Volume 6, Issue 6, June 2016 ISSN: 2277 128X
International Journal of Advanced Research in Computer Science and Software Engineering
Research Paper
Available online at: www.ijarcsse.com

Information Retrieval using Dice Similarity Coefficient

Manoj Chahal*
Master of Technology, Department of Computer Science and Engineering,
GJUS & T, Hisar, Haryana, India

Abstract: ***There is large Volume of data available on digital world. To retrieve relevant and useful information in this digital world is one of the challenging tasks. It is impossible to user to retrieve efficient data manually. To solve this problem search engine is used. Search engine use Information Retrieval System and Genetic Algorithm to retrieve relevant information. In this paper Dice Similarity Function is used to retrieve relevant information .Dice similarity is used in genetic Algorithm to get efficient data from digital world.***

Keywords: Genetic Algorithm, Dice Similarity Coefficient, Search Engine, optimization, Vector Space Model.

I. INTRODUCTION

With the explosive growth of the number of Web pages on Internet the requirement of search engine to assisting users in finding the best and newest information has been increased. Search engine technology has had to scale dramatically to keep up with the growth of the web. Gerard Salton was the father of modern search technology. His team developed the smart information retrieval system [10]. In 1994 one of the first web search engines the World Wide Web Worm had an index of 110,000 web pages and web accessible documents. As of November 1997 the top search engines claim to index from 2 million to 100 million web documents [8].Google search engine began in January 1996 by Larry page and Sergey Brin as a research project. It was first incorporated as a privately held company on 1998 and its initial public offering on August 2004. In January of 2013 Google announced it had earned $50 billion in annual revenue for the year of 2012. The first time Google had reached this feat topping their 2011 total of $38 billion. [11]

The various part of search engine are:-

- Crawler
- Pagerank
- Repository
- Indexer

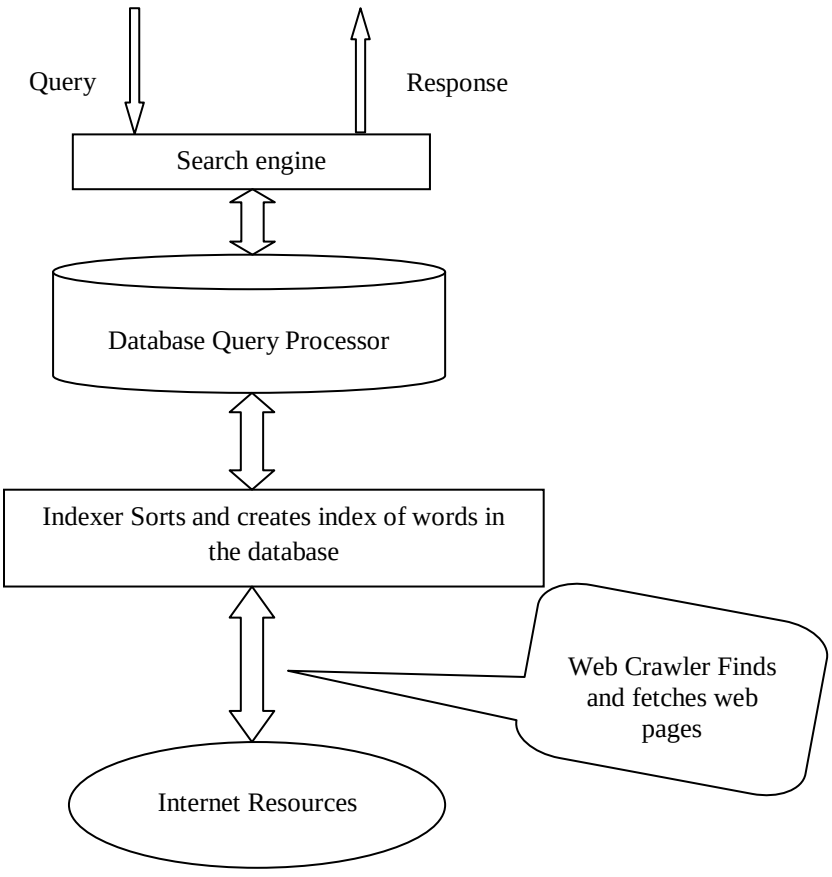


Figure 1 Search Engine [9]

1. Information Retrieval System

Information retrieval is a system used to retrieve information or document which is relevant to the user query. The goal of information retrieval system is to help a user to locate the most similar document that have the potential to satisfy the user needs.

