ISSN:0975-9646

Applying Genetic Algorithm in Text to Matrix Generator

Manish Sharma[1] Mr. Rahul Patel[2],
[1]*PG Scholar, CSE, AITR, Indore,* [2]*Assistant professor, CSE, AITR, Indore*

***Abstract*-This article presents an information retrieval system (IRS) using genetic algorithm to increase the performance and efficiency of Text to Matrix Generator (TMG).This paper presents an extension in previous work of Text to Matrix Generator (TMG). In this paper, we proposed a genetic algorithm approach in Text to Matrix Generator for information retrieval. This experimental result shows an improved version of Text to Matrix Generator (TMG).**

***Keywords*-Information Retrieval (IR), Genetic algorithm (GA), *Text to Matrix Generator (TMG)*, Vector Space Model (VSM).**

I. INTRODUCTION

Information retrieval is generally considered as a subfield of computer science that deals with the representation, storage, and access of information [1]. Information retrieval is concerned with the organization and retrieval of information from large database collections [2]. Information Retrieval (IR) is the process by which a collection of data is represented, stored, and searched for the purpose of knowledge discovery as a response to a user request (query) [3].this process involves various stages initiate with representing data and ending with returning relevant information to the user. Intermediate stage includes filtering, searching, matching and ranking operations.

The main goal of information retrieval system (IRS) is to "finding relevant information or a document that satisfies user information needs". To achieve this goal, IRSs usually implement following processes:

In indexing process the documents are represented in summarized content form.

In filtering process all the stop words and common words are remove.

Searching is the core process of IRS. There are various techniques for retrieving documents that match with users need.

There are two basic measures for assessing the quality of information retrieval [2].

Precision: This is the percentage of retrieved documents that are in fact relevant to the query. Recall: This is the percentage of documents that are relevant to the query and were, in fact, retrieved.

In this paper, we present an approach to improve the performance of text to matrix generator (TMG). Text to Matrix Generator (TMG) is a MATLAB Toolbox that can be used for various Data Mining (DM) and Information Retrieval (IR) tasks.

The structure of this paper is as follows. A brief literature review is presented in Section II, followed by vector space model in section III. Followed by genetic algorithm in Section IV. Followed by proposed method in section V, Finally, Section VI covers conclusions.

II. LITERATURE REVIEW

Bangorn Klabbankoh and Ouen Pinngern [4] analyzed vector space model to boost information retrieval efficiency. In vector space model, IR is based on the similarity measurement between query and documents.

Md. Abu Kausar and Md. Nasar [14] give the details on Information retrieval system using genetic algorithm.

Bangorn Klabbankoh and Ouen Pinngern [15] applied genetic algorithm in information retrieval.

Marıa J. Martın-Bautista and Marıa-Amparo Vila and Henrik Legind Larsen [13] address A Fuzzy Genetic Algorithm Approach to an Adaptive Information Retrieval Agent.

Wafa. Maitah, Mamoun. Al-Rababaa and Ghasan. Kannan [11] address improving the effectiveness of information retrieval system using adaptive genetic algorithm.

Vaclav Snasel, Ajith Abraham et al. [5] Optimize Information Retrieval Using Evolutionary Algorithms and Fuzzy Inference System.

Mohammad Othman Nassar et al. [6] investigate Genetic algorithms to optimize the user query in the vector space model.

S.Siva Sathya and Philomina Simon [7] describe Review on Applicability of Genetic Algorithm to Web Search.

Priya I. Borkar and Leena H. Patil [9] address Web Information Retrieval Using Genetic Algorithm-Particle Swarm Optimization.

S.Siva Sathya and Philomina Simon address [8] A Document Retrieval System with Combination Terms Using Genetic Algorithm.

Mohammad Othman Nassar, Feras Al Mashagba, and Eman Al Mashagba [12] Improving the User Query for the Boolean Model Using Genetic Algorithms

Praveen Pathak Michael Gordon Weiguo Fan [16] address an Effective Information Retrieval using Genetic Algorithms based Matching Functions Adaptation

J. Usharani, and Dr K Iyakutti address [17] A Genetic Algorithm based on Cosine Similarity for Relevant Document Retrieval.

