GENETIC ALGORITHM BASED APPROACH FOR THE SELECTION OF PROJECTS IN PUBLIC R&D INSTITUTIONS

SANJAY S , PRADEEP S , MANIKANTA V , KUMARA S.S , HARSHA P

Department of Human Resource Development
CSIR-Central Food Technological Research Institute, Mysore-20, India

Manilal P [*]

Department of Planning, Monitoring and Coordination
CSIR-Central Food Technological Research Institute, Mysore-20, India
chirakkal@yahoo.com

Abstract

Identification and selection of new project concepts are one of the crucial steps in the project selection process in research establishments. It is essential to ensure that objectives of proposed projects are evaluated for duplicity and redundancy against databases. In this context, implementation of a genetic algorithm (GA) based approach is described.

Keywords: Genetic Algorithm; Information processing; Project Management.

1. Introduction

Public R&D laboratories are in constant expedition to formulate innovative solutions to achieve a sustainable growth, wide visibility while positioning the organisation strategically in their domain areas. Risks and uncertainties associated with R&D project selection is an unenviable task for the decision making process. Viable and potential problems need to be picked up to ensure effective delivery with a sizeable impact on the national and global scenario. Projects are aimed to meet a number of objectives focused on various topics under exploration. By adopting a pragmatic screening and evaluation mechanism, projects with identical and redundant objectives are to be avoided to save various resources. Role of the decision makers in R & D laboratory include assessing the potential goals while new projects are considered. Selecting appropriate tools and mechanism assumes significance in this context. It is thus desirable to evolve a scientific mechanism in the selection and screening of projects.

Conventionally information retrieval methods depend on boolean queries for search process. The similarities between a query and documents collection are measured by various retrieval methods that are based on more frequent terms found in both the document as well as the query. Measures of retrieval effectiveness include precision and recall values in which retrieved documents are compared against relevant items[1]. Information retrieval process involves representation, storage, searching and retrieving the documents which are relevant. GA represents one of the artificial intelligence algorithms to improve performance in information retrieval systems. First pioneered by John Holland in 1975 [4] , GA has been widely studied, experimented and applied in many fields of engineering and management sciences. GA has been found to be useful to various domain areas such as information retrieval, engineering design, image processing, robotics, routing problems, encryption, computer aided molecular design and gene expression profiling.

2. Theory of Genetic Algorithm

GA is a search and optimization technique which is used to find exact or approximate solutions for search problems. It is the simulation of evolution process and they are probabilistic optimization methods. In GA, the search procedure has a direct analogy to the set of competing individual chromosomes represented by a string of binary codes. A fitness value of individual is computed using a fitness function. Population is a collection of chromosomes together with their fitness value. Genetic algorithm performs a series of computations on the population to evolve a new generation. GA is an adaptive heuristic search algorithm based on the evolutionary ideas of natural selection and genetics. GA provides an alternative method for solving problems and can outperform many traditional methods [4]. A screen view of the algorithm is given (Annexure I).

3. Functionalities

The functionalities and steps involved in a GA process are:

3.1. Chromosome Representation: Translate the real problem into biological terms in which a chromosome is represented as a string of zeroes and ones. If a particular keyword is present in the document, then the position in the string is marked as '1' or else '0'. These chromosomes are termed as initial population and length of chromosome depends on number of document keywords retrieved from a document vector.

3.2. Fitness Evaluation**:** Fitness function is a measure of performance, which evaluates the suitability of the solution. The fitness of the chromosomes is evaluated (1) using Jaccard's coefficient [3],

$$\frac{x \cap y}{x + y - x \cap y} \quad . \; . \; . \; . \; . \; . \; . \; . \; . \; . \; . \; . \; . \; . \; . \; . \; (1)$$

Where, x = number of keywords in document X
y = number of key-words in document Y
$x \cap y$ = number of key-words common in both X and Y

Results of these fitness functions fall in the range 0-1. The value 1.0 means document match exactly to the query and the value which is near to 1.0 means retrieved documents are more relevant to the query.

