International Journal of Research (IJR)
e-ISSN: 2348-6848, p- ISSN: 2348-795X Volume 3, Issue 01, January 2016
Available at http://internationaljournalofresearch.org

A Survey of Evolutionary Approaches for Information Retrieval

Ankit Naik & CA Dhote

Student, CSE, PRMIT&R, Badnera, India
ankitnaik@live.com

Professor, IT, PRMIT&R, Badnera, India
vikasdhote@rediffmail.com

Abstract

The survey paper attempts to explain the use of various Evolutionary approaches which can be used to optimize and enhance the Information Retrival System (IRS) performance. The paper describes the types of Evolutionary Approaches and their application in Information Retrieval. Comparison of various approaches is done to give a clarity where each approach can be applied to get better results.

Key Words: Information Retrieval; volutionary Approaches; Genetic Algorithm; Genetic Programming.

Introduction

Information Retrieval (IR) is a field of study that helps the user to extract useful information from a large collection of documents. Information retrieval (IR) tries to make a suitable use of these data bases, allowing the users to access to the information which is really relevant in an appropriate time interval.[1]

An Information Retrieval System (IRS) is a software tool for data representation, storage and information search. IRS manages large collection of documents and provides easy, efficient and accurate information search. General architecture of an information retrieval system is shown in Fig. 1. [2]

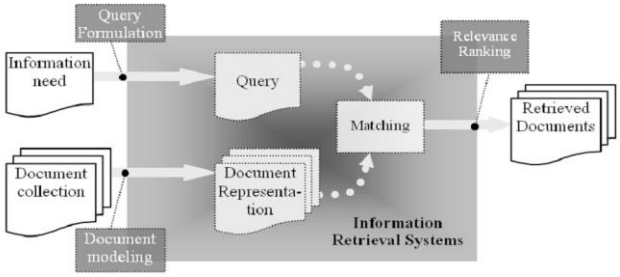

Fig1 Information Retrieval System (IRS)[2]

Need for IR using Evolutionary Approaches

There are three main classes of retrieval models. Exact Match Models which form the basis of most commercial retrieval systems, Vector Space Models which view documents and queries as vectors in a high- dimension vector space and use distance as a measure of similarity, and probabilistic models which view retrieval as a problem of estimating the probability that a document representation matches or satisfies a query.The vector space and probabilistic models have been shown experimentally to offer significant improvements in retrieval performance over exact-match models.[4][5]

The performance of information retrieval can be enhanced by using evolutionary approaches which can improve the quality of query and obtain more developed queries that fit the searcher's needs.[3]

Evolutionary Algorithms

Evolutionary Algorithms(EA) uses Evolutionary Computation(EC) based on computational

International Journal of Research (IJR)
e-ISSN: 2348-6848, p- ISSN: 2348-795X Volume 3, Issue 01, January 2016
Available at http://internationaljournalofresearch.org

models of evolutionary processes as key elements in the design and implementation of computer-based problem solving systems.[1]

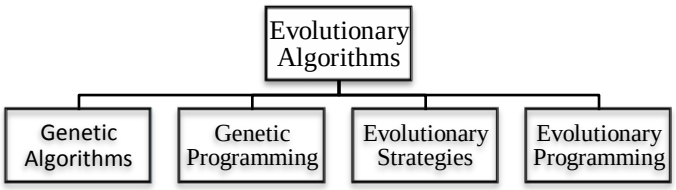

Fig. 2. Types of Evolutionary Approaches

Genetic Algorithms (GA)

A genetic algorithm is a search procedure inspired by principles from natural selection and genetics. It is often used as an optimization method to solve problems where little is known about the objective function. The operation of the genetic algorithm is quite simple. It starts with a population of random individuals, each corresponding to a particular candidate solution to the problem to be solved. Then, the best individuals survive, mate, and create offspring, originating a new population of individuals. This process is repeated a number of times, and typically leads to better and better individuals. [6][8]

Genetic Programming (GP)