III. VECTOR SPACE MODEL

The vector space model can best be characterized by its attempt to rank documents by the similarity between the query and each document [10].In the Vector Space Model(VSM), documents and query are represent as a Vector and the angle between the two vectors are computed

using the similarity cosine function. Similarity Cosine function can be defined as:
Where,

$$\mathrm{sim}(d_j, q) = \frac{\mathbf{d_j} \cdot \mathbf{q}}{\|\mathbf{d_j}\| \, \|\mathbf{q}\|} = \frac{\sum_{i=1}^{N} w_{i,j} w_{i,q}}{\sqrt{\sum_{i=1}^{N} w_{i,j}^2} \sqrt{\sum_{i=1}^{N} w_{i,q}^2}}$$

Documents and queries are represented as vectors.

$$d_j = (w_{1,j}, w_{2,j}, \ldots, w_{t,j})$$
$$q = (w_{1,q}, w_{2,q}, \ldots, w_{t,q})$$

Vector Space Model have been introduce term weight scheme known as if-idf weighting. These weights have a term frequency (tf) factor measuring the frequency of occurrence of the terms in the document or query texts and an inverse document frequency (idf) factor measuring the inverse of the number of documents that contain a query or document term [4].

IV. GENETIC ALGORITHM

Genetic Algorithm (GA) is a global optimization algorithm derived from evolution and natural selection. Although genetic algorithm cannot always provide optimal solution, it has its own advantages and is a powerful tool for solving complex problems.

Genetic algorithm is a powerful search mechanism and it is suitable for the information retrieval for the following reasons [18].
The document search space represents a high dimensional space. GAs are one of the powerful searching mechanisms known for its robustness and quick search capabilities. So they are suitable for information retrieval. In comparison with the classical information retrieval models, GA manipulates a population of queries rather than a single query. Each query may retrieve a subset of relevant documents that can be merged. The traditional methods of query expansion manipulate each term independent of other. GA contributes to maintain useful information links representing a set of terms indexing the relevant documents. The traditional methods of relevance feedback are not efficient when no relevant documents are retrieved with the initial query.
Genetic algorithm operations can be used to generate new and better generations. The genetic algorithm operations include:

A. *Reproduction:* the selection of the fittest individuals based on the fitness function.

B. *Crossover:* is the exchange of genes between two individual chromosomes that are reproducing. In one point cross over a chunk of connected. Genes will be swapped between two chromosomes.

C. *Mutation:* is the process of randomly altering the genes in a particular chromosome. There are two types of mutation:

1) *Point mutation:* in which a single gene is changed.

2) *Chromosomal mutation:* where some number of genes is changed completely.

As shown in figure 1 a simple GA works as follows:

1. Start with a randomly generated population.

2. Evaluate the fitness of each individual in the population

3. Select individuals to reproduce based on their fitness

4. Apply crossover

5. Apply mutation

6. Replace the population by the new generation of individuals

7 Go to step 2.

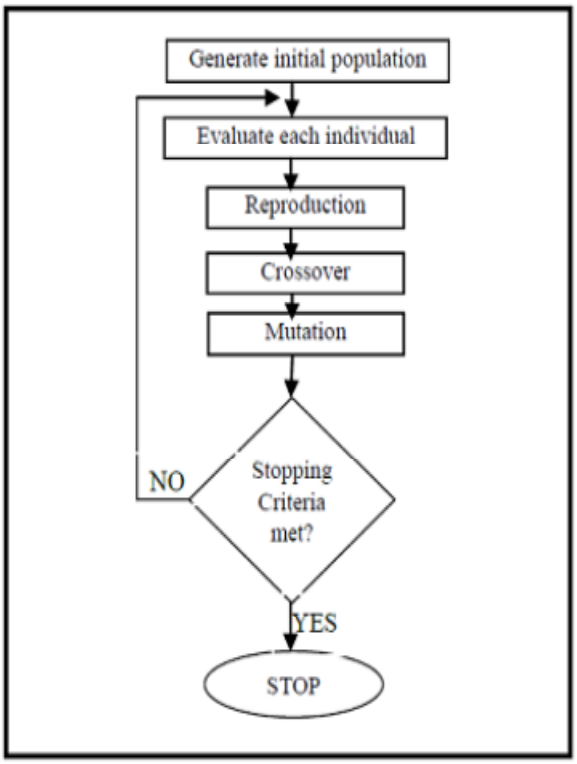

Fig 1 Flowchart of typical Genetic algorithm [19]

V. PROPOSED METHOD

We are using a Genetic Algorithm approach in Text to Matrix Generator (TMG) to improve the performance by optimizing the objective function of vector space model(VSM).The aim of this proposed work is to retrieve the relevant documents by using vector space model from the given set of documents. Here we are using genetic algorithm to optimize the objective function of vector space model (VSM) in Text to Matrix Generator (TMG). The advantage of this proposed work is to save time and retrieve the most relevant document when a query is given.

