International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 2 Issue 4, April - 2013

# Introducing GA Based Information Retrieval System For Effectively Retrieving News Article

Prof. Anuradha Thakare    Pallavi Kapare    Prajakta Chaudhari    Manisha Kambale

*Department of Computer Engineering, Pimpri Chinchwad College of Engineering, Pune, India*

## Abstract

*One of the challenge of information retrieval system is to retrieve the relevant news articles from particular newspaper dataset in minimum time. To overcome this challenge we used genetic algorithm in information retrieval system. In information retrieval system matching functions like Jaccard's coefficient, Dice coefficient, and Cosine measures has been used to determine the retrieval effectiveness. Effectiveness of information retrieval system is calculated in terms of precision and recall. Information retrieval system helps user to retrieve the most relevant news article according to user query.*



## 1. Introduction

Information Retrieval System (IRS) is used to store set of information that need to be processed and generate a ranking which reflect relevance between query and retrieves information corresponding to user's query. The goal of an Information retrieval system is help to user to locate the relevant documents that have potential to satisfy user's query. An Information Retrieval System consists of a software program that help to user to find information as per their needs.IRS have to extract the keywords from the text documents of news articles and assign weights for each keyword. Matching functions such as dice coefficient, cosine coefficient, jaccard coefficient are used to calculate matching score. The information retrieval efficiency measures from recall and precision. The recall is defined as the proportion of relevant document

## 2. Proposed System Architecture

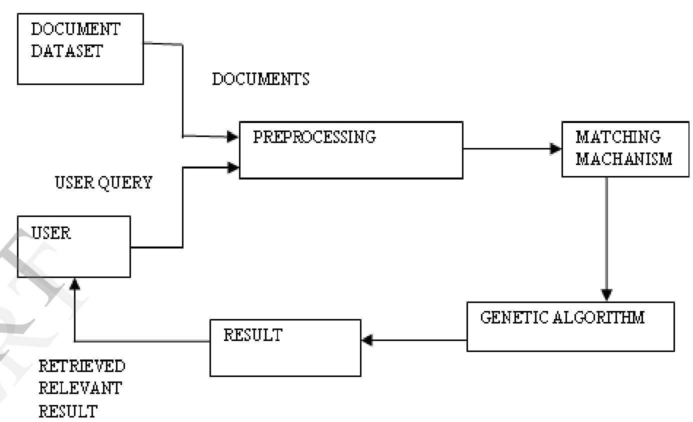


Fig.1 Information retrieval system framework

The information retrieval system consists of various components such as database, pre-processing, matching mechanism, Genetic Algorithm.

### 2.1. Database

This component stores the text documents of news articles. The dataset contains news of October 2011.

### 2.2 Pre-processing

This component preprocesses the text documents and query. The query and documents are converted into weighted vector format by using TF-IDF.

### 2.3. Matching mechanism

Matching mechanism compares set of news article in the dataset with the query given by user. Those news article matches with user query are called relevant news articles [3].

### 2.4. Genetic Algorithm

Genetic algorithms are adaptive heuristic search algorithm [2]. Genetic algorithm is based on evolutionary ideas of natural selection and genetics. Genetic algorithm is often used to solve problems and looking for best solution. In genetic algorithm query entered by user and text documents of news articles stored in data repository represented as chromosome. Each chromosome has specific matching score. By deciding some threshold value we compare the threshold value with matching score and if matching score is greater than threshold value then that chromosomes will be fittest. And if matching score is less than threshold value then that chromosomes will be taken to generate the next generation. The next generation is generated by using genetic operators such as crossover, mutation.

## 3. Proposed Information Retrieval Model

We are representing the information retrieval model of our work by the notations of set theory. S={Q,D,P,F,K,B,GA,Vo}
Where, Q is input query which is entered by user. D is set of documents, d1,d2,d3……….dn. P is parser functions that will return to required keywords .F is TF-IDF algorithm. GA is Genetic Algorithm and Vo denotes Output of weighted vector. We have to compute parser functions for both input query as well as set of documents using following functions. Tq=P(Q) and Td=P(d). Tq is the result of parser function for input query. Td is the result of parser function for set of documents.

### 3.1. Document Ranking

TF-IDF Algorithm is a ranking algorithm. In TF-IDF algorithm, there are two terms TF and IDF. TF stands for term frequency.IDF stands for inverse document frequency.TF is used to determine how many number of times a keyword is present in documents. Inverse document frequency (IDF) can be calculated by log(N/n) where 'N' as number of documents in a collection and 'n' as number of documents containing a query term. Term Frequency Inverse Document Frequency (TF-IDF) to determine what words in a collection of documents might be more favorable to use in a query. TF-IDF can find documents that are relevant to user query.

TF-IDF algorithm is applied as K=F(T). Where F is TF-IDF algorithm applied on T which is calculated for input query and set of documents and result is stored in K.

### 3.2. Genetic Algorithm Function

The genetic algorithm for our work is represented as**:** Vo=GA(Vq,Vd, FE), Where Vo is Output of weighted vector. Vq is Query weighted vector, Vd is Document of weighted vector, and FE is Fitness evaluation. For calculating fitness evaluation, we are using matching functions such as cosine coefficient, dice coefficient, jaccard coefficient.

