Importance of Similarity Measures in Effective Web Information Retrieval

Shagun Giridhar
Department of Computer Science
Manav Institute of Technology & Management
Hisar, India
e-mail: shagun.success@gmail.com

Kanika Bhutani
Department of Computer Science
Manav Institute of Technology & Management
Hisar, India
e-mail:kbhutani2012@gmail.com

Abstract— Information Retrieval (IR) manages recovering and showing data inside the WWW and online databases and furthermore looks through the web reports The quick development of site pages accessible on the Internet as of late, seeking applicable and coming data has turned into a pivotal issue. Data recovery is a standout amongst the most essential segments in web crawlers and their improvement would greatly affect enhancing the looking productivity because of dynamic nature of web it turns out to be much hard to discover applicable and late data. That is the reason an ever increasing number of individuals began to utilize centered crawler to get correct data in their uncommon fields today.

The information retrieval field mainly deals with the grouping of similar documents to retrieve required information to the user from huge amount of data. The researchers proposed different types of similarity measures and models in information retrieval to determine the similarity between the texts and for document clustering.

This research intends the study of genetic algorithm based information retrieval using similarity measures like cosine coefficient, jaccard coefficient, dice coefficient.

Keywords*- Information Retrieval, Similarity Measures, Genetic Algorithm; Fitness Function; Crossover; Mutation.*

*****

I. INTRODUCTION

The basic aim of information retrieval is retrieval of most relevant documents for a given user query. Web searches are the perfect example for this application [1]. Many algorithms was developed for this purpose, which take an input query and match it with the stored documents or text snippets and rank the documents based on their similarity score respective to the given query. Such algorithms rely on matching the indexed documents, which maintain the information concerning term frequencies and positions, against the individual query terms. A score is assigned to each document based on its similarity value [2].

Data Retrieval (IR) goes for demonstrating, outlining, and executing frameworks ready to give quick and successful substance based access to a lot of data. The point of an IR framework is to assess the significance of data things, for example, content archives, pictures and video, to a client's data require. Such data require is spoken to as an inquiry, which as a rule relates to a pack of words. The essential objective of an IR framework is to recover all the data things that are important to a client question while recovering as few non-applicable things as could be expected under the circumstances.

II. INFORMATION RETRIEVAL SYSTEM

.

Information Retrieval System (IRS) is a system used to store data that is to be processed, searched and retrieved corresponding to a user generated query. Most IRSs use keywords to retrieve documents [2]. First of all , before the retrieval process can be initiated, it is necessary to define the text database and this is done by the database manager which specify the operation to be performed on text and generated model. The text operation transform the original document to a logical view and them index of the text is generated by the database manager because it is a critical data structure and it allows fast searching over large numbers of data

Components of Information Retrieval System The basic component of Information retrieval system. They are User, Documentary Database, Query Subsystem and matching function.

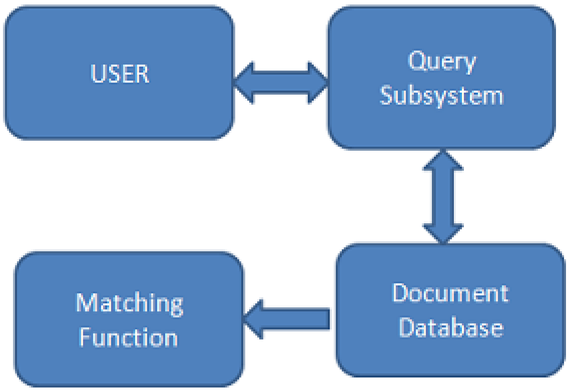

Fig:1 Basic Model of IRS

1) *User:-* User is a person who put the request on the information retrieval system. On the bases of this request information is retrieved from the database.

2) *Query subsystem:-*An input is submitted by the user and corresponding data is retrieved by the system.

3) *Document database: -* It is the storage room where every one of the records are put away. Alongside reports it additionally speaks to their data content. Coordinating function think about every one of the archives of report

database with the client inquiry and concentrate important record from database [3].

4) *Matching function:* - A novel data retrieval calculations utilizing hereditary calculation to expand the execution of data retrieval framework. The novel coordinating capacities called Overall Matching Function (OMF) and Virtual Center based Matching Function (VCF) are proposed for enhancing the retrieval execution. By and large Matching Function gives the outcomes by finding the normal of coordinating scores from established coordinating capacities and VCF depends on finding the virtual focus from the arrangement of centroids exhibit in grouping space. VCF based Genetic Algorithm (VCGA) are utilized for data recovery. Working of both the coordinating capacities is contrasted with check the execution.