3.3. Selection**:** This operation selects chromosomes from the population for reproduction. Fitter the chromosome, more likely a chromosome will be selected for the reproduction and the Roulette wheel selection procedure is used here.

3.4. Crossover**:** Crossover is the genetic operator that mixes two chromosomes together to form a new offspring with certain probability. Chromosomes that are not subjected to crossover are carried to offspring without any modifications. Though there are different types of crossovers used for the information retrieval, two-point crossover is considered appropriate [3]. In a Two-point crossover two points are selected on the parent strings, and bits between these two points are swapped rendering two child organisms.

3.5. Mutation**:** Mutation involves modification of each gene with a certain probability. While changing some bit values of chromosomes, a new breed is created. The mutation can make genetic algorithms fast approach to the global optimum and quickly get out of premature convergence. Depending on the nature of the problem and encoding type, implementation of crossover and mutation may vary greatly. A pictorial representation of the algorithm is shown in Fig.1.Genetic algorithms offer many solutions and search multiple points simultaneously [5].

4. Methodology

Genetic Algorithm based retrieval procedure was tried for a dataset of 70 R & D projects having around 300 objectives in the area of food science and technology. The queries were subjected for a match in the dictionary. The extracted keywords from these projects were arranged alphabetically and the resultant template was used to represent the chromosome. Fitness value of each of the chromosomes (project) was calculated using Jaccard's fitness function (1) and the average fitness values were recorded. Using roulette wheel method, cumulative sum of the fitness was computed. The crossover was carried out for two of the randomly selected documents using two-point crossover method. The newer chromosome from this step replaces the older ones and evolutionary process were continued. New chromosomes were then subjected to mutation process and the process was repeated until one of the terminating conditions such as (a) previous generation was fitter than the present generation (b) average fitness value of a generation equal to 1 (c) number of generations <= 150 were fulfilled. Chromosomes with the highest fitness value of the final generation is compared bitwise against the initial population and the documents with predefined threshold probability are considered as the matching document or project.

The GA parameters were set as the number of generations = 100; type = two point crossover; crossover probability = 0.8; mutation type = bit flip and mutation probability = 0.01 based on various trials [3]. Efficacy of the retrieval was compared using the precision and recall parameters.

5. Result and Conclusion

Customized software with a user-friendly interface and improved functionality was developed. Modules developed have provisions to add new project, objectives and to view intermediate and final results. Five sets of keywords (s1, s2, s3, s4, s5) were chosen and the results are showed (Table 1). Also the graph (Fig. 2) shows the quick convergence of the retrieval process. The precision and recall values were calculated and compared against that of conventional method. Though improved precision was visible in the case of a GA based approach compared to keyword based conventional approach (Table 2), the recall has not been satisfactory for the selected dataset. It could be mainly due to quality of the initial query set as well. The software could be used while sanctioning new R&D proposals by funding agencies and to identify unique and non-redundant project topics which could guarantee maximum gains to public. The project selection process could be made more scientific and robust by incorporating project abstracts to the database to cover wider scope while using a GA –assisted selection mechanism for new R&D projects.

6. Acknowledgements

The authors thank Director, CFTRI, Mysore for his encouragement in undertaking this study. The co-authors, Sanjay S and Pradeep S acknowledges Head, Human Resource Development, CFTRI, Mysore for enabling to carry out the project for partial fulfillment of their graduate degree programme.

References

[1] Ahmed A A Radwan; Abdel Mgeid A Ali;Bahgat A Abdel Latef & Osman A Sadek. (2006).*Using genetic algorithm to improve information retrieval systems*, Proceedings of the *World Academy of Science Engineering and Technology*, 17, pp.6-12.

[2] Andrew Troelsel,*C# and .NET Platform, (Apress, USA)*, 2nd edition, 2003.

[3] Bangorn Klabbankoh & Quen Pinngern.(2005). *Applied genetic algorithms in information retrieval, Proceedings of International Journal of Production Research (King Mongnut's Institute of Technology, Ladkrabang, Bangkok)*, 43, pp.4083-4101.