### 3.3. Matching Function

Fitness function is used to measure the performance of solution. It evaluates how solution is good. We also use the fitness functions to calculate the distance between document and query.
We define X = (x1 , x2 , x3 ,….., xn)
| X | = number of terms occur in X. Matching score of fitness function is in between 0 to 1 [1]. 1.0 means document and query is similar [1]. Matching score near 1.0 mean documents and query are more relevant and values near 0.0 mean documents and query are less relevant [1]. Values evaluate from fitness functions are called "fitness". We are using following matching functions [4].

## 4. IR Using Genetic Algorithm

```
Begin
D=text document
K=TF-IDF on keywords
Preprocess (D,K)
Generation=0
P=initialization_population
F=evaluation(P)
While not(required_fitness(F) and not
termination_condition do
      Begin
             Generation=generation+1
             I=selection(P,F)%for genetic operators
P=new_generation(P,I)
F=evaluation(P)
      End
end
```

### 4.1. Working Steps of System

1. User enters query into information retrieval system.
2. Match keywords from user query with list of keywords.
3. Preprocess keywords and text documents. In preprocessing the text document of newspaperare

converted into vectors of weight using IF-IDF approach.
4. Encode documents retrieved by user query to chromosomes (initial population)
5. New Population is generated using genetic operator such as mutation, selection, crossover.
6.Repeat step 5 till generation reaches to maximum value. We will get an effective query chromosome for text document retrieval.
7. Decode optimize query chromosome to query and retrieve newspaper documents from database.

### 4.2. Flow Diagram

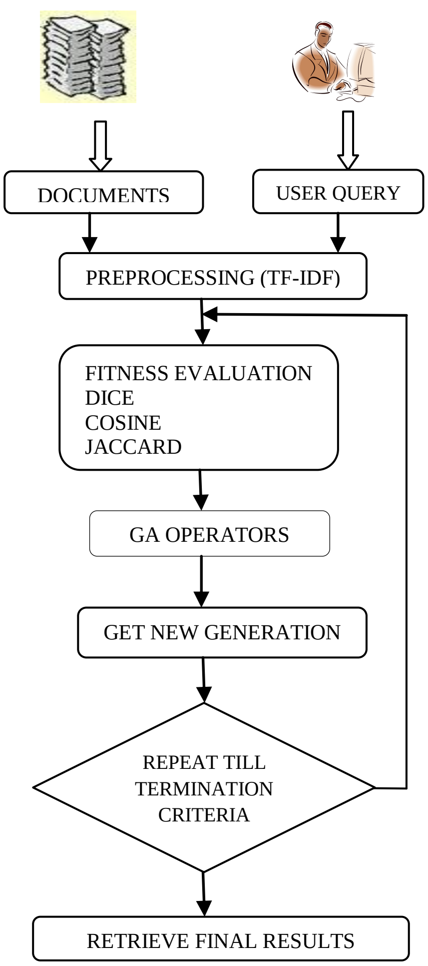


Fig: Flow Propagation

### 4.3. Precision and Recall

Precision and Recall is used for measuring efficiency of information retrieval system [1].Recall is defined as the proportion of relevant document retrieved [1]. It is the fraction of the documents that are relevant to the query that is successfully retrieved.

Recall= $(|\{P\}\cap\{Q\}|)$ / $(|\{P\}|)$
P= relevant documents
Q= retrieved documents
Precision is defined as the proportion of retrieved document that is relevant [1]. It is the fraction of the documents retrieved that are relevant as per the user's query.
Precision= $(|\{P\}\cap\{Q\}|)$ /$(|\{Q\}|)$
P= relevant documents
Q= retrieved documents

## 5. Advantages of Proposed System

The proposed information retrieval system will provide more relevant result and will require less time to search document according to users query.

## 6. Conclusion

The proposed information retrieval system is more efficient within a specific domain as it retrieves more relevant results. The system performance has been verified using the evaluation measures recall and precision.



## 7. References

[1] Bangorn Klabbankoh, Ouen Pinngern "Applied Genetic Algorithms in Information Retrieval", 2010.

[2] Eman Al Mashagba, Feras Al Mashegba, Mohammad Othman Nassar, "Query Optimization Using Genetic Algorithms in the Vector Space Model", September 2011.

[3] Ahmed A. A. Radwan, Bahgat A. Abdel Latef, Abdel Mgeid A. Ali, and Osman A. Sadek," Using Genetic Algorithm to Improve Information Retrieval Systems", World Academy of Science, Engineering and Technology 17 2008.

[4] Mr. Vikas Thada, Mr. Sandeep Joshi."A genetic algorithm approach for improving the average relevancy of retrieved documents using Jaccard Similarity Coefficient", August, 2011.

[5] Huda Yasin, Mohsin Mohammad Yasin, Farah Mohammad Yasin, "Automated Multiple Related Documents Summarization via Jaccard's Coefficient", January 2011.