III. LITERATURE REVIEW

Data on the Web is available as content records (organized in HTML), and that is the reason many Web Document processing systems are worked in data mining approaches [4]. Because of the development of data in web prompts radical increment in field of data recovery. Effective data retrieval and navigation is given by record clustering. Record clustering is the procedure of naturally gathering the related reports into groups. Rather than scanning whole records for important data, these cluster will enhance the effectiveness and redundancy of data. Relevant data can be proficiently recovered and gotten by methods for document clustering.

IV. CHALLENGING ISSUES IN WEB CLUSTERING BASED ON SIMILARITY

The World Wide Web is very large, widely spreaded , global information service centre in this retrieving exact information for users in Search Engine faces lots of trouble.

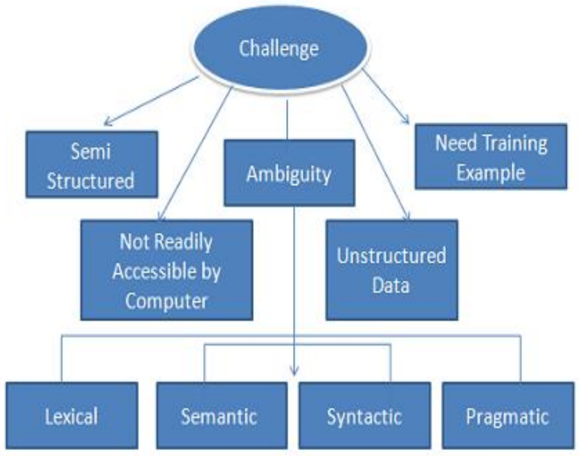

Fig 2.Challenging Issues in Web Document Clustering

This is because of accurately measuring the semantic and syntactic similarity between words is an important problem, such as word sense disambiguation, textual entailment, and automatic text summarization. In information retrieval, one of the major problems is to retrieve a set of documents that is semantically similar to a given user query. Retrieving accurate information to users to these kind of realated words is very challenging. Some previously designed system proposed an model and method to measure semantic similarity between words, which consists of snippets, page-counts and two class support vector machine.

V. SIMILARITY MEASURES

Similarity Measures is a function that is used to measure the amount of similarity between query and documents. It measures how much the query and document is similar to each other. It generate a value which decides the level of similarity.

Clustering is a powerful system that compose a substantial amount of unordered text document into small important and intelligent records. Exact grouping requires an exact meaning of the closeness between a couple of items ,in terms of the pair wise similarity or distance.

A wide variety of similarity or distance measure have been proposed and frequently applied. Some of important similarity measures are:

- Euclidian Distance (ED)
- Cosine Similarity Measure (CSM)
- Jaccard Similarity Measure (JSM)
- Dice Coefficient Measure (DCM)

A. *Euclidian Distance (ED)*

Euclidian distance measure is an ordinary distance measure to compute the distance between two points in two and three dimensional space. This measure is used in document clustering to group the documents into similar clusters based on the distance between the documents [9]. Euclidian Distance measure is represented in equation

$$\mathrm{ED}(d_x,d_y)=\sqrt{\sum_{i=1}^{m}(w(t_i,d_x)\text{-}\mathrm{w}(t_i,d_y))}\ \ldots\ldots\ldots\ldots(1)$$

$\mathrm{ED}(d_x,d_y)$ is the Euclidian Distance between d_x and d_y documents. $(t_1, t_2, \ldots, t_m)$ is the set of terms, $\mathrm{w}(t_i, d_x)$, $\mathrm{w}(t_i, d_y)$ is the weights of term t_i in document x and y respectively.

B. *Cosine Similarity Measure (CSM)*

Cosine similarity measure computes similarity as a function of the angle made by the vectors. If two vectors are close, the angle formed between them would be small and if the two vectors are distant, the angle formed between them would be large [10].

The cosine value varies from +1 to -1 for angles ranging from 0 to 180 degrees respectively, making it the ideal choice for these requirements. A score of 1 evaluates to the angle being 0 degree, which means the document are similar. While a score of 0 evaluates to the angle being 90 degree, which means the documents are entirely dissimilar. The cosine weighting measure is implemented on length normalized vectors for making their weights comparable. Equation (2) gives the formula for Cosine Similarity.

$$\text{CSIM}(q,d_j)=\frac{\sum_{i=1}^{m} w(t_i,q)\times(t_i,d_j)}{\sqrt{\sum_{i=1}^{m}(t_i,q)^2}\times\sqrt{\sum_{i=1}^{m}(t_i,d_j)^2}}\text{........}(2)$$

Where, w(t_i,q), w(t_i, d_j) are the weights of the term t_i in query q and document d_j respectively.

