A Genetic Algorithm based on Cosine Similarity for Relevant Document Retrieval

J. Usharani, Assistant Professor, Dept. of Computer Science, M. K. U. College, Madurai, India
Dr K Iyakutti, Professor, Dept of Physics and Nanotechnology, SRM University, Chennai, India

Abstract

As the web is growing at a very rapid rate, the amount of information and the pages that are similar to each other are also increasing. One central problem of information retrieval is to determine the relevance of document based on user query. Relevancy of pages can be calculated with the use of similarity measures. In this paper we propose a genetic algorithm based method for finding similarity of web document based on cosine similarity. Using our proposed method we first calculate the average relevancy of document retrieved from Google search engine based on query. We then expand the query with new keywords mostly found in the retrieved document and show that average relevancy is increases.

1. Introduction

The growth of the World Wide Web has lead 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.

2. Genetic Algorithm

GA belongs to the class of evolutionary computational algorithm that mimics the natural process of evolution to discover solution to problems. GAs are good at effectively solving large search and optimization problem Genetic algorithm is a powerful search mechanism and it is suitable for the information retrieval for the following reasons The document search space represents a high dimensional space. GAs are one of the powerful searching mechanism known for their robustness and quick search capabilities. So they are suitable for information retrieval. GA exploits the idea of the survival of the fittest and an interbreeding population to create a novel and innovative search strategy In GA, the search space is composed of candidate solutions to the problem, each represented by a string termed as a chromosome. Each chromosome has an objective function value, called fitness. A set of chromosomes together with their

associated fitness is called the population. This population, at a given iteration of the genetic algorithm, is called a generation.

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 with probability Pc
5. Apply mutation with probability Pm
6. Replace the population by the new generation of individuals
7 Go to step 2.

3. Vector space Model

The vector space model (VSM) is an IR model that represents the documents and queries as vectors in a multidimensional space, whose dimensions are the terms used to build an index to represent the documents Document retrieval is based on the measurement of the similarity between the query and the documents. This means that documents with a higher similarity to the query are judged to be more relevant to it and should be retrieved by the IRS in a higher position in the list of retrieved documents. In this method, the retrieved documents can be orderly presented to the user with respect to their relevance to the query. Suppose there are t index terms in collection of documents. Then document $D_i = (di_1, di_2, di_3, .., di_k)$

$Q = (wq_1, wq_2, wq_3, .., wq_k)$

where d_{ij}(j=1 to t) are term weights in document D_i and wq_j(j=1 to t) are term weights in the query Q.

4. Similarity Measures

A similarity measures is a function which computes the degree of similarity between a pair of text objects .Similarity Measures rely heavily on terms occurring in both query and the document. If the query and document do not have any term is common then similarity score is very low. Different similarity measures have been suggested to match the query document. Some of popular measures are cosine, jaccard, dice etc. In this paper we apply the cosine similarity.

4.1 Cosine Similarity

Cosine similarity is one of the most popular similarity measure applied to text document, in numerous information retrieval applications. When documents are represented as term vectors, the similarity of two documents corresponds to the correlation between the vectors. Basically, cosine similarity is a measure of similarity that can be used to compare documents with respect to a given vector of query words. This is quantified as the cosine of angle between vectors.

Given two document x,y and their cosine similarity is

$$\cos(x, y) = \frac{x.y}{|x||y|}$$

where x, y is m dimensional vector over the term set T= $\{t_1,t_2..t_m\}$
Each dimension represent a term with a weight in the document which non negative. As a result cosine similarity is nonnegative and bounded between [0, 1].

5. Experiment work and Results

In our experiment we have initially selected few queries and retrieved first 10 documents from the Google search engine. We generate chromosomes and extract keywords with highest frequency from each of these pages. The length of chromosome depends upon number of keywords extracted from the 10 documents.

Average relevancy of each document for query was calculated using cosine similarity as fitness function and applying selection, crossover and mutation operation. We have selected roulette function for selection of parents for crossover. We use binary encoding for chromosomes wherein a 1 is put if the term appears in the query and a 0 is put otherwise.
The values of various parameters for GA are as follows:

- Chromosome length - 15
- No of Generations - 150
- Cross over probability – 0.7
- Mutation Probability – 0.02

The Fitness Function used is the cosine similarity function:

$$\frac{\sum_{i=1}^{n} A_i \times B_i}{\sqrt{\sum_{i=1}^{n} (A_i)^2} \times \sqrt{\sum_{i=1}^{n} (B_i)^2}}$$

where A is the document vector and B is the query vector.

Table 1: Average relevancy using cosine similarity

Query	Average relevancy	New keywords added	Average relevancy after adding new keywords using proposed method	% increase in relevancy
Share Loan Bank	0.8131	Money	0.9645	18.62%
Genetic Algorithm Neural Network	0.5624	Soft computing	0.6179	9.87%
Mysql Dbms database	0.7285	Data	0.8124	11.52%
Ilayaraja Music mp3	0.6425	Download	0.6915	7.63%
Object-oriented Multithread Platform	0.7135	Java	0.8790	23.20%
Cluster Association Classification	0.8145	Data Mining	0.9512	16.78%

We first calculate the average relevancy of document retrieved from Google search engine based on the query. We then expand the query with new keywords mostly found in the retrieved document and as can be observed from the table, the average relevancy is increases.

6. Conclusion and Future Work

It is observed that the usage of GA increases the relevancy of retrieved documents. As a part of future work, the effect of adjusting the values of various parameters of GA such as mutation probability, cross over probability can be studied.

7. References

[1] D. E. Goldberg, Genetic Algorithms in Search, Optimization, and Machine Learning: Addison-Wesley, 1989

[2] Imran, Hazra, Sharan, Aditi, "Genetic Algorithm based Model for Effective Document Retrieval"

[3] Dallal, A.L., Wahab, Abdul Q.S., "Genetic Algorithm based to improve Document Retrieval"
[4] Thada, Vikas, Joshi, Sandeep, "A Genetic Algorithm Approach For Improving The average Relevancy Of Retrieved Documents Using Jaccard Similarity Coefficient", International Journal of Research in IT & Management, Volume 1, Issue 4, August, 2011.