performance of different information retrieval models are governed by some conditions which are to be outlined. In the results, it shown that tokenization with pre-processing generates better tokens, as it is with less number of token generated and less storage space is required and it facilitates with more accuracy in results retrieval and this is also responsible for reducing the overall-time value of information retrieval model. This algorithm performs better than traditional algorithm of tokenization because of the accuracy in token identification phase.

## REFERENCES

[1] G. Salton, M.J. Mcgill, "Introduction to Modern Information Retrieval", Mcgraw-Hill Book Co., New York, 1983.

[2] R. Baeza-Yates, B. Ribeiro –Neto, "Modern Information Retrieval", Harlow: Acm Press 1999.

[3] B. Saini, V. Singh, S. Kumar, "Information Retrieval Models And Searching Methodologies: Survey", In International Journal Of Advance Foundation and Research in Computer, pp. 57-62, 2014.

[4] H. Dong, F.K. Husain, E. Chang, "A Survey in Traditional Information Retrieval Models", IEEE International Conference on Digital Ecosystems and Technologies, Pp. 397-402, 2008.

[5] S. Raman, V. Kumar, S. Venkatesan, "Performance Comparison of Various Information Retrieval Models Used in Search Engines", IEEE Conference on Communication, Information and Computing Technology, Mumbai, India, 2012.

[6] J. Hua, "Study on the Performance of Information Retrieval Models", In 2009 International Symposium on Intelligent Ubiquitous Computing and Education, Pp. 436-439, 2009.

[7] J. Qui, C. Tang, "Topic Oriented Semi-Supervised Document Clustering", In Proceedings of SIGMOD, Workshop on Innovative Database Research, pp- 57-52, 2007.

[8] M. Karthikeyan, P. Aruna, "Probability Based Document Clustering and Image Clustering using Content-Based Image Retrieval", In Elsevier Journal of Applied Soft Computing, Pp.959-966, 2012.

[9] Ning Zhong, Yuefeng Li, Sheng-Tang Wu, "Effective Pattern Discovery for Text Mining", IEEE TRANSACTIONS ON KNOWLEDGE Fadi Yamout, "Further Enhancement to the Porter's Stemming Algorithm", Issue 2006

[10] V. Srividhya, R. Anitha, "Evaluating Preprocessing Techniques in Text Categorization - International Journal of Computer Science and Application" Issue 2010.

[11] C.Ramasubramanian, R.Ramya, Effective Pre-Processing Activities in Text Mining using Improved Porter's Stemming Algorithm, International Journal of Advanced Research in Computer and Communication Engineering, Vol. 2, Issue 12, December 2013

[12] Karbasi, S, Boughanem, M. (2006) "Document length normalization using effective level of term frequency in large collections, Advances in Information Retrieval, Lecture Notes in Computer Science", Springer Berlin / Heidelberg, Vol. 3936/2006,Pp.72-83.

[13] Diao, Q, Diao, H. (2000) “Three Term Weighting and Classification Algorithms in Text Automatic Classification”, The Fourth International Conference on High-Performance Computing in theAsia-Pacific Region, Vol. 2, P.629.

[14] S. K. M. Wong, W. Ziarko, P. C. N. Wong, "Generalized vector space model in information retrieval," in the 8th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, New York, 1985, pp. 18-25.

[15] Xue, X, Zhou, Z. (2009) “Distributional Features for Text Categorization”, IEEE Transactions on Knowledge and Data Engineering, Vol.21, No. 3, Pp. 428-442.

[16] Shalini Puri, Sona Kaushik, “A Technical Study and anlaysis on fuzzy similarity based model for text classification”, International Journal of Data Mining & Knowledge Management Process (IJDKP) Vol.2, No.2, March 2012

[17] Meryeme Hadni, Said Alaoui Ouatik, Abdelmonaime Lachkar, “Effective Arabic Stemmer Based Hybrid Approach for Arabic Text Categorization”, International Journal of Data Mining & Knowledge Management Process (IJDKP) Vol.3, No.4, July 2013

[18] A. P. Siva kumar, Dr. P. Premchand, Dr. A. Govardhan, “Query-Based Summarizer Based on Similarity of Sentences and Word Frequency”, International Journal of Data Mining & Knowledge Management Process (IJDKP) Vol.1, No.3, May 2011

[19] Y.V. Haribhakta, Dr. Parag Kulkarni, “Learning Context for Text Categorization”, International Journal of Data Mining and Knowledge Management Process, ) Vol.1, No.3, May 2012

[20] Zhi Teng., Ye Liu., Fuji Ren ., Seiji Tsuchiya., and Fuji Ren “Single Document Summarization Based on Local Topic Identification and Word Frequency”, In Seventh Mexican International Conference on Artificial Intelligence 2008.

[21] Md. Mohsin Ali ., Monotosh Kumar Ghosh., and Abdullah-Al-Mamun., “Multi-document Text Summarization: SimWithFirst Based Features and Sentence Co-selection Based Evaluation” In International Conference on Future Computer and Communication 2009.

[22] Lucy Vanderwende, Hisami Suzuki, Chris Brockett, Ani Nenkova, “Beyond Sum Basic: Task-focused summarization with sentence simplification and lexical expansion (pp. 1606-1618). Information Processing and Management, 43 2007 volume 6 November 2007, Tarrytown, NY, USA: Pergamon Press.