The various part of information retrieval is:-

- User
- Query Subsystem
- Matching Mechanism
- Document Database

2. Similarity Measures:-

It measures similarity between the two documents. It gives the probability or degree of similarity between the two or more documents.

There are three information retrieval models in the information retrieval area are

- Boolean model
- Vector space model
- Probabilistic model

Dice Similarity measure come under Vector space model. In this model a document is viewed as a vector in n-dimensional document space and each term represents one dimension in the document space. Document retrieval is based on the measurement of similarity between the query and document.

Dice formulation as shown below:

$$\mathbf{D}(\boldsymbol{A},\boldsymbol{B}) = \frac{\mathbf{2}|\boldsymbol{A} \cap \boldsymbol{B}|}{|\boldsymbol{A}| + |\boldsymbol{B}|}$$

II. GENETIC ALGORITHM

It is a probabilistic search algorithm used to get the optimal solution of a given problem. It based on the principle of Darwinian principle of natural selection and using operations that are patterned after naturally occurring genetic operations such as crossover and mutation.

Genetic algorithm operations can be used to generate new and better generations. Generate Initial Population for Query Chromosome. Initially query and document is converted into chromosome for giving input to Genetic algorithm. These initial chromosomes make initial population for Genetic algorithm.

III. PREVIOUS WORKS ON INFORMATION RETRIEVAL

There are several studies that used genetic algorithm in information retrieval system to optimize the user query.

Siti Nurkhadijah Aishah Ibrahim [1] presented a retrieval model of hybrid GA-Particle Swarm Optimization based query optimization for Web information retrieval. The keywords are used to produce new keywords that are related to the user search. Seung-Seok Choi, Charles C. Tappert [2] described various similarity and distance measures. Each of them is differently defined by its own synthetic properties. Some include negative matches and some do not. Some use simple count difference and some utilize complicated correlation. J.Allaan, Jay Aslam et al [3] described Information retrieval with the structure, analysis, organization, storage, searching and retrieval of information.

Philomina Simon and S. Siva Sathya [4] described a general frame work of information retrieval system. The applicability of genetic algorithm was discussed in different areas of information retrieval such as genetic mining, query optimization, document clustering, and query optimization etc. Chahal et al[5] describe the various similarity function also described genetic algorithm and their application in the field of information retrieval. Zhengyu Zhu, Xinghuan Chen et al. [6] described relevance feedback technique to retrieve relevant information. They applied Genetic Algorithm to optimize user query and retrieve web information. M. Zolghadri-jahromi, and M.R. Valizadeh [7] described a query sensitive similarity measure mechanism to measure the similarity of two documents. In the first step they identified the sources of information that may be used for this purpose. In the second step they proposed a query sensitive similarity measure based on these information sources. Finally it was proposed that a query sensitive similarity measure parametric that simultaneously makes use of the product and weighted sum to fuse the information from the identified sources. S. Brin and L. Page [8] described architecture of search engine and also described the working of search engine.

