AN EFFECTIVE TOKENIZATION ALGORITHM FOR INFORMATION RETRIEVAL SYSTEMS

Vikram Singh and Balwinder Saini

Department of Computer Engineering,
National Institute of Technology, Kurukshetra, Haryana, India
viks@nitkkar.ac.in
me7saini@gmail.com

ABSTRACT

In the web, amount of operational data has been increasing exponentially from past few decades, the expectations of data-user is changing proportionally as well. The data-user expects more deep, exact, and detailed results. Retrieval of relevant results is always affected by the pattern, how they are stored/ indexed. There are various techniques are designed to indexed the documents, which is done on the token's identified with in documents. Tokenization process, primarily effective 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 of a given documents helps to satisfy user's information need more precisely and reduced search sharply, is believed to be a part of information retrieval. Tokenization 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. No of Token generated is the parameters used for result analysis.

KEYWORDS

Information Retrieval (IR), Indexing/Ranking, Stemming, Tokenization.

1. INTRODUCTION

Information retrieval is always attracted immense research interest and huge possibility in field of data mining. An IR model concerning with representation, storage, access and retrieval of data relevant to user's query [1] [2]. Following are some current research trends [3] in the area of IR:

- Information Searching
- Ranking/Indexing of user's query results.
- Elaborating representation and storage of information
- Classification of documents (i.e. Pre-defined groups)
- Clustering of documents (i.e. Automatically creates clusters)

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

David C. Wyld et al. (Eds) : COSIT, DMIN, SIGL, CYBI, NMCT, AIAPP - 2014
pp. 109–119, 2014. © CS & IT-CSCP 2014 DOI : 10.5121/csit.2014.4910

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 is a process of identification of token/topics within input documents and it helps to reduced search with significant degree [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 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 [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.

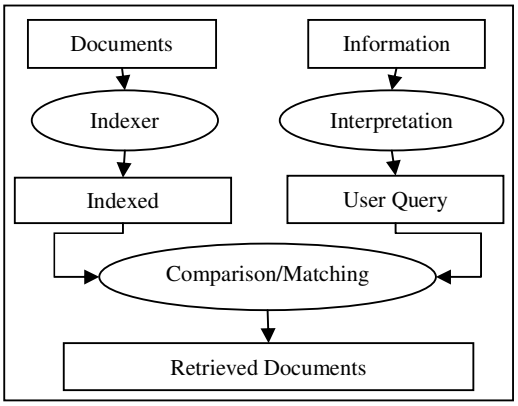

Fig. 1 Formal IR Model System [3]

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 can be used as indexing/ranking process. 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.

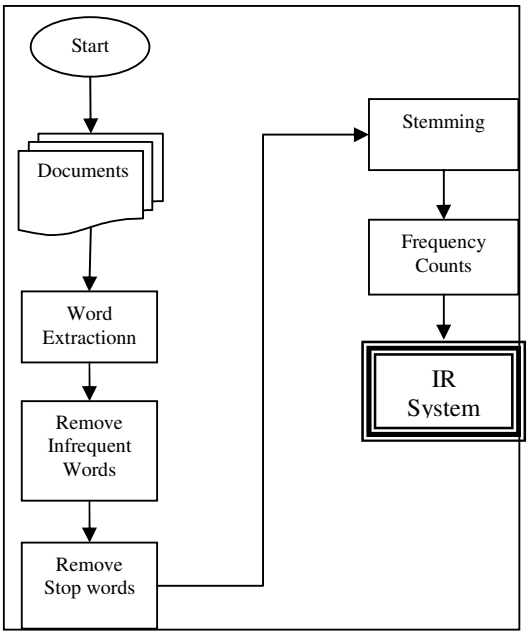

Fig. 2 Tokenization Process

Clearly, the central 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 view 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 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.
- A similarity function can be used to estimate the relevance of a document to the information need based on the document model and on the query model.

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 listed and removed for example remove words having frequency less than two. Intermediate results are passed to the stop word removal phase. In this phase remove those english words which are useless in information retrieval these english words are known as stop words.

For example stop words [2] include "the, as, of, and, or, to etc. this phase is very essential in the tokenization because it has some advantages: It reduces the size of indexing file and it also improve the overall efficiency and make effectiveness.

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.

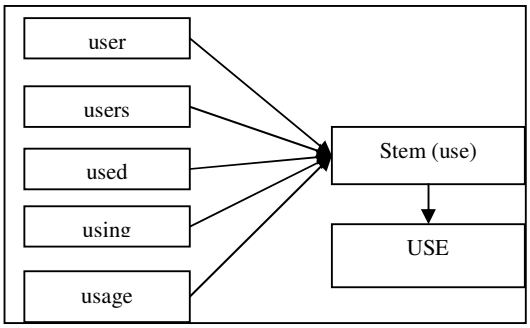

Fig.3 Stemming Process

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].

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”.

Fig. 4: Input Document

After applying tokenization process to the figure 4 then the output is formed like:
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>

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

In the proposed algorithm, tokenization is done based on the set of training vectors which are initially provided into the algorithm to train the system. The training documents are of different knowledge domain, are use to create vectors. The created vector helps algorithm to process the input documents. The tokenization on documents is performed with respect to the vectors, use of vectors in pre tokenization helps to make whole tokenization process more precise and successful. The effect on tokenization of vectors is shown in results section also, where the no of token generated & time consumed for the process significantly differ. Following figure-5 shows the proposed algorithm for the tokenization of documents.

