Available Online at www.ijcsmc.com

International Journal of Computer Science and Mobile Computing

A Monthly Journal of Computer Science and Information Technology

ISSN 2320–088X

International Conference on Mobility in Computing- ICMiC13, Organized by Mar Baselios College of Engineering and Technology during December 17-18, 2013 at Trivandrum, Kerala, India, pg.80 – 86

SURVEY ARTICLE

Fitness Function in Genetic Algorithm based Information Filtering - A Survey

Shini Renjith[1], Anjali C[2]

[1,2]Department of Computer Science and Engineering, Mar Baselios College of Engineering & Technology, Trivandrum, India

[1]shinirenjith@gmail.com;[2]anjalichandrika@gmail.com

Abstract **–Information filtering systems have a vital role in exploiting the flurry of information that keeps on growing exponentially. This study focuses on improving the performance of genetic algorithms based information filtering. In GA, the fitness function is the performance measure for relevance judgement. Being the backbone of the evaluation process the selection of fitness function is vital in the design of a GA. The secondary data gathered from past studies lead to the conclusion that Cosine Coefficient is the better similarity measure among other alternatives like Euclidean Distance, Dice Coefficient, Jaccard Coefficient and Inner Product.**

Keywords **–Information Filtering; Genetic Algorithm; Similarity Measure; Fitness Function; Vector Space Model**

I. Introduction

Information filtering focuses on the management of overwhelming information, and performs the presentment of information that is highly relevant (interesting or useful) to the user. Content-based filtering is one of the two major approaches used for information filtering. A content-based filtering model works based on the correlation between the user preferences and the content of a given set of objects under consideration. Mostly these objects are documents, but other types of data like image, video, sounds etc. also can be considered for an information filtering system. Further the same concept can be extended to recommendation systems in day to day life such as shopping, travel, entertainment, etc.

Usually statistical approaches are considered to the backbone of information filtering systems. However artificial intelligence models like Genetic Algorithm (GA)can also take up the crucial role here which can improve the performance. GA is one of the evolutionary algorithms based on the principle of 'Survival of the fittest" and mimic the natural process of evolution like reproduction, mutation, recombination and selection to generate solution to problems. It is developed by John Holland [1] of University of Michigan in 1975 and got very popular in late 80's. It is widely used in business, scientific and engineering circles to solve a variety of problems that are not easy to solve using other techniques.

Section II deals with the background study of genetic algorithm. Section III explains the methodology and different similarity measures. The discussion is given in section IV and finally the paper concludes in section V.

II. Antecedents

Genetic algorithms come under the evolutionary algorithms used for content based filtering of information from past user behaviour. The algorithm [2] reproduces the process of natural selection in living organisms. The state space in genetic algorithm constitutes of candidate keys to the queries. A population constitutes the set of a solution. A chromosome represents the string whereas the gene resembles the bit pattern. The Fitness function is a fact based function value of each chromosome for evaluating how good a solution is. A Population obtained after certain iteration is called as generation.

However better generations can be obtained by an effective implementation of the above said algorithm. This includes reproduction, crossover and mutation. Reproduction is a process in which the fittest individual is selected based on the fitness function. Crossover deals with the exchange of genes between two individual chromosomes that are reproducing. Mutation is the arbitrary altering of the genes in a specific chromosome. The two types of mutation include Point mutation where a single gene is altered and Chromosomal mutation where few numbers of genes are altered completely.

A. ENCODING TECHNIQUES

Genetic Algorithms can be effectively implemented using various encoding techniques [3]. Encoding symbols effectively describes a method to encode the potential solution to a problem. Among the different encoding techniques, binary encoding is the most common and has straightforward nature. This encoding scheme is considered to be relatively simple but has a limitation that it is often artificial in nature and needs correction after operations like crossover and mutation. The next method of encoding is permutation encoding which can be used in task ordering problems –a good example is travelling salesman problem. In this method, a string of numbers represents the position of chromosome in a sequence. Operations like crossover and mutation will help in making the chromosome more consistent. Yet another method is the value encoding which can handle more complicated values. Here, every chromosome represents a sequence of values. Tree encoding can be used for expressions or evolving programs in genetic programming. Here every chromosome is considered as a tree of objects. Tree encoding is useful for evolving programs which can be encoded to trees – like Programing language LISP.