C. Jaccard Similarity Measure (JSM)

Jaccard Similarity measure is another measure for calculating the similarity between the queries and documents [11]. In this measure, the index starts with a minimum value of 0 (completely dissimilar) and goes to a maximum value of 1 (completely similar).

Jaccard similarity measure measures similarity between the two documents. The value is between 0 and 1. 0 show that documents are dissimilar and 1 shows those documents are identical with each other. Value between 0 and 1 show the probability of similarity between the documents. Equation (3) represents the Jaccard Similarity measure

$$\text{JSIM}(q,d_j)=\frac{\sum_{i=1}^{m} w(t_i,q)\times(t_i,d_j)}{\sqrt{\sum_{i=1}^{m}(t_i,q)^2}+\sqrt{\sum_{i=1}^{m}(t_i,d_j)^2}-\sum_{i=1}^{m} w(t_i,q)\times(t_i,d_j)}.(3)$$

Where, w(t_i,q), w(t_i, d_j) are the weights of the term t_i in query q and document d_j respectively.

D. Dice Coefficient Measure (DCM)

Dice Coefficient measure is used to compare the similarity between two samples of text [14]. Equation (4) shows the Dice Coefficient measure.

$$\text{DSIM}(q,d_j)=\frac{\sum_{i=1}^{m} w(t_i,q)\times(t_i,d_j)}{\alpha\sum_{i=1}^{m}(t_i,q)^2+(1-\alpha)\sum_{i=1}^{m}(t_i,d_j)^2}\text{.........}(4)$$

Where, w(t_i , q), w(t_i, d_j) are the weights of the term t_i in query q and document dj respectively. α Parameter range is from 0 to 1. α control the magnitude of penalties of false negative versus false positive errors. In general Alpha value is 0.5. if α > 0.5, DCM measure gives more significance to precision and if α < 0.5, this measure gives more significance to recall.

VI. GENETIC ALGORITHM

A genetic algorithm is a adaptive heuristic search algorithm based on the idea of natural selection and genetics[2]. The operation of the genetic algorithm is very simple. They represent an intelligent exploitation of a random search used to solve optimized problem .It is better than conventional AI and it is more robust [12,13].

GAs simulate the survival of the fittest among individuals over consecutive generation for solving a problem. Each generation consist of population.

Genetic Algorithm basic components are:

A. Chromosome Representation

Chromosomes are the basic input applied to GA . All the data and query are converted into chromosome and supplied as an input to the genetic algorithm. The chromosome is represented in the form of binary values. It uses fixed-length binary strings to represent chromosome, where each position corresponds to a query term.

B. Fitness Function

Fitness Function gives a value which is used to calculate the similarity between query and document. Based on this value chromosome is selected for selection mechanism. The similarity function works as a mapping between both of them. The most popular similarity measure is cosine similarity measure which is cosine of the angle between document vector and query vector, and respond with high results. Other popular similarity measures are Dice coefficient and Jaccard coefficient.

C. Selection operator

Selection is the process in which chromosomes is selected for next step or generation in genetic algorithm based on fitness value of chromosomes. During each successive generation a portion of the existing population is selected to generate a new population. Poor chromosome or lowest fitness chromosome selected very few or rejected.

D. Crossover operator

Crossover is one of the basic operators of Genetic algorithm. The performance of GA depends on them. In crossover two or more parent chromosomes is selected and a pair of genes are interchanging with each other.

Different crossovers have been employed in GA algorithm by researchers.

- *Single-point crossover*
- *Two-point* and *Multipoint crossover*

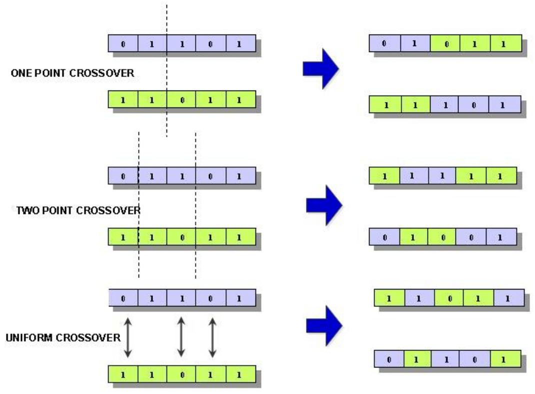

Fig 3:Crosover of Chromosomes

E. Mutation operator

Mutation is a process in which gene of the chromosome is changed. In one point mutation if gene is 0 then change it into 1 and if gene is 1 then change it into 0.