IV. EXPERIMENT

Step for conducting experiment:-

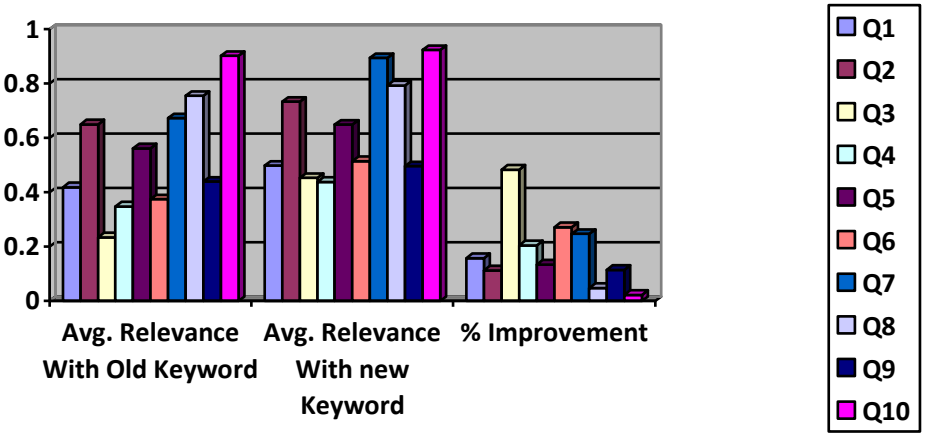
- First document are taken from database.
- Convert these documents into initial chromosome.
- This is used as input to Genetic Algorithm.
- Applying Dice similarity function
- Then apply selection, crossover and Mutation

V. RESULT

Adding new Keyword and Calculating Percentage of Improvement:

Table 1.1: Percentage Improvement in Average Relevance after Adding New Keyword.

Query	Avg. Relevance With Old Keyword	Avg. Relevance With new Keyword	% Improvement
Q1	0.4190	0.4976	15.79%
Q2	0.6502	0.7329	11.28%
Q3	0.2341	0.4532	48.34%
Q4	0.3482	0.4382	20.53%
Q5	0.5619	0.6490	13.42%
Q6	0.3745	0.5143	27.18%
Q7	0.6722	0.8934	24.75%
Q8	0.7549	0.7932	4.82%
Q9	0.4392	0.4962	11.48%
Q10	0.9020	0.9234	2.31%



VI. CONCLUSION AND FUTURE WORKS

It is observed that average relevance of documents increases by applying Dice Similarity Function in GA. It means Dice Similarity Function explore and exploit our search space. Average relevance of document can be increased by applying other methods. In this paper Dice Similarity Function is applied but this work can also be done by applying other similarity measure and compare the result with each other. In this paper binary vector is applied but this work can also be done with weighted vector.

REFERENCES

[1] Siti Nurkhadijah Aishah Ibrahim, Ali Selamat, Mohd Hafiz Selamat , "*Query Optimization in Relevance Feedback using Hybrid GA-PSO for EffectiveWeb Information Retrieval* " , *Third Asia International Conference on Modeling & Simulation,2009.*

[2] Seung-Seok Choi, Sung-Hyuk Cha, Charles C. Tappert, "*A Survey of Binary Similarity and Distance Measures*" , *Department of computer science, Pace University*

[3] J. Allaan, Jay Aslam et al., "Challenges in information retrieval and language modeling", *Report of a Workshop held at the Center for Intelligent Information Retrieval, University of Massachusetts Amherst*, Sept. 2002.

[4] Philomina Simon and S. Siva Sathya,"*Genetic Algorithm for Information Retrieva*l", *Ramanujan School of Mathematics and Computer Sciences, Pondicherry, India.*

[5] Manoj Chahal and Jaswinder Singh ," Effective Information Retrieval Using Similarity Function: Horng and Yeh Coefficient", *International Journal of Advanced Research in computer science and software Engineering* , vol 3 Issue 8 ,pp- 401-406 ,Aug 2013.

[6] Zhengyu Zhu, Xinghuan Chen, Qihong Xie, Qingsheng Zhu, "A GA based query optimization for web information retrieval", *International Conference on Intelligent Computing*, pp. 2069-2078, Aug. 2005.

[7] M. Zolghadri-jahromi, and M.R. Valizadeh, "A proposed query-sensitive similarity measure for information retrieval", *Iranian Journal of Science & Technology,* Shiraz University, vol. 30, no. B2, pp.171-180, 2006.

[8] S. Brin and L. Page, "The Anatomy of a Large-Scale Hypertextual Web Search Engine," *Proc. Seventh Int'l Conf. World Wide Web* (WWW '98), pp. 107-117, 1998.

[9] http://www.google.co.in/searchenginediagram.

[10] http://www.searchenginehistory.com.

[11] http://en.wikipedia.org/wiki/Google.