B. INFORMATION FILTERINGSYSTEM

Information filtering system screens and stores information after processing, searching and retrieval of data based on the circumstances [4]. Three specific components are prerequisites for driving the filtration process. The first component is the population and its initialization is the starting point for the information filtering process. The initial population is selected based on a trainer sample. The second constituent is the query subsystem where a person articulates his desires through queries formulated using some query language. The last and most important component is the selection process which evaluates the level of match / similarity score of items qualified by the query. The query is iteratively refined over generations.

C. INFORMATION FILTERINGMODELS

The most prevailing information retrieval paradigms are Boolean, Vector space and Probabilistic models [5]. In the Boolean model, user queries are expressed using a query language along with the logical operators. The outcome is either the set of items that totally match with the query or irrelevant for the user. In Vector space model an item is represented as a vector in n-dimensional space. Here n is the number of exclusive attributes which explains the item in a collection and each attribute corresponds to one dimension in the document space. The query also can be represented in the similar manner. The retrieval process is based on the similarity measures between the query and the items. The documents with a higher fitness to the query are considered to be more relevant. Probabilistic model is based on the estimated probability of relevance of an item to the user.

III. Methodology

Genetic Algorithms follow iterative process to refine the population of possible results by continuous evolution through a fitness function that ranks the solutions. The best solutions are retained and the worst ones are removed as the iteration continues. In GA, the population comprises of potential solutions to the problem/ query which is represented by the term chromosome. Each of these chromosomes is associated with an objective function value - the fitness. A generation is the population at a specific iteration of the genetic algorithm.Fig.1shows the pseudo code and flowchart for genetic algorithm.

Vector Space Model (VSM) is used for representing any objects as vectors in a multi-dimensional space. Each dimension corresponds to a term/ attribute that are used to represent the document / item under consideration. The retrieval process is based on the correlation between the documents/ items and the query. The highly relevant documents with respect to the query– the ones with a higher similarity - are considered as more relevant and should get positioned first in the result set.

Fitness Function/Similarity Measure are the objective function that computes the degree of similarity between the query and the document. If the query and the document do not have any attribute/term in common, then similarity score should be very low. There are multiple similarity measures that are available to find the correlation between query and document. This paper attempts to perform comparison between these similarity metrics.

```
{
    Initialize Population
    Evaluate Population
    While Termination_Condition_Not_Met
    {
        Select Parents for Reproduction
        Apply Crossover/ Recombination
        Apply Mutation
        Replace the Population by the new generation of
      individuals
        Evaluate Population
    }
}
```

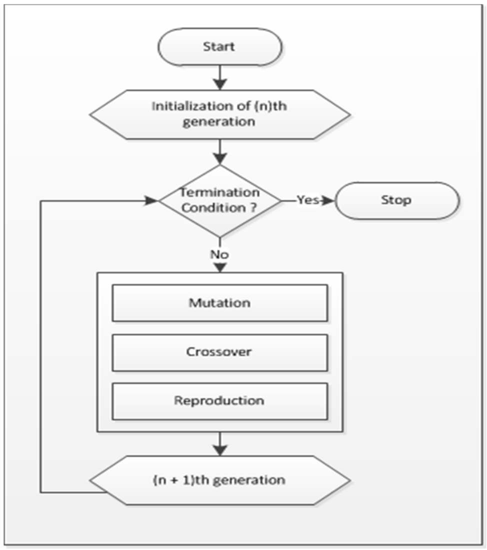

Fig. 1A Simple Genetic Algorithm

A. EUCLIDEAN DISTANCE

The simplest method of measuring similarity is by using the Euclidean distance [6], where the common attributes of the objects are being compared. The similarity between two objects is calculated as the sum of all differences (between of each term/ attribute of the objects) is squared. The drawback of this model is the two dimensional nature of the Euclidean distance. Mathematically the similarity between A, the Document vector and B, the Query vector, is represented as

$$\mathrm{ES} = \sqrt{\sum_{i=1}^{n} (Ai\text{-}Bi)^2}$$

B. COSINE COEFFICIENT

In cosine coefficient [6], the attributes/terms are represented in vector space model to calculate the normalized dot product of two documents. By ascertaining the cosine similarity, the user can find the cosine of the angle between the two objects. For cosine similarities with a value of 0, the documents do not have any common attributes (or words) and the angle between the objects is 90 degrees in the vector space model. Mathematically the same can be represented as