Input (Di)
Output (Tokens)
Begin
Step 1:
Collect Input documents (Di) where i=1, 2, 3….n;
Step2:
For each input Di;
Extract Word (EWi) = Di;
// apply extract word process for all documents i=1, 2, 3…n in and extract words//
Step 3:
For each EWi;
Stop Word (SWi) =EWi;
// apply Stop word elimination process to remove all stop words like is, am, to, as, etc. //
Stemming (Si) = SWi;
// It create stems of each word, like "use" is the stem of user, using, usage etc. //
Step 4:
For each Si;
Freq_Count (WCi)= Si;
// for the total no. of occurrences of each Stem Si. //
Return (Si);
Step 5:
Tokens (Si) will be passed to an **IR System**.
End

Example:-**Phase 1:**

Input Documents:

S.No.	Documents Contents
doc1	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.
doc2	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.

doc3	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.
doc4	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.

Input four documents to the tokenization process, the process will complete the action in following steps,

Phase 2:

In this phase, all the words are extracted from these four documents as shown below:

Name: doc1

[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.]

Name: doc2

[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.]

Name: doc3

[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.]

Name: doc4

[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.]

Phase 3 and Phase 4:

After extracting all the words, next phases is to remove all stop words and stemming, as shown below:

Name: doc1

[militari, good, option, for, career, builder, for, youngster, militari, not, cover, onli, defens, it, also, includ, it, sector, it, variou, form, ar, armi, navi, air, forc, it, satisfi, sacrific, need, youth, for, their, country]

Name: doc2

[cricket, most, popular, game, in, india, in, crocket, player, us, bat, to, hit, ball, score, run, it, plai, between, two, team, team, score, maximum, run, win, game]

Name: doc3

[scienc, essenti, educ, what, we, ar, watch, our, ey, happen, non, happen, all, includ, scienc, variou, scientist, work, on, differ, topic, help, to, understand, scienc, in, our, live, scienc, continu, evolutionari, studi, each, dai, someth, new, determin]

Name: doc4

[engine, make, develop, ani, countri, engin, ar, manufactur, benefici, thing, dai, by, dai, profession, engin, softwar, develop, program, which, reduc, man, work, civil, engin, give, their, knowledg, to, construct, to, form, build, hospit, etc, everyth, can, be, control, by, comput, system, nowadai]

Now, as above mentioned, the documents are ready to process by information retrieval model. All the comparative improvement on the performance in algorithm is discussed in 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 of 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 Tokens 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) ***Strategy***: 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 6 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 200 numbers of tokens generated while for same set of input documents another strategy (without pre-processing) generates more than 300 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.

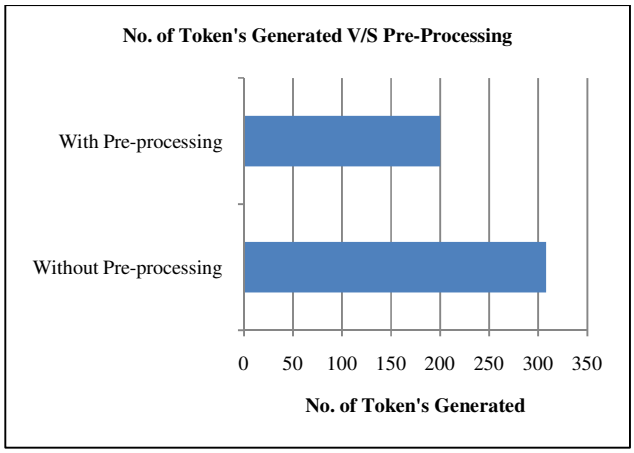

Fig. 6: Document Tokenization Graph

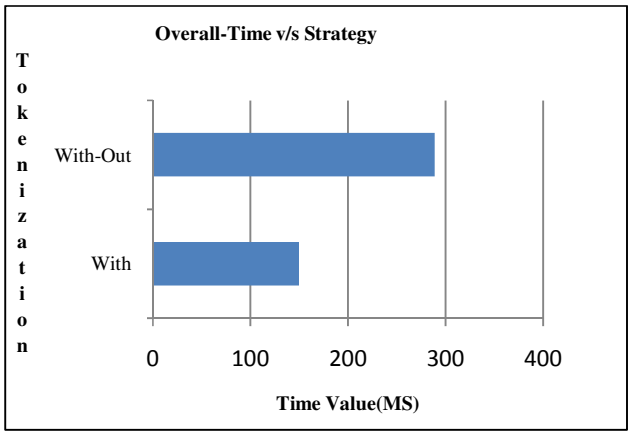

Fig.7: Overall-Time Graph

Another result graph is shown in figure 7, 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 100 input documents and generate 200 distinct and accurate tokens in 156 (ms), while processing same set of documents in another strategy takes 289 (ms) and generates more than 300 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 n 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.

AUTHORS

Vikram Singh (Assistant Professor at NIT Kurukshetra) received his Master of Technology from School of Computer& Systems Sciences, JNU (New Delhi), 2009. He is currently pursuing PhD in Computer Engineering department at National Institute of Technology, Kurukshetra, Haryana, India. His areas of research includes: Distributed Databases System, Query Processing, Ontology and Semantic Web & Big Data Mining.

Balwinder Saini received his bachelor degree in Computer Science and Engineering from Kurukshetra University, Haryana, India, 2011. He is currently pursuing M.Tech degree in Computer Engineering at National Institute of Technology, Kurukshetra, India. His research interests include: Advance Data Base Management System, Knowledge base system and Information Retrieval.