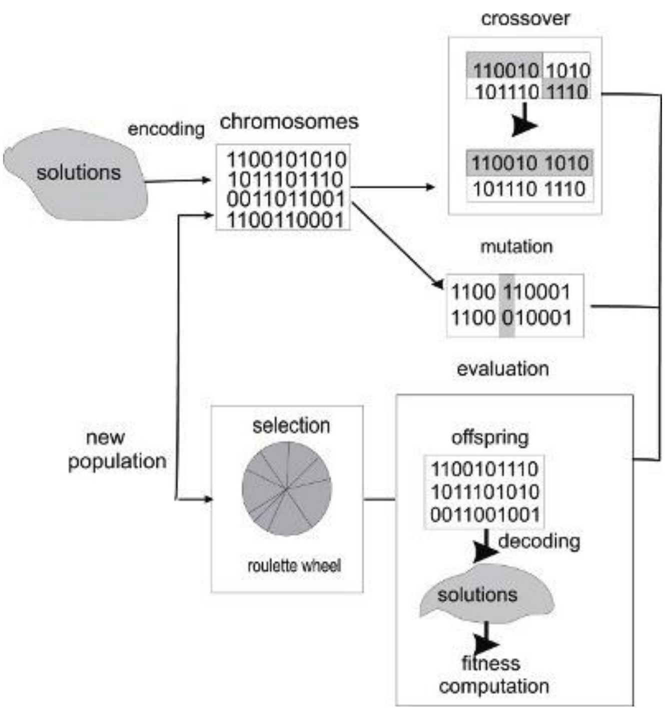

Fig:3 General Structure of Genetic Algorithm

VII. Experiment & Result

This experimentation is done on 10 queries with fitness function jaccard coefficient, dice and cosine. Let's take an example of first query. Process of experiment is done as follows:

- Enter Query
- Extraction of Keywords with Textalyser Tool
- Encoding of Chromosomes
- Calculation of average relevancy with old keywords
- Applying Genetic Algorithm Operators
- Result of Genetic Algorithm
- Decode Optimized Query Chromosome to Query
- Calculation of Average Relevancy with New Keyword.

The effect of different cross over and mutation probability on chromosome average relevancy in the form of generation. At Pm=0.1, (Pc=0.7, Pc=0.8, Pc=0.9) no query converge at one chromosome. At Pm=0.05, (Pc=0.7, Pc=0.8, Pc=0.9) no query converge at one but this gives more relevance than Pm=0.1. At Pm=0.01 all query converge at one in less number of generation. All this shows that less mutation rate is best for queries.

It is concluded that Information Retrieval having genetic algorithm gives more relevant results as compared to classical Information Retrieval System. Efficiency of Informational Retrieval increased after applying genetic algorithm.