$$\cos\theta = \frac{|A \cap B|}{|A|^{1/2}.|B|^{1/2}} = \frac{\sum_{i=1}^{n} Ai.Bi}{\sqrt{\sum_{i=1}^{n}(Ai)^2}\ .\sqrt{\sum_{i=1}^{n}(Bi)^2}}$$

where A is the Document vector and B is the Query vector

C. DICE COEFFICIENT

Also known as Sorensen Index, Dice coefficient [6] is an indicator used for measuring the similarity between two objects. When A and B are the Document vector and Query Vector, respectively; the quotient of similarity, QS ranges from 0 to 1 and is mathematically expressed as below

$$\mathrm{DS} = \frac{2|\mathrm{A} \cap \mathrm{B}|}{|\mathrm{A}| + |\mathrm{B}|} = \frac{2\sum_{i=1}^{n} \mathrm{Ai.Bi}}{\sum_{i=1}^{n}(\mathrm{Ai})^2 + \sum_{i=1}^{n}(\mathrm{Bi})^2}$$

D. JACCARDCOEFFICIENT

When negative values do not give any relevance, Jaccard coefficient [6] can be used as a similarity measure. Here each attribute/ term is considered as binary - each bit represents the absence or presence of a characteristic. In this the similarity can be measured via the overlap, or intersection, of the sets. Mathematically the similarity ratio, SR can be expressed as below

$$\mathrm{SR} = \frac{|\mathrm{A} \cap \mathrm{B}|}{|\mathrm{A}| + |\mathrm{B}| - |\mathrm{A} \cap \mathrm{B}|} = \frac{\sum_{i=1}^{n} \mathrm{Ai.Bi}}{\sum_{i=1}^{n}(\mathrm{Ai})^2 + \sum_{i=1}^{n}(\mathrm{Bi})^2 - \sum_{i=1}^{n} \mathrm{Ai.Bi}}$$

E. INNER PRODUCT

Inner product or dot product or scalar product [6] takes two equal-length sequences of attributes (coordinate vectors) and returns a single attribute. It is the cross product of two vectors, which produces a pseudo vector as the result.

$$\mathrm{Sim} = |\mathrm{A} \cap \mathrm{B}| = \sum_{i=1}^{n} \mathrm{Ai.Bi}$$

IV. Discussion

The efficiency of information retrieval can be measured in terms of recall and precision. Recall is defined as ratio of the number of relevant documents retrieved over the total number of relevant documents in the population. Precision is defined as ratioof the number of relevant documents retrieved over the total number of documents retrieved.

$$Recall = \frac{Number\ of\ relevant\ documents\ retrieved}{Total\ number\ of relevant\ documents\ in\ population}$$

$$Precision = \frac{Number\ of\ relevant\ documents\ retrieved}{Total\ number\ of\ documents\ retrieved}$$

Given below the secondary data collected from the experimental research conducted by Safaa I. Hajeer in 2012 using a collection of 800 English documents from National Physics Laboratory, UK and tested using 20 queries as the sample [6]. The details are summarized in Fig.2.

Similarity Measure	Recall	Precision
Cosine Coefficient	0.650	0.250
Inner Product	0.643	0.236
Jaccard Coefficient	0.618	0.204
Euclidean Distance	0.600	0.200
Dice Coefficient	0.571	0.180

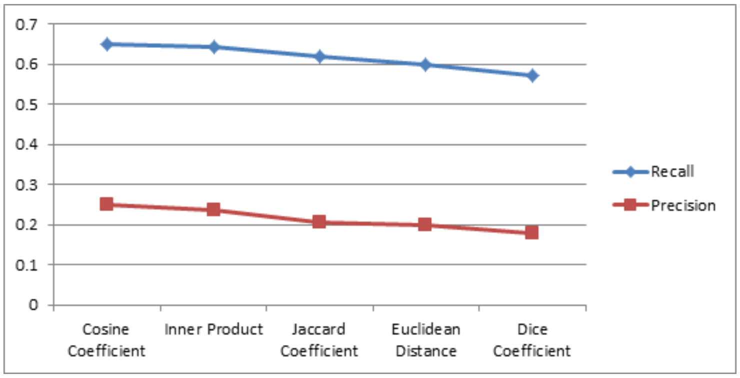

Fig.2Summary of Accuracy Measurement for English documents

The best results are achieved using Cosine Similarity measure while comparing against the other four measures, with 65% recall and25% precision. Next one is the Inner Product having 64.3% recalland23.6% precision. The third measure is Jaccard Coefficient with 61.8% recall and 20.4% precision. The fourth is Euclidean distance with 60% recall and 20% precision. The last one is Dice Coefficient with 57.1% recall and 18% precision.