[4] Goldberg D E, *Genetic Algorithms: In search, optimization and machine learning*, New York: Addison-Wesley Publishing Co. Inc., 1989, pp. 59-75.

[5] Simon mardle & Sean pascol.(1999) .*An overview of genetic algorithms for the solution of optimization problems, Computers in Higher Education Economics Review*, 13(1),pp.16-20.

[6] Suhail S J Owai;, Pavel Kromer & Vaclav Snasel.(2005). Evolutionary learning of boolean queries by genetic programming, Proceedings of *ADBIS Research Communications*, 152.

[7] Suhail S J Owais, Pavel Kromer & Vaclav Snasel, Query optimization by genetic algorithms, In *Dateso* 2005, Edited by K Richta, V Snasel & J Pokorny, Vol 2005, 2005, 125-137.

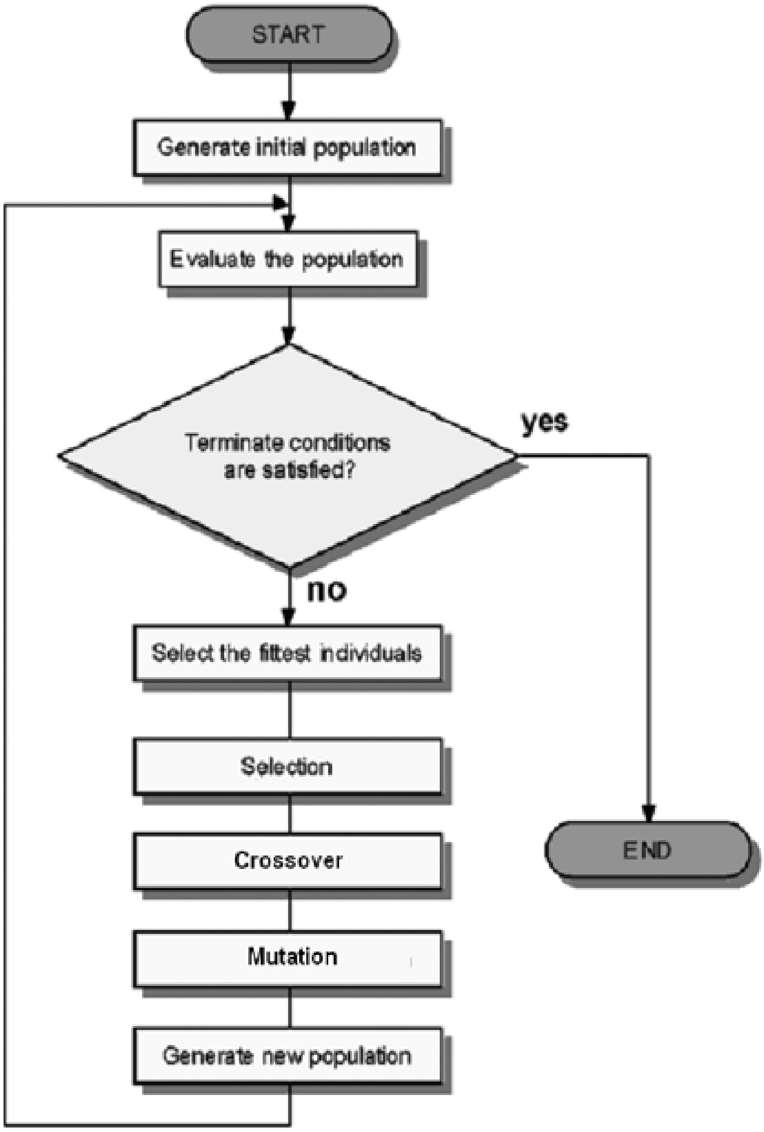

START
Generate initial population
Evaluate the population
Terminate conditions are satisfied?
yes
no
Select the fittest individuals
Selection
Crossover
Mutation
Generate new population
END