VI. CONCLUSION

Genetic algorithm is an excellent optimization tool. In this paper, we proposed a genetic algorithm in Text to matrix Generator tool to improve the results. The algorithm use fitness function which is represented by the equation gives more sophisticated result.

By using genetic algorithm in Text to Matrix Generator (TMG) the number of Iterations will be increases which will improve the performance.

REFERENECES

[1] Mohameth-François Sy, Sylvie Ranwez, Jacky Montmain, Armelle Regnault, Michel Crampes, Vincent Ranwez Pezzoli , User centered and ontology based information Retrieval system for life sciences, BMC Bioinformatics, 2012, 1471-2105.

[2] R. Sagayam, S.Srinivasan, S. Roshni, A Survey of Text Mining: Retrieval, Extraction and Indexing Techniques, IJCER, Vol. 2 Issue. 5, Sep 2012, PP: 1443-1444.

[3] Anwar A. Alhenshiri, Web Information Retrieval and Search Engines Techniques, Al- Satil journal PP: 55-92.

[4] Djoerd Hiemstra, Arjen P. de Vries, Relating the new language models of information retrieval to the traditional retrieval models, published as CTIT technical report TR-CTIT-00-09, May 2000.

[5] Vaclav Snasel, Ajith Abraham2, Suhail Owais3, Jan Platos, and Pavel Kromer, Optimizing Information Retrieval Using Evolutionary Algorithms and Fuzzy Inference System, pages: 1-23.

[6] Mohammad Othman Nassar, Feras Fares Al Mashagba, and Eman Fares Al Mashagba, Investigating Genetic algorithms to optimize the user query in the vector space model, Australian Journal of Basic and Applied Sciences, 7(2): 47-53, 2013.

[7] S.Siva Sathya and Philomina Simon, Review on Applicability of Genetic Algorithm to Web Search, International Journal of Computer Theory and Engineering, Vol. 1, No. 4, October2009 1793-8201.

[8] S.Siva Sathya and Philomina Simon, A Document Retrieval System with Combination Terms Using Genetic Algorithm, International Journal of Computer and Electrical Engineering, Vol. 2, No. 1, February, 2010 1793-8163.

[9] Priya I. Borkar and Leena H. Patil, Web Information Retrieval Using Genetic Algorithm-Particle Swarm Optimization, International Journal of Future Computer and Communication, Vol. 2, No. 6, December 2013.

[10] G. Salton and M.J. McGill, editors. Introduction to Modern Information Retrieval. McGraw-Hill 1983

[11] Wafa. Maitah, Mamoun. Al-Rababaa and Ghasan. Kannan, improving the effectiveness of information retrieval system using adaptive genetic algorithm, International Journal of Computer Science & Information Technology (IJCSIT) Vol 5, No 5, October 2013

[12] Mohammad Othman Nassar, Feras Al Mashagba, and Eman Al Mashagba, Improving the User Query for the Boolean Model Using Genetic Algorithms, IJCSI International Journal of Computer Science Issues, Vol. 8, Issue 5, No 1, September 2011.

[13] Marıa J. Martın-Bautista and Marıa-Amparo Vila and Henrik Legind Larsen [13] address A Fuzzy Genetic Algorithm Approach to an Adaptive Information Retrieval Agent, Journal of the American society for information science.50(9):760–771, 1999.

[14] Md. Abu Kausar and Md. Nasar, the detailed study on Information retrieval system using genetic algorithm, Journal of Industrial and Intelligent Information Vol. 1, No. 3, September 2013.

[15] Bangorn Klabbankoh and Ouen Pinngern, applied genetic algorithm in information retrieval.

[16] Praveen Pathak Michael Gordon Weiguo Fan, Effective Information Retrieval using Genetic Algorithms based Matching Functions Adaptation, Proceedings of the 33rd Hawaii International Conference on System Sciences – 2000.

[17] J. Usharani, and Dr K Iyakutti, A Genetic Algorithm based on Cosine Similarity for Relevant Document Retrieval, International Journal of Engineering Research & Technology (IJERT) Vol. 2 Issue 2, February- 2013 ISSN: 2278-0181.

[18] M.Boughanem, C. Chrisment, L. Tamine, Multiple query evaluation based on an enhanced genetic algorithm‖, Information Processing and Management 39,215–231, 2003.

[19] Priya I. Borkar, Leena H. Patil, A model of Hybrid Genetic Algorithm Particle Swarm Optimization (HGAPSO) based query optimization for web Information Retrieval, IJRET Volume: 2 Issue: 1 JAN 2013, ISSN: 2319 – 1163 pages 59-64.

[20] Dimitrios Zeimpekis, Efstratios Gallopoulos, Text to matrix generator (TMG).