Comparison of Crossover Types to Build Improved Queries Using Adaptive Genetic Algorithm

Khaled Almakadmeh, Assistant Professor
Department of Software Engineering
Hashemite University
P.O. Box 330136 Zarqa (13115) Jordan
Email: khaled.almakadmeh@hu.edu.jo

Wafa' Alma'aitah, Lecturer
Department of Basic Sciences
Hashemite University
P.O. Box 330136 Zarqa (13115) Jordan
Email: wafaa_maitah@hu.edu.jo

***Abstract*— this paper presents an information retrieval system that use genetic algorithm to improve information retrieval efficiency and vector space model to measure similarity between query and documents retrieved. Therefore, documents with high similarity to query retrieved first. Using the genetic algorithm, each query represented by a chromosome, these chromosomes are fed into genetic operator process: selection, crossover, and mutation until a query chromosome generated for document retrieval. The proposed approach is experimented using a data set of (242) proceedings abstracts collected from a Saudi Arabian national conference. Experimental results show that information retrieval with adaptive crossover probability set to two-point type crossover and roulette wheel as selection type yields the highest recall.**

Keywords*— *Adaptive genetic algorithm; vector space model; Cosine similarity, Crossover.

I. Introduction

A. Information Retrieval Systems

Information retrieval is a field of study that helps the user to find needed information from a collection of large documents. Retrieving information simply means finding a set of documents that is relevant to the user query [1]. A ranking of these documents is also performed in accordance to their relevance scores to the query. The user with information need issues a query to the retrieval system through the query operational module. Information retrieval systems deal with documentary bases containing textual, pictorial or vocal information and process user queries trying to allow the users to access the relevant information within an acceptable time interval [2].

An IRS consists of three basic components: documentary database, query subsystem, matching mechanism [3]. The documentary database stores documents along with representation of their information content. It is typically associated with the indexer module, which automatically generates a representation of each document by extracting the document contents [3].

The query subsystem allows the user to specify their information needs and presents the relevant documents retrieved by the system. The efficiency of an information retrieval system significantly depends upon query formation [4].

The matching mechanism evaluates the degree to which documents are relevant to user query giving a retrieval status value (RSV) for each document. The relevant document is ranked based on this value [4].

B. Vector Space Model

In vector space model, a document is viewed as a vector in n-dimensional document space (where n is the number of distinguishing terms used to describe contents of the documents in a collection) and each term represents one dimension in the document space. A query is also treated in the same way and constructed from the terms and weights provided in the user request. Document retrieval is based on the measurement of the similarity between the query and the documents [5]. This means that documents with a higher similarity to the query are judged more relevant to it and should be retrieved by the information retrieval system in a higher position in the list of retrieved documents. Using this model, the retrieved documents can be orderly presented to the user with respect to their relevance to the query [6].

C. Genetic Algorithms

A genetic algorithm is a probabilistic search algorithm, which is used for optimization of difficult problem. It is based on Darwinian principle of natural selection. It exploits and explores the document search space [7]. The basic operators used by genetic algorithm are selection, crossover and mutation. By using these operators, complex problems can be easily solved. Genetic Algorithm basic components are [8] [9]:

- Chromosome Representation: chromosomes are the initial input given to GA. All the documents and query are first converted into chromosome. This is given as input to the genetic algorithm [10].
- 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 [10].
- Selection operator: selection is the process in which chromosomes are selected for next step or generation in genetic algorithm based on fitness value of chromosomes. Poor chromosome or lowest fitness chromosome selected few or not at all [11].