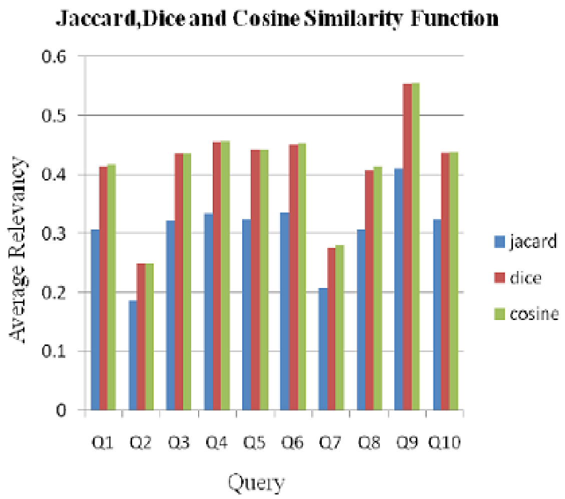

Fig :4 Graph of Old keywords of Jaccard, Dice and Cosine Similarity Function

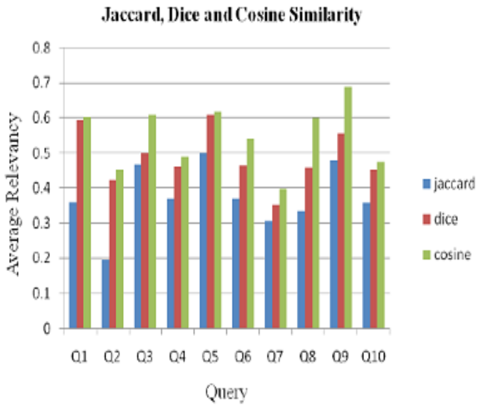

Fig :5 Graph of New keywords of Jaccard, Dice and Cosine Similarity Function

Fig.4 shows the graph of old keywords having jaccard, dice and cosine similarity function before applying genetic algorithm. Cosine function gives slightly better relevancy which is nearby equal to dice function and relevancy of dice function is high as compared to jaccard function. Fig.5 shows the comparisons of jaccard, dice and cosine similarity function after applying genetic algorithms.

VIII. Conclusion

This paper deals with the fundamental and genetic algorithm. We have defined various methods and process to find solution of similarity arises during a search query. GAs starts with a limited numbers of individuals from initial population and generate new population by selection, crossover and mutation.

Fitness function plays an important role to find out similarity measures between the query and document.

IX. REFRENCES

[1] Impact of Similarity Measures in Information Retrieval ISSN (e): 2250 – 3005 || Volume, 08 || Issue, 6|| Jun – 2018.

[2] Effective Information Retrieval Using Similarity Function: Horng and Yeh Coefficient- IJARCSSE Volume 3, Issue 8, August 2013 ISSN: 2277 128X.

[3] An Overview of Genetic Algorithm Based Information Retrieval International Journal on Recent and Innovation Trends in Computing and Communication ISSN: 2321-8169. Volume: 5 Issue: 5 652 – 655.

[4] G.Thilagavathi, J.Anitha, K.Nethra,"Sentence Similarity Based Document Clustering Using Fuzzy Algorithm", International Journal of Advance Foundation and Research in Computer (IJAFRC) Volume 1, Issue 3, March 2014. ISSN 2348 – 4853.

[5] Ann Gledson & John Keane, (2008) "Using Web-Search Results to Measure Word-Group Similarity", 22nd International Conference on Computational Linguistics), pp. 281-28.

[6] Hughes T & Ramage D (2007) "Lexical Semantic Relatedness with Random Graph Walks", Conference on Empirical Methods in Natural Language Processing Conference on Computational Natural Language Learning, (EMNLP-CoNLL07), pp. 581-589.

[7] Lin D, (1998) "An Information-Theoretic Definition of Similarity", 15th International Conference on Machine Learning, pp. 296-304.

[8] P.Ilakiya, "Discovering Semantic Similarity between Words Using Web Document and Context Aware Semantic Association Ranking" ,International Journal of Advanced Research in Computer Engineering & Technology (IJARCET) ,Volume 2, Issue 6, June 2013.

[9] Komal Maher, Madhuri S. Joshi, "Effectiveness of Different Similarity Measures for Text Classification and Clustering",International Journal of Computer Science and Information Technologies, Vol. 7, No.4, pp.1715-1720, 2016.

[10] Moheb Ramzy Girgis, Abdelmgeid Amin Aly & Fatima Mohy Eldin Azzam, "The Effect Of Similarity Measures On Genetic Algorithm-Based Information Retrieval", International Journal of Computer Science Engineering and Information Technology Research, Vol. 4, Issue 5, pp. 91-100, Oct 2014.

[11] Abhishek Jain, Aman Jain, Nihal Chauhan, Vikrant Singh, Narina Thakur , "Information Retrieval using Cosine and Jaccard Similarity Measures in Vector Space Model", International Journal of Computer Applications, Volume 164, No 6, PP.28-30, 2017.

[12] A Detailed Study on Information Retrieval using Genetic Algorithm Journal of Industrial and Intelligent Information Vol. 1, No. 3, September 2013.

[13] An Overview of Genetic Algorithm Based Information Retrieval International Journal on Recent and Innovation Trends in Computing and Communication ISSN: 2321-8169 Volume: 5 Issue: 5 652 – 655.

[14] E man Al Mashagba , Feras Al Mashagba and Mohammad Othman Nassar, "Query optimization using genetic algorithm in the vector space model", International Journal of Computer Science, vol. 8, no. 3, pp.450-457, Sept. 2011.