Fig.1. Steps in a simple genetic algorithm process

Number of Generations	**Average Fitness**				
	s1	**s2**	**s3**	**s4**	**s5**
0	0.0472708	0.041408583	0.240535	0.109205	0.183776
1	0.13237353	0.181489833	1	0.4138186	0.769231
2	0.2448646	0.30328025		0.4499248	0.85
3	0.28813073	0.374810583		0.536153	
4	0.31513487	0.419421917		0.719804	
5	0.38541687	0.450491833			
6	0.41409387	0.469521333			
7	0.50891753	0.471422083			
8		0.563688083			
9		0.573468667			
10		0.58372075			
11		0.591537			

Table 1: Average fitness Vs generations

Input Sets	**Precision (GA)**	**Precision (keyword based)**	**Recall (GA)**
s1	0.66	0.10	0.2
s2	0.166	0.33	0.2
s3	1	0.75	0.33
s4	0.5	0.29	1
s5	1	0.50	1

Table 2: Precision Vs recall

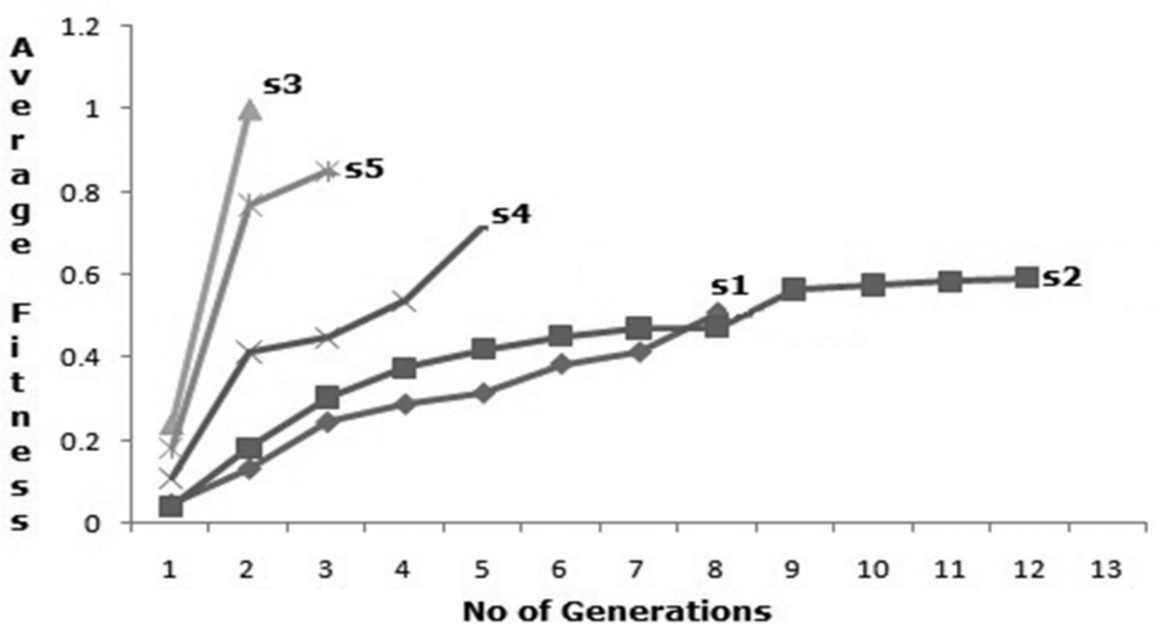


Fig.2: Average fitness against number of generations

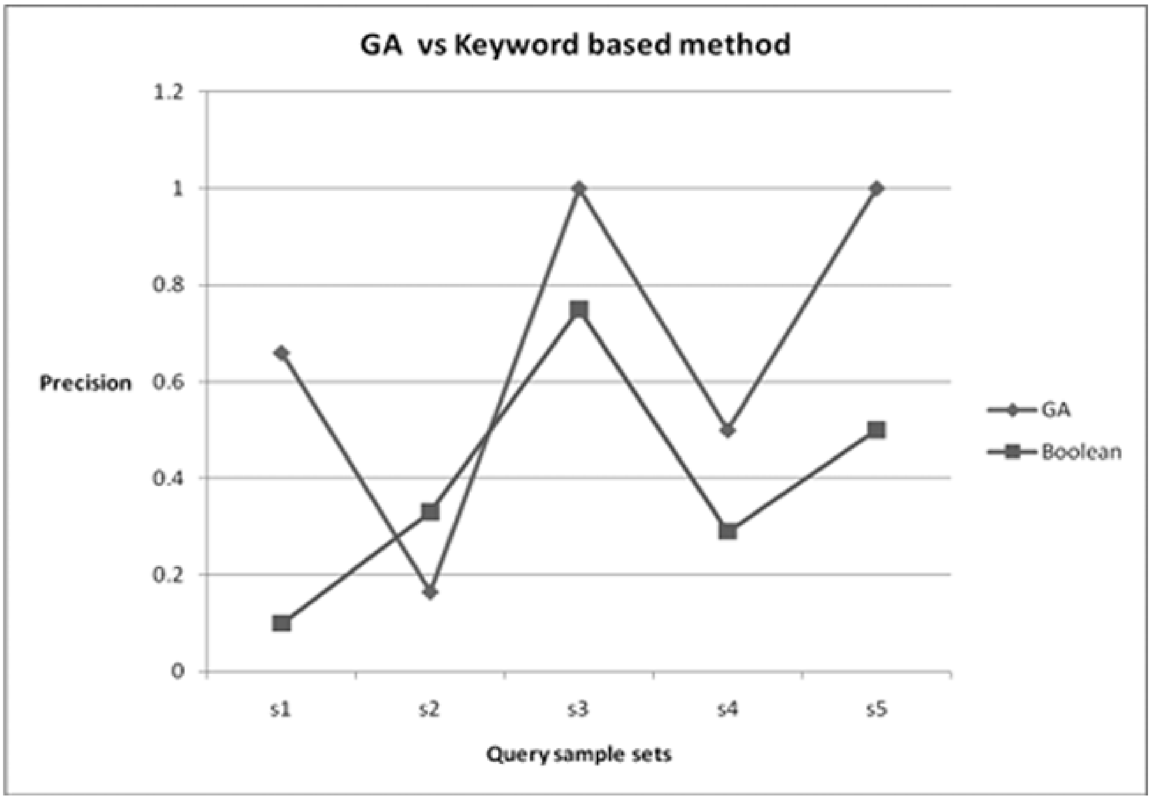
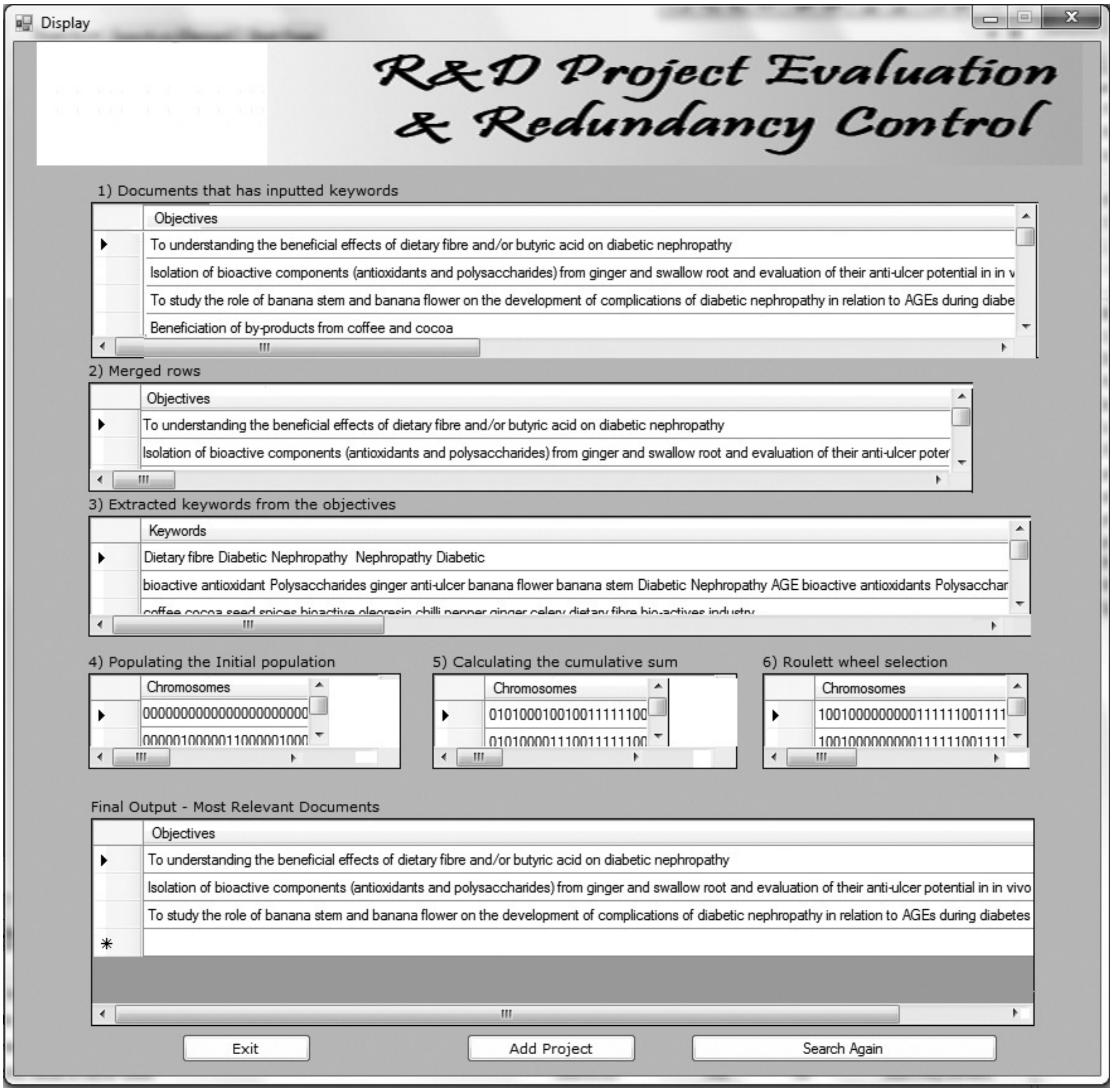