A similar study is performed using Arabic documents [6] as well and the results are as given in Fig.3.

Similarity Measure	Recall	Precision
Cosine Coefficient	0.532	0.380
Inner Product	0.530	0.376
Jaccard Coefficient	0.526	0.360
Euclidean Distance	0.525	0.357
Dice Coefficient	0.522	0.356

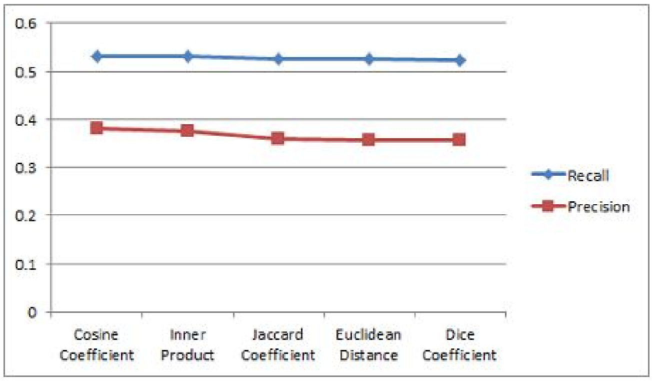

Fig.3Summary of Accuracy Measurement for Arabic documents

The best results are achieved using Cosine Similarity measure while comparing against the other four measures, with 53.2% recall and38% precision. Next one is the Inner Product having53% recalland37.6% precision. The third measure is Jaccard Coefficient with 52.6% recall and 36% precision. The fourth is Euclidean distance with 52.5% recall and 35.7% precision. The last one is Dice Coefficient with 52.2% recall and 35.6% precision.

From the experimental research conducted by Eman Al Mashagba et.al, in 2011 using a collection of 242Arabic abstracts collected from the Proceedings of the Saudi Arabian National Conference and tested using59 queries as the sample, it is found that Inner Product is comparatively better than Dice Coefficient in vector space model [7]. This is also in line with the other researches covered under this study.

Studies conducted by VikasThada et.al, in 2013 used 10 different generations of 10 distinct queries on the varying number of pages from Google search results [8]. The research results confirmed that the best fitness values are achieved while using the Cosine similarity coefficients and then by Dice and Jaccard(as shown inFig.4).

Query	Cosine	Jaccard	Dice
Query 1	0.498964	0.392462	0.245556
Query 2	0.500266	0.414344	0.254118
Query 3	0.43258	0.308178	0.190834
Query 4	0.478768	0.417308	0.257926
Query 5	0.549814	0.453104	0.288336
Query 6	0.510478	0.372908	0.267144
Query 7	0.517844	0.381148	0.263516
Query 8	0.513668	0.392446	0.240868
Query 9	0.464326	0.397308	0.23856
Query 10	0.492134	0.387372	0.247892

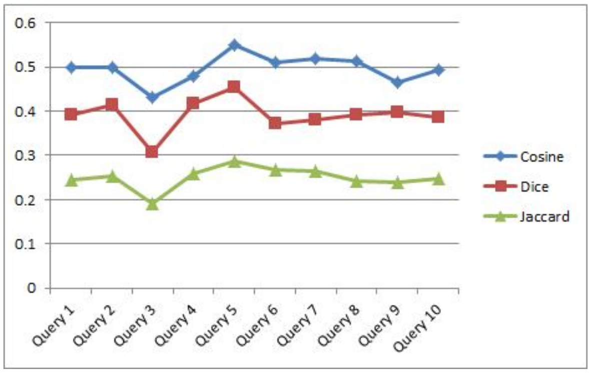

Fig.4Summary of the research done by VikasThada et.al

Yet another experimental study conducted by Mohammad Othman Nassaret.al, in 2013using a collection of 242Arabic abstracts collected from the Proceedings of the Saudi Arabian National Conference and tested using59 queries as the sample, it is found that Inner Product is comparatively better than Dice Coefficient followed by Jaccard Coefficient in vector space model [9] (shown inFig.5). However Cosine Coefficient found to be having the worst efficiency as per this study which is a controversy while comparing with other researches.