Genetic programming is an extension of the genetic algorithms, which uses the principle of natural evolution to describe the design evolution. Its biggest difference from the genetic algorithm is that genetic programming is an executable program, rather than a string. A frequent objective of genetic programming is to achieve human-type machine intelligence with little direct human involvement.[9]

Evolutionary strategies (ES) and Evolutionary programming (EP)

The basic difference between EP (or ESs) and GAs [1] is the variation operator used for producing offspring. Both EP and ES use only mutation operator to produce offspring, while GAs use both crossover and mutation operators. Since mutation is the main operator in EP , a number of innovative mutation operators have been proposed such as Cauchy mutation, a combination of Cauchy and Gaussian mutation, and Levy mutation. The aim of these mutations is to introduce large variations for producing offspring so that a population can globally explore wider regions of a search space. This means that the improvement of EP has been sought by increasing its exploration capability. However, both exploration and exploitation are necessary, depending on whether an evolutionary process becomes trapped in a local optima or finds more promising regions in the search space.[10]

GENETIC ALGORITHMS(GA)

Genetic Algorithms(GA) works on the principle of natural selection. GA is an iterative process that operates on a population, i.e., a set of candidate solutions. Initially, the population is randomly generated. Every individual in the population is assigned, by means of a fitness function, a fitness value that reflects its quality with respect to solving the particular problem. The reproductive operators like crossover and mutation are then applied to the individuals in this population yielding a new population. The whole process is repeated until a certain termination criterion is achieved. [11][12]

International Journal of Research (IJR)
e-ISSN: 2348-6848, p- ISSN: 2348-795X Volume 3, Issue 01, January 2016
Available at http://internationaljournalofresearch.org

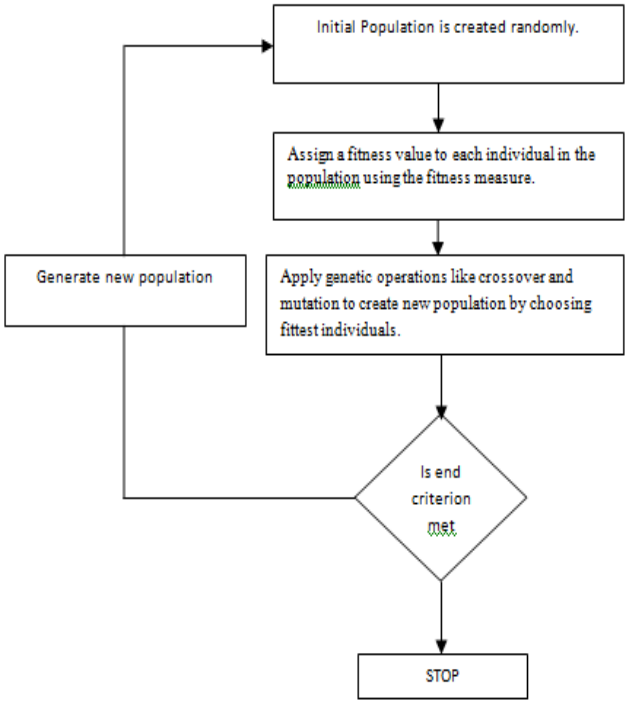

Fig.3. GA process Flow chart.

GENETIC ALGORITHM (GA) FOR INFORMATION RETRIEVAL

The GA's are used to solve some of the Information Retrieval problems like query optimization and formulation , document indexing etc. Some of the techniques are

A. Genetic Mining

The process of mining of data from web documents with the help of Genetic Algorithms is termed as Genetic Mining. A Genetic Algorithm which ranks the document based on the internal structure of the documents was proposed by Sun Kim[15]. Ardil [16] proposed another genetic algorithm in concept weighting and topic identification, based on the concept of standard deviation.[13][14]

B. Internet Search

The Genetic Algorithm that searches for Web pages based on keywords and provides relevant pages in an Internet search to make search process easier is called a Web Crawler or Web Spider. Chen et al. implemented Internet personal spiders based on best first search and GA techniques. [14]

C. Query Optimization

To make an information retrieval system effective by improving the query using mathematical techniques and Genetic Algorithms operators like crossover and mutation.

