GA on IR: Study the Effectiveness of the Developed Fitness Function on IR

Ammar Al-Dallal, School of Information Systems Computing and Mathematics, Brunel University, West London, UK

Rasha S. Abdul-Wahab, College of Information Technology, Ahlia University, Manama, Bahrain

ABSTRACT

Increasing the growth rates of websites’ number has led to the challenge of assisting Web customers in finding appropriate details from the Internet using an intelligent search engine. Information retrieval (IR) is an essential and useful strategy for Web users; thus, different strategies and techniques are designed for such purpose. Currently, the focus on the usefulness of Artificial Intelligence (AI) has been improved with IR. One AI area is Evolutionary Computation (EC), which is based on designs of natural selection. A traditional and important strategy in EC is Genetic Algorithm (GA); this paper adopts the GA technique to enhance the retrieval of HTML documents. This improvement is obtained by creating a modern evaluation function and applying a hybrid crossover operator. The proposed evaluation function is based on term proximity, keyword probability within the document, and HTML tag weight query. Experimental results are compared with two well known evaluation function functions applied in IR domain which are Okapi-BM25 and Bayesian interface network model. The results demonstrate a good level of enhancement to the recall and precision. In addition, the documents retrieved by the proposed system were more accurate and relevant to the queries than that retrieved by other models.

Keywords: Artificial Intelligence (AI), Evolutionary Computation (EC), Genetic Algorithm (GA), HTML Document, Information Retrieval (IR)

1. INTRODUCTION

Rapid growth to the number of Web pages needs continuous challenges for helping Web users to find relevant information from the Internet. Information Retrieval (IR) is an essential and useful technique for Web users so studying of such system have increased since the coming of the World Wide Web.

In recent years, emphasis in the applicability of Artificial Intelligence (AI) has been increased with IR. One of the AI areas is Evolutionary Computation (EC) which is based on models of natural selection. A classical and important technique in EC is Genetic Algorithm (GA). The GA is biologically inspired and has many mechanisms inspired by natural evolution. Because of its parallel mechanism with high-dimensional space, GA has been used to solve many of scientific and engineering problems.

DOI: 10.4018/jalr.2012040101
This in turn led to encourage researchers for using this algorithm in IR. Besides, GA plays an important approach to provide suitable information for the user’s needs.

IR and GA integrated to avoid web users suffering from specific problems when trying to retrieve useful information such that:

- Many of the retrieved documents are not related to the user query.
- Some of relevant documents have not been retrieved yet (Picarougne et al., 2002).

However, retrieving relevant information is not a simple process. However, the complexity of this process is further increased by the fact that more and more of this information appears in natural language and not in structured formats (Liu, 2006).

The rest of this paper is organized as follows: in section 2, we describe problem statement. Section 3 describes the main objectives of this research. Section 4 discusses related works. Document representation is introduced in section 5. Section 6 introduces the proposed approach. Section 7 represents the experimental results of the proposed method. Section 8 gives the conclusions of this study.

2. PROBLEM STATEMENT AND OBJECTIVES

Every internet user wishes to have satisfied results when using any web search engine. Satisfaction in sense that all the retrieved results are relevant, and all relevant documents are retrieved; in another words, the web user is satisfied when the search engine retrieves all and only the relevant documents. In spite of several enhancements are achieved in such searching techniques, still web users suffer from two major problems when trying to retrieve useful information.

The objective of this paper is integrate IR and GA to avoid web users suffering from such problems when trying to retrieve useful information such that: Many of the retrieved documents are not related to the user query which is called low precision, and many of relevant documents have not been retrieved yet which is called low recall. In order to enhance precision and recall, this paper proposes a novel information retrieval GA based searching mechanism. In addition this paper explains how the new developed crossover operator can affect the efficiency of the proposed GA system. Another objective presented in this paper is to develop a new evaluation function. This new function will be used to evaluate documents and retrieve the related documents at high rank.

3. RELATED WORK

The long history of GAs with IR is presented by integrating GA in IR with the aim of solving IR problems. One of these approaches is the one developed by Kim and Zhang (2003; 2000). They proposed a GA-based retrieval method which is used to learn the importance of HTML tags. This method shows an improvement of average precision when using tagged information over non-tagged information.

Picarougne et al (2002) reported their experience in designing different fitness function for web search using Genetic Programming (GP). They tested this set of fitness functions with other existing order-based fitness functions on the task of ranking function discovery. Their results show the design of such fitness functions lead to an increase in performance.

An integrated framework for discovering and optimizing ranking functions is proposed by Weiguo et al (2009). These novel ranking functions are discovered by using GP technique. The structural and statistical information available in HTML documents collected intelligently in their work. However, combining the scores of various well-known ranking functions is achieved by using GA. Their framework is tested on the well-known TREC collection of web documents. They viewed a significant increase in retrieval performance and their results outperformed the other well-known ranking functions.
Related Content

Analysis of a Step-Based Watershed Algorithm Using CUDA
[www.irma-international.org/article/analysis-step-based-watershed-algorithm/52612/](www.irma-international.org/article/analysis-step-based-watershed-algorithm/52612/)

Guiding Self-Organization in Systems of Cooperative Mobile Agents
[www.irma-international.org/chapter/guiding-self-organization-systems-cooperative/4983/](www.irma-international.org/chapter/guiding-self-organization-systems-cooperative/4983/)

Mapping with Monocular Vision in Two Dimensions
[www.irma-international.org/article/mapping-monocular-vision-two-dimensions/52616/](www.irma-international.org/article/mapping-monocular-vision-two-dimensions/52616/)

A Parallel Hardware Architecture based on Node-Depth Encoding to Solve Network Design Problems
[www.irma-international.org/article/a-parallel-hardware-architecture-based-on-node-depth-encoding-to-solve-network-design-problems/104694/](www.irma-international.org/article/a-parallel-hardware-architecture-based-on-node-depth-encoding-to-solve-network-design-problems/104694/)

A Family of Superstable n-D Mappings
[www.irma-international.org/article/family-superstable-mappings/38934/](www.irma-international.org/article/family-superstable-mappings/38934/)