Recall	Cosine	Jaccard	Dice	Inner Product
0.1	0.165	0.141	0.141	0.146
0.2	0.164	0.199	0.197	0.208
0.3	0.182	0.288	0.298	0.301
0.4	0.166	0.277	0.277	0.283
0.5	0.179	0.387	0.402	0.405
0.6	0.191	0.401	0.408	0.409
0.7	0.193	0.398	0.396	0.413
0.8	0.244	0.415	0.412	0.437
0.9	0.251	0.467	0.441	0.487

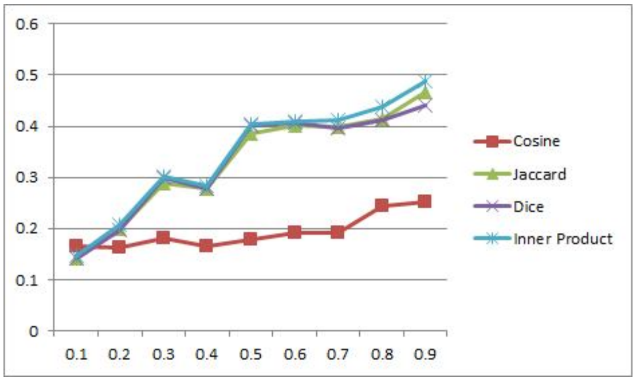

Fig.5Summary of the research done by Mohammad Othman Nassar et.al

V. Conclusion

The major challenge in information filtering process is the determination of the relevance of an item based on the user query. Information filtering using genetic algorithms follows an iterative process to produce the next generation solutions by continuous evolution. A fitness function or similarity measure is used to rank the solutions in each generation. This study considered five different similarity measures for comparison - Euclidean Distance, Cosine Coefficient, Dice Coefficient, Jaccard Coefficient and Inner Product. The studies on the secondary data from past researches lead to the conclusion that Cosine Coefficient is the most efficient. Precision and recall are taken as the measures for evaluating the efficiency. As a future work, the implementation of a GA based recommender system is planned using cosine coefficient as the fitness function.

Acknowledgment

We acknowledge support from the Department of Computer Science and Engineering, Mar Baselios College of Engineering and Technology for all guidance, reviews, valuable suggestions and very useful discussions. Also we would like to thank Mr. Renjith Ranganadhan for his motivation and the encouragements afforded.

References

[1] John H. Holland, *Adaptation in Natural and Artificial Systems*, 1st ed., University of Michigan Press, Ann Arbor, 1975.

[2] BangornKlabbankoh and OuenPinngern, “Applied Genetic Algorithms in Information Retrieval,” in Proc. IEEE vol 92(4), 2004, p. 702-711.

[3] Rakesh Kumar and Jyotishree, “Novel Encoding Scheme in Genetic Algorithms for Better Fitness,” *International Journal of Engineering and Advanced Technology (IJEAT)*, vol. 1(6), pp. 214–218, Aug. 2012.

[4] Ahmed A. A. Radwan, Bahgat A. Abdel Latef, Abdel Mgeid A. Ali and Osman A. Sadek, “Using Genetic Algorithm to Improve Information Retrieval Systems,” in *Proc. World Academy of Science*, *Engineering And Technology, Vol 17*, 2006, p. 6-12.

[5] Abdelmgeid A. Aly, “Applying Genetic Algorithm in Query Improvement Problem,” *International Journal "Information Technologies and Knowledge*, vol.1, pp. 309–316, 2007.

[6] Safaa I. Hajeer, “Comparison on the Effectiveness of Different Statistical Similarity Measures,” *International Journal of Computer Applications (0975 – 8887)*, vol. 53(8), pp. 14–19, Sep. 2012.

[7] Eman Al Mashagba, Feras Al Mashagba and Mohammad Othman Nassar, “Query Optimization Using Genetic Algorithms in the Vector Space Model,” *IJCSI International Journal of Computer Science Issues*, vol. 8(5.3), pp. 450–457, Sep. 2011.

[8] VikasThada and VivekJaglan, “Comparison of Jaccard, Dice, Cosine Similarity Coefficient To Find Best Fitness Value for Web Retrieved Documents Using Genetic Algorithm,” *International Journal of Innovations in Engineering and Technology (IJIET)*, vol. 2(4), pp. 202–205, Aug. 2013.

[9] Mohammad Othman Nassar, Feras Fares Al Mashagba andEman Fares Al Mashagba, “Investigating Genetic algorithms to optimize the user query in the vector space model,” *Australian Journal of Basic and Applied Sciences*, vol. 7(2), pp. 47–53, Feb. 2013.