D. Document Clustering

Document Clustering helps to find relationships among different documents having some related pieces of information relevant to user's needs. [14]

Table 1 : Application of GA in Different Areas of IR

GA Application Area	Purpose of GA	Population	Fitness Function	Genetic Operators

International Journal of Research (IJR)
e-ISSN: 2348-6848, p- ISSN: 2348-795X Volume 3, Issue 01, January 2016
Available at http://internationaljournalofresearch.org

Genetic Mining[19]	Learn the internal structure of HTML documents	Tag weights are encoded as chromosomes	Fitness function measures the performance of retrieval results using tag weights	Roulette Selection Truncation selection One point cross over Uniform cross over Single point Mutation
Internet Search[20]	Intelligent Searching	The search space of the problem is represented as a collection of individuals which are referred as chromosomes	Jaccards fitness function	Stochastic selection Heuristic based crossover and mutation operators
Query Optimization	Retrieve more relevant Chromosome documents with respect to with respect to user query	Chromosome encoding done with boolean query where it is represented by tree structure	Precision Recall	Single Point CrossOver Single Point Mutation
Document Clustering	GA grouping the terms without maintaining initial order	Two different coding schemes used Separator Method Division Assignment Method	Measure of relative entropy Pratt's measure	Roulette Selection Order based, Position based, One and two point cross over Inversion, random sublist and position Mutation operators

GENETIC PROGRAMMING(GP)

Genetic Programming(GP) is an extension of Genetic Algorithm(GA) . Genetic programming is a branch of genetic algorithms. The main difference between genetic programming and genetic algorithms is the representation of the solution. Genetic programming creates computer programs in the lisp or scheme computer languages as the solution. Genetic algorithms create a string of numbers that represent the solution. [21][22]

GP uses four steps to solve problems:
1) Generate an initial population of random compositions of the functions and terminals of the problem (computer programs).

2) Execute each program in the population and assign it a fitness value according to how well it solves the problem.

3) Create a new population of computer programs.

i) Copy the best existing programs
ii) Create new computer programs by mutation.
iii) Create new computer programs by crossover(sexual reproduction).

4) The best computer program that appeared in any generation, the best-so-far solution, is designated as the result of genetic programming.[21][22]

International Journal of Research (IJR)
e-ISSN: 2348-6848, p- ISSN: 2348-795X Volume 3, Issue 01, January 2016
Available at http://internationaljournalofresearch.org

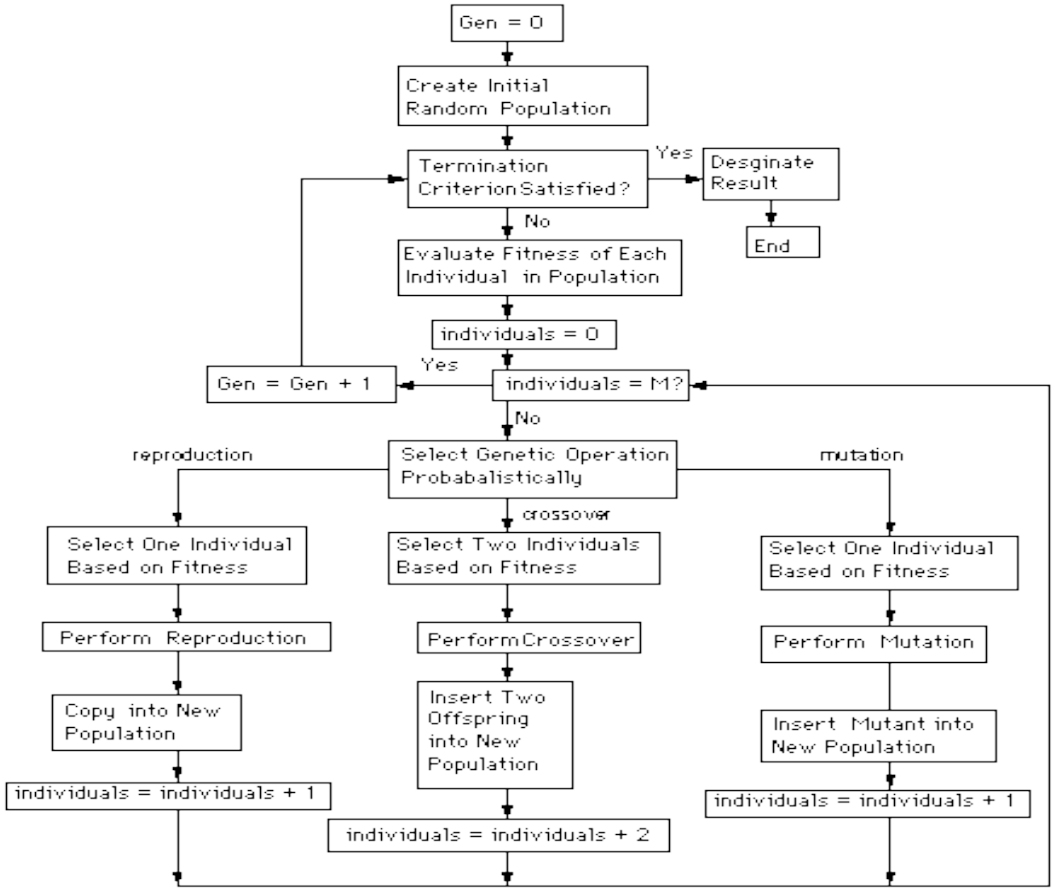

Figure 2: Genetic Programming Flowchart.

GENETIC PROGRAMMING (GP) FOR INFORMATION RETRIEVAL

Ranking for Information Retrieval

One of the main problem of information retrieval(IR) is to determine which documents are relevant and which are not to the user information needs. The ranking is done based on the keywords(i.e. index terms). [23]

RankGP is Genetic Programming based algorithm which starts with a set of individuals as the initial population, then

each individual's fitness is calculated. The most fit individual is added to the output set and new population is created by reproducing the most fit individual into it. Genetic operators like crossover and mutation are applied to generate individuals. Evaluate the performance and output best one.[23]

Term Weighting in Information Retrieval

Vector space model represents each document in the collection as a vector of terms with weights associated to each term. The weight of each term is based on the frequency of the term in the documents and collection. The query (user need) is also modeled as a vector and a matching function is used to compare each document vector to the query vector. Once the documents are compared, they are sorted into a ranked list and returned to the user.[24][25]

International Journal of Research (IJR)
e-ISSN: 2348-6848, p- ISSN: 2348-795X Volume 3, Issue 01, January 2016
Available at http://internationaljournalofresearch.org

The GP approach adopted in this work evolves the weighting scheme over a number of generations. An initial population is created randomly by combining a set of primitive measures using a set of operators. The average precision, used as the fitness function, is calculated for each scheme by comparing the ranked list returned by the system for each weighting scheme against the human determined relevant documents for each query. Average precision is calculated over all points of recall and is frequently used as a performance measure in IR systems. The matching function used in all experiments is the inner-product matching function. [24][25]

EVOLUTIONARY STRATEGIES (ES)

Evolution strategies are based on the principal of strong causality, which states that similar causes have similar effects. It uses only Mutation as genetic operator in contrast to GA which uses both Crossover and Mutation.

The process of evolution strategy can be summarized by a relatively simple algorithm:

1. Generate some random individuals.

2. Select the p best individuals based on some selection algorithm (fitness function)

3. Use these p individuals to generate c children (using mutation or recombination)

4. Go to step 2, until the ending condition is satisfied (i.e. little difference between generations, or maximum number of iterations completed).

There are two types of ES :

1. **$(\mu + \lambda)$ – ES**

$(\mu + \lambda)$ - ES specifies that μ parents produce λ descendants, where $\lambda > \mu$. The descendants compete with their parents in the selection of the best μ individuals to the creation of the next generation. It is an elitist strategy.[26][27]

2. **(μ, λ) – ES**

(μ, λ) - ES is very similar to $(\mu + \lambda)$-ES with the exception that only descendants survive and go through next generation. This strategy is more greedy than $(\mu + \lambda)$ and it allows for more diversity in the population, thus avoiding the algorithm to get trapped in local optima.[26][27]

EVOLUTIONARY PROGRAMMING(EP)

Evolutionary Programming(EP) differs substantially from GA and GP, in that EP emphasizes the development of behavioral models and not genetic models. EP is derived from the simulation of adaptive behavior in evolution.
The main components of an EP are :

1. Initialization

2. Evaluation
   - Fitness function measures the "behavioral error" of an individual with respect to the environment of that individual.
   - provides an absolute fitness measure of how well the problem is solved
   - Survival in EP is usually based on a relative fitness measure.
   - A score is computed to quantify how well an individual compares with a randomly selected group of competing individuals
   - Individuals that survive to the next generation are selected based on this relative fitness
   - The search process in EP is therefore driven by a relative fitness measure, and not an absolute fitness measure

3. Mutation as the only source of variation

International Journal of Research (IJR)
e-ISSN: 2348-6848, p- ISSN: 2348-795X Volume 3, Issue 01, January 2016
Available at http://internationaljournalofresearch.org

4. Selection
   - Main purpose to select new population
   - A competitive process where parents and offspring compete to survive.

COMPARISON OF EVOLUTIONARY APPROACHES FOR INFORMATION RETRIEVAL

	Genetic Algorithm (GA)	Genetic Programming (GP)	Evolutionary Strategies (ES)	Evolutionary Programming (EP)
Problem representation Scheme and Genetic Operator	Fixed-size bit-strings using crossover as its main operator	Trees of flexible size	Vectors of real values for representation and mutation as the main operator	Manipulates graphs using mutation as the single genetic operator
Selection Scheme	Proportional Selection	Proportional Selection	Ranking based selection	Tournament selection
Type of Problem Solved	Optimization Problem	Computational Problem	Empirical Experiments	Optimization problem possessing many local optimal solutions

Conclusion This paper describes and compares various Evolutionary Approaches currently being used in Information Retrieval. The approaches are mainly based on the "Theory of Evolution". These approaches help in optimizing and enhancing the performance of current approaches and getting better results. These approaches can be used in vast areas like Web Search Engines, Ranking of Documents, Document Clustering etc.

References:

[1] O. Cord_on , E. Herrera-Viedma , C. L_opez-Pujalte , M. Luque , Zarco A review on the application of evolutionary computation to information retrieval International Journal of Approximate Reasoning 34 (2003) 241–264

[2] Vaclav Snasel, Ajith Abraham, Suhail Owais, Jan Platos, and Pavel Kromer Optimizing Information Retrieval Using Evolutionary Algorithms and Fuzzy Inference System

[3] Wafa. Maitah, Mamoun. Al-Rababaa and Ghasan. Kannan Improving The Effectiveness Of Information Retrieval System Using Adaptive Genetic Algorithm International Journal of Computer Science & Information Technology (IJCSIT) Vol 5, No 5, October 2013

[4] Howard R. Turtle And W. Bruce Croft A Comparison of Text Retrieval Models. The Computer Journal, Vol. 35, No. 3, 1992

[5] Information Retrieval Models Djoerd Hiemstra . Published in: Goker, A., and Davies, J. Information Retrieval: Searching in the 21st Century. John Wiley and Sons, Ltd., ISBN-13: 978-0470027622, November 2009.

[6] A Detailed Study on Information Retrieval using Genetic Algorithm Md. Abu Kausar, Md. Nasar, Sanjeev Kumar Singh Journal of Industrial and Intelligent Information Vol. 1, No. 3, September 2013

[7] The History of Information Retrieval Research Sanderson, M.; Croft, W.B. Proceedings of the IEEE

Year: 2012, Volume: 100, Issue: Special Centennial Issue

[8] Hybrid evolutionary algorithms for data classification in intrusion detection systems Hedar, A.-R.; Omer, M.A.; Al-Sadek, A.F.; Sewisy, A.A. Software Engineering, Artificial Intelligence, Networking and Parallel/Distributed Computing (SNPD), 2015 16th IEEE/ACIS International Conference on Year: 2015

[9] Dynamic simulations of nonlinear multi-domain systems based on genetic programming and bond graphs Di, Wenhui; Sun, Bo; Xu, Lixin Tsinghua Science and Technology Year: 2009, Volume: 14, Issue: 5

[10] Recurring Two-Stage Evolutionary Programming: A Novel Approach for Numeric Optimization Alam, M.S.; Islam, M.M.; Yao, X.; Murase, K. Systems, Man, and Cybernetics, Part B: Cybernetics, IEEE Transactions on Year: 2011, Volume: 41, Issue: 5 Pages: 1352 -1365

[11] Genetic Algorithm And Programming Based Classification: A Survey Dharminder Kumar, Sunita Beniwal Journal Of Theoretical And Applied Information Technology 10th August 2013. Vol. 54 No.1

[12] Survey Of Genetic Algorithms And Genetic Programming John R. Koza ISBN# 0-7803-2636-9

[13] Using Genetic Algorithm to Improve Information Retrieval Systems Ahmed A. A. Radwan, Bahgat A. Abdel Latef, Abdel Mgeid A. Ali, and Osman A. Sadek International Science Index, Computer and Information Engineering Vol:2, No:5, 2008 waset.org/Publication/8821

[14] Genetic Algorithm for Information Retrieval Philomina Simon S. Siva Sathya 978-1-4244-4711-4/09/ ©2009 IEEE

[15] Sun Kim and Byoung, Genetic Mining of HTML structures for effective information retrieval , Applied Intelligence , 18, 243256,2003

[16] S. M. Khalessizadeh, R. Zaefarian, S.H. Nasseri, and E. Ardil , Genetic Mining, Using Genetic Algorithm for Topic based on Concept Distribution, PROCEEDINGS OF World academy of science, engineering and technology ,143 -147, 2006

[17] O.Cordon, .E.Herrera , C. Lopez- Pujalte, M.Luque, C.Zarco ,A review on the application of evolutionary computation to information retrieval, International Journal of Approximate reasoning34, 241- 264, 2003.

[18]Michael Gordon, Applying probabilistic and genetic algorithms for document retrieval, Computer Practies, 1208 -1218 , 1988

[19] Michael Gordon, Applying probabilistic and genetic algorithms for document retrieval, Computer Practies, 1208 -1218 , 1988

[20] H. Chen, C. Yi-Ming, M. Ramsey, C. Yang, "An intelligent personal spider (agent) for dynamic Internet/Intranet searching", Decision Support Systems 23 (1998) 41-58.

[21]Koza, John R. 1992. Genetic Programming: On the Programming of Computers by Means of Natural Selection. Cambridge, MA: The MIT Press.

[22]Cramer, Nichael Lynn: "A Representation for the Adaptive Generation of Simple Sequential Programs", Proceedings, International Conference on Genetic Algorithms and their Applications, July 1985 [CMU], pp183-187.

[23] Learning to rank for information retrieval using Genetic Progamming Jen-Yuan Yeh, Jung -Yi Len, Hao-Ren Ke, Wei- Pang Yang copyright 2007 ACM 1-58113-000-0/00/0004

[24] Determining general term weighting schemes for the Vector Space Model of Information Retrieval using Genetic Programming Ronan Cummins and Colm O'Riordan.

[25] Term-Weighting in Information Retrieval using Genetic Programming: A three stage process Ronan Cummins and Colm O'Riordan

[26] Enhancing Information Retrieval By Using Evolution Strategies Abdelmgeid Amin Aly International Conference Knowledge-Dialogue-Solutions 2007

[27] Evolution strategies A comprehensive introduction HANS-GEORG BEYER and HANS-PAUL SCHWEFEL Natural Computing 1: 3–52, 2002. © 2002 Kluwer Academic Publishers. Printed in the Netherlands.