УДК 004.6-048.34

STATISTICAL SEARCH ENGINE OPTIMIZATION

A. HASSAN
Instructor-teache Lebanese University
– Lebanon, MS computer science-4th
year PhD

Lebanese University, Lebanon

Abstract. This report provides a view point of the development of optimization methods for the Retrieval of relevant information from a documentary database. As Genetic Algorithms (GA) are robust and efficient search and optimization techniques, they can be used to search the huge document search space. In this search, a general frame work of information retrieval system and a development of optimization methods are to be discussed. With the rapid growth of the amount of data available in electronic libraries, through Internet and enterprise network mediums, advanced methods of search and information retrieval are in demand. Information retrieval systems, designed for storing, maintaining and searching large-scale sets of unstructured documents. One step in optimizing the information retrieval experience is the deployment of Genetic Algorithms, a widely used subclass of Evolutionary Algorithms that have proved to be a successful optimization tool in many areas. The Evolutionary computation, Evolutionary Search Process, Genetic Operators, Genetic Programming, Evolutionary Techniques and Fuzzy Logic Principles in IRS, Fuzzy principles in Information are all discussed. Establishment of statistical regularities in the study of search information in documentary database and a great focus on the Search engine optimization (SEO) is the core of this study , meanwhile the genetic indexed along with the indexing methods creation in the process of information search on the base of data obtained as a result of statistical analysis performed queries to documentary database will be fully illustrated.

Key words: optimization, Internet, documentary database, Genetic Algorithm

Introduction. Information retrieval dealt with the representation, storage, organization, and access to information items. The representation and association of the information items will provide the user with effortless access to the information in which he will be interested. Unfortunately, characterization of the client information need is not a simple problem. Find all the pages containing information on college football teams which are maintained by a university in one country and participate in any tournament. To be relevant, the result must include information on the national ranking of the team in the last few years and the email or mobile number of the team participants and coach. Clearly, this complete description of the player information need not be used directly to request information using the recent Web search engines. The user must first translate this information need into a query which will be processed by the IR system. In its most general form, this translation gives a set of keywords which summarize the description of the user information. Given the user query, the key goal of IR system is to retrieved information which may be relevant to the user.

1 Genetic Algorithm (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. General structure of genetic algorithms It starts with the operation and the basic theory of the genetic algorithm, and then moves on to the key aspects of current genetic algorithm theory.

1.1 Generic Algorithm Theory. This theory is centered around the notion of a building block. The study will be talking about deception, population sizing studies, the role of parameters and operators, building block mixing, and linkage learning. These studies are motivated by the desire of building better GAs, algorithms that can solve difficult problems quickly, accurately, and reliably. It is therefore a theory that is guided by practical matters.

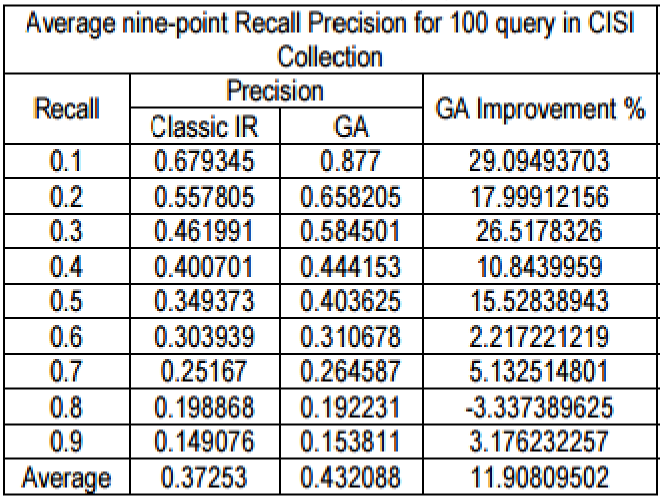
1.2 Genetic Algorithm Operation. A Genetic Algorithm Operation This section describes the operation of a simple genetic algorithm . The exposition uses a step-oriented style and is written from an application perspective. The steps of applying a GA are: To Choose an encoding , Choose a fitness function, Choose operators, Choose parameters, Choose initialization method and stopping criteria (Multiagent system ,2009).

1.3 IR system factors

The Internet has brought far more information than anybody can absorb. Similarly, organizations store a large amount of information in manuals, procedures, documentation, expert knowledge, e-mail archives, news sources, and technical reports. Such a large amount of information serves as a huge information repository for organizations. However, it also makes finding relevant information from it extremely difficult. How to help users find their required information is the central task of any information retrieval (IR) system or search engine. However, precision and recall, the two most commonly used performance measures, of commonly used search engines are usually very low .

Table 1

Shows the experimental result of applying GA on CISI collection



Average nine-point Recall Precision for 100 query in CISI Collection

Recall	Precision		GA Improvement %
	Classic IR	GA	
0.1	0.679345	0.877	29.09493703
0.2	0.557805	0.658205	17.99912156
0.3	0.461991	0.584501	26.5178326
0.4	0.400701	0.444153	10.8439959
0.5	0.349373	0.403625	15.52838943
0.6	0.303939	0.310678	2.217221219
0.7	0.25167	0.264587	5.132514801
0.8	0.198868	0.192231	-3.337389625
0.9	0.149076	0.153811	3.176232257
Average	0.37253	0.432088	11.90809502

Retrieval performance of an IR system can be affected by many factors: the ambiguity of query terms, unfamiliarity with system features, as well as factors relating to document representation. Many approaches have been proposed to address these issues. For example, query expansion techