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

IJCSMC, Vol. 4, Issue. 12, December 2015, pg.255 – 260

Survey: Genetic Algorithm Based Cosine Similarity

Aniket M. Akarte[1], Prof. Mrs. Pradnya V. Kulkarni[2]

[1]Dept. of Computer Engineering
MAEER'S MIT Pune, India
[2]Dept. of Computer Engineering
MAEER'S MIT Pune, India
[1] a1.aniket@hotmail.com; [2] pradnya.kulkarni@mitpune.edu.in

Abstract— Typically large amount of information is present on the webpages rather than the main information which is actually useful to the user. so there is need to separate men content block from other content for that we use algorithm based on the content structure tree(CST) , which part of CST is more important and which part of the CST is less important according to the information required to the user is decided by the cosine similarity and the search based genetic algorithm measure in the fallowing paper is about detailed working of the cosine similarity and the different type of cosine similarity measure along with the genetic algorithm.

Keywords— Genetic Algorithm, Cosine Similarity, Fitness Function, Web Content Mining.

I. INTRODUCTION

The growth of the World Wide Web has leads to a massive increase in the amount of information. As more and more information becomes available on the web, the data on the web is likely to have an exponential growth. In this situation, the retrieval of documents relevant to the user request is of utmost importance. One of the ways to find the relevancy is to calculate the similarity of the user query with the retrieved documents. The cosine similarity function is one of the most popular similarity functions for handling web data. Genetic Algorithms have wide range of applications in search and optimization problems. The application of genetic Algorithm to Information Retrieval holds interesting promises in Information Retrieval and the paper is an attempt in this direction

II. GENETIC ALGORITHM

Genetic algorithm is the fast optimization technique to look for nearly best solution according to the fitness criteria so it avoid local optima and surah for global fitness and genetic algorithm optimization technique based on Darwin's principle of natural selection, the genetic algorithm begins with the population of the randomly

garneted structure where each structure encodes the solution to the suntan task it preside to evolves generation, during each generation the genetic algorithm improves structure of its current population by performing selection followed by crossover or mutation ,genetic algorithm is differ from traditional search optimization method in three significant point

1) They search parallel from a population of point there for it has the ability to avoid being trapped in local optimal solution like traditional method

2) Genetic algorithm work on the chromosome which is an the encoded version of potential solutions parameters rather than optimizing parameters themselves

3) Genetic algorithm use fitness score which is obtained from objective functions without other artificial black box mathematics

The initial population is usually represented as a number of individuals called chromosomes. The goal is to obtain a set of qualified chromosomes after some generations. The quality of a chromosome is measured by a fitness function. Each generation produces new children by applying genetic crossover and mutation operators. Usually, the process ends while two consecutive generations do not produce a significant fitness improvement or terminates after producing a certain number of new generations. Working process of genetic algorithm can be explain as following flow chart:

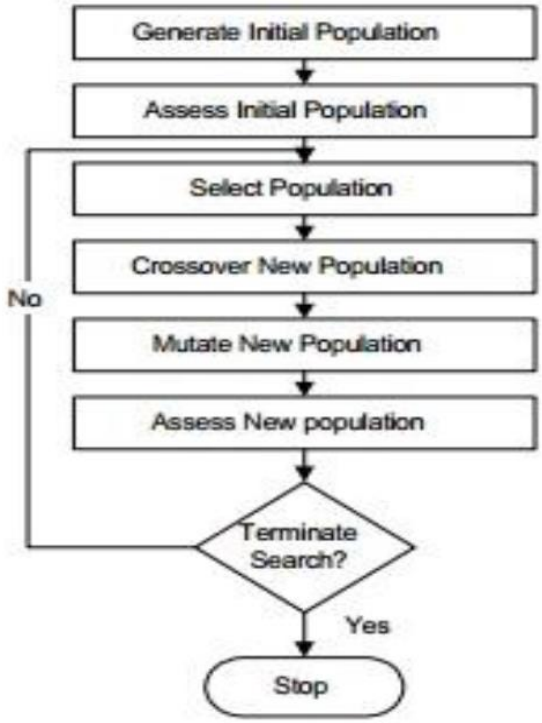

Fig1: genetic algorithm flowchart

A. Fitness Function :

In 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. Genetic algorithms come under the evolutionary algorithms used for content based filtering of information from past user behavior. The algorithm 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. 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.

Moreover, the fitness function must not only correlate closely with the designer's goal, it must also be computed quickly. Speed of execution is very important, as a typical genetic algorithm must be iterated many

times in order to produce a usable result for a non-trivial problem. Fitness approximation may be appropriate, especially in the following cases

- Fitness computation time of a single solution is extremely high
- Precise model for fitness computation is missing
- The fitness function is uncertain or noisy.

Two main classes of fitness functions exist: one where the fitness function does not change, as in optimizing a fixed function or testing with a fixed set of test cases; and one where the fitness function is mutable, as in niche differentiation or co-evolving the set of test cases. Another way of looking at fitness functions is in terms of a fitness landscape, which shows the fitness for each possible chromosome.

Fitness function work with cosine similarity measure the term of information retrieval of the for selecting most useful document is 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 ratio of the number of relevant documents retrieved over the total number of documents retrieved.

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

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

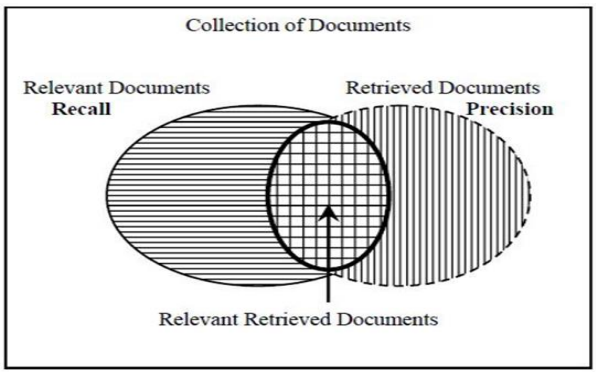

Fig2: Document Retrieval

III. COSINE SIMILARITY

Cosine similarity is the measure of similarity between tow vectors in an inner products it measures the cosine of the angle between them. the cosine of zero degree is one and it is less than one for any other angle.it is thus the judgment of orientation but not the magnitude the two vectors with the same orientation have cosine similarity of one ,two vectors at 90 degree have similarity zero and two vector which are diametrically opposite have similarity of -1 independent of their magnitude cosine similarity is particularly use for the positive space where there outcome is neatly bounded in 0 and 1 ,the cosine similarity is basically use in high dimensional positive spaces such as information retrieval ,data and text mining each term is initially assigned to different dimension under document is characterize by vector with value of each dimension corresponds to the number of times that term appear in the document. The cosine similarity that gives the useful measure of how tow document are similar likely to be in terms of the subject matter that technique is also use to measure the collision with in cluster in the filled of data mining, cosine distance is often use as compliment in positive space it important to note however this is not the proper distance matrix that it does not have the triangle inequality property violate the coincidence axiom to repair the triangle inequality property want's to mean ten same ordering it is necessary to convert angular distance. one of the resin of popularity of cosine similarity that it is very efficient to evaluate, especially for space vector as in the non-zero dimension are need to be consider "cosine of two vectors can be divided by Euclidean dot product formula given two vectors as an attribute in that zero indicating the independent document and intermediate value shows the similarity or dissimilarity of the document "in the text

matching operation the attribute vectors a & b are the frequency vectors of the document. Cosine similarity can be seen as the method of normalizing document link during comparison.

IV. COSINE SIMILARITY WITH GANENTIC ALGRITHM

The growth of the World Wide Web has led to a massive increase in the amount of information. As web is likely to have an exponential growth. In this situation, the retrieval of documents relevant to the user request is of utmost importance. One of the ways to find the relevancy is to calculate the similarity of the user query with the retrieved documents. The cosine similarity function is one of the most popular similarity functions for handling web data. Genetic Algorithms have wide range of applications in search and optimization problems. The application of genetic Algorithm to Information Retrieval holds interesting promises in Information Retrieval more and more information becomes available on the web, the data on the template is designed so that author affiliations are not repeated each time for multiple authors of the same affiliation. Please keep your affiliations as succinct as possible (for example, do not differentiate among departments of the same organization). This template was designed for two affiliations.

DIFFERENT SIMILARITY MATCHING METHODS

Sno	Author Name	Title	Methods	Advantages	Dis Advantages
1	Alfred V. Aho and Margaret J. Corasick	Efficient String Matching An Aid to Bibliographic Search	Pattern matching algorithm Construction of go to, output and failure functions Time complexity of algorithms	Locates keyword in a text string Directed graph begins at the state 0 Time complexity is large	Substrings may overlap with one another Partially computed output function Failure function stored in one dimensional array
2	Arvind Arasu, Venkatesh Ganti, et al.	Efficient Exact-Set Similarity Joins	Threshold based SSJoin Hamming SSJoin Jaccard SSJoin	Threshold parameter is high Vector representation between two sets Similarity value is 0 or 1	Different similarity sets Dimension is differ Common elements
3	Thomas Bocek, Burkhard Stiller, et al.,	Fast Similarity Search in Large Dictionaries	Edit distance NR|-Grep N-grams and Cosine Similarity	Minimum operations required from one string to one string to another Reverse pattern matching Offline approach	Dictionary size is low Avoids number of searching words in NR-grep method Similarity is shared
4	Kaushik Chakrabarti, Dong Xin, et al.,	An Efficient Filter for Approximate Membership Checking	Pruning condition Filtering by ISH Weighted signatures	Three similarity measures are identified Sub string search is quick Weighted signature is in decreasing order	Lower bound value is not identified String similarity is less Different number of signatures

5	Aristides Gionis, Piotr Indyk, et al.,	Similarity Search in High Dimensions via Hashing	Locality Sensitive Hashing Color Histograms Texture Features	Better run time Dependence on data size To measure the performance	Value is small and there is resort needed One index is not sufficient
6	Daniel Karch,Dennis Luxen,etal.,	Improved Fast Similarity Search in Dictionaries	Preprocessing Space Preprocessing Time Query Performance	String Split Parameter based on query time Ten Times Faster Maximum Distance calculated	Speed is low Does not Store any information's Query time and search space size is average.
7	Amit Singhal	Modern Information Retrieval: A Brief Overview	Vector Space Model Probabilistic Model Inference Network Model	Calculate using the Term Weighting Relevance feedback based on user queries ,Retrieval effectiveness	Boolean systems are less effective Poor stemming Style of phrase generation is not critical

V. ADVANTAGES

- GA increases the relevancy of retrieved documents.
- Cosine Similarity is a powerful way to predict user working and performance.
- Genetic Algorithm has been proposed. The proposed information retrieval system is more efficient within a specific domain as it retrieves more relevant result.
- Genetic algorithm with cosine similarity has been verified using the evaluation measures, precision and recall.
- Cosine Coefficient is the most efficient. Precision and recall are taken as the measures for evaluating the efficiency.
- GA based recommender system is planned using cosine coefficient as the fitness function.

CONCLUSION

- Cosine Coefficient is the most efficient. Precision and recall are taken as the measures for evaluating the efficiency
- Cosine Similarity showed The Most Relevant Page When Compared to the Jaccard Similarity Cosine Similarity retrieved web pages having most of the searched token than the Jaccard Similarity.
- In vector space model, the research compares different genetic algorithm strategies by Calculating evaluation using average recall formula. We noticed that the vector space model with Cosine fitness represent the best strategy

- We compared 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.

REFERENCES

[1] David C. Anastasiu and George Karypis "L2AP: Fast Cosine Similarity Search With Prefix L-2 Norm Bounds" IEEE 2014
[2] Swe Swe Nyein, "Mining Contents in Web Page Using Cosine Similarity", IEEE, Copyright 2011.
[3] Roopak.S, Tony Thomas "A Novel Phishing Page Detection Mechanism Using HTML Source Code Comparison and Cosine Similarity" IEEE 2014
[4] Digvijay B. Gautam, Pradnya V. Kulkarni "Cosine Similarity Measure and Genetic Algorithm for extracting main content from web documents" IRD 2014
[5] J. Usharani, Dr K Iyakutti, "A Genetic Algorithm based on Cosine Similarity for Relevant Document Retrieval" IJERT 2013
[6] Vikas Thada and Dr Vivek Jaglan "Comparison of Jaccard, Dice, Cosine Similarity Coefficient To Find Best Fitness Value for Web Retrieved Documents Using Genetic Algorithm" IJIET 2013
[7] Renjith, Shini, and C. Anjali. "Fitness function in genetic algorithm based information filtering-A survey." International Journal of Computer Science and Mobile Computing, ICMIC13 (2013): 80-86.
[8] Wafa. Maitah, Mamoun. Al-Rababaa and Ghasan. Kannan "Improving The Effectiveness Of Information Retrieval System Using Adaptive Genetic Algorithm" IJCSIT 2013
[9] Alfred V. Aho and Margaret J. Corasick Bell Laboratories, Efficient String Matching An Aid to Bibliographic Search, communications of the ACM, Vol. 18 No.6, June 1975
[10] Amit Chandel, P.C.Nagesh, Suita Sarawagi, Efficient Batch Top-k for Dictionary-based Entity Recognition, Proc. 22nd International Conference Data Engineering., pp.28, 2006.
[11] Amit Singhal, Modern Information Retrieval: A Brief Overview, IEEE Computer Society Technical Committee on Data Engineering, pp 1-9, 2001
[12] C. Li, J. Dong, and J. Chen, "Extraction of Informative Blocks from Web Pages Based on VIPS", 1553-9105/ Copyright January 2010
[13] Wafa. Maitah, Mamoun. Al-Rababaa and Ghasan. Kannan," Improving The Effectiveness Of Information Retrieval System Using Adaptive Genetic Algorithm" IJCSIT 2013
[14] S.Balan, Dr. P.Ponmuthuramalingam," A Survey on String Similarity Matching Search Techniques" IJETCS.