Fig.3: Comparison of precision



1) Documents that has inputted keywords

Objectives
To understanding the beneficial effects of dietary fibre and/or butyric acid on diabetic nephropathy
Isolation of bioactive components (antioxidants and polysaccharides) from ginger and swallow root and evaluation of their anti-ulcer potential in in v
To study the role of banana stem and banana flower on the development of complications of diabetic nephropathy in relation to AGEs during diabe
Beneficiation of by-products from coffee and cocoa

2) Merged rows

Objectives
To understanding the beneficial effects of dietary fibre and/or butyric acid on diabetic nephropathy
Isolation of bioactive components (antioxidants and polysaccharides) from ginger and swallow root and evaluation of their anti-ulcer poter

3) Extracted keywords from the objectives

Keywords
Dietary fibre Diabetic Nephropathy Nephropathy Diabetic
bioactive antioxidant Polysaccharides ginger anti-ulcer banana flower banana stem Diabetic Nephropathy AGE bioactive antioxidants Polysacchar
coffee cocoa seed spices bioactive oleoresin chilli pepper ginger celery dietary fibre bio-actives industry

4) Populating the Initial population

Chromosomes
000000000000000000000
0000010000011000001000

5) Calculating the cumulative sum

Chromosomes
0101000100100111111100
0101000011100111111100

6) Roulett wheel selection

Chromosomes
1001000000000111111001111
1001000000000111111001111

Final Output - Most Relevant Documents

Objectives
To understanding the beneficial effects of dietary fibre and/or butyric acid on diabetic nephropathy
Isolation of bioactive components (antioxidants and polysaccharides) from ginger and swallow root and evaluation of their anti-ulcer potential in in vivo
To study the role of banana stem and banana flower on the development of complications of diabetic nephropathy in relation to AGEs during diabetes

Annexure I**:** Intermediate **r**esults from a GA based retrieval session

This page is intentionally left blank

This page is intentionally left blank

This page is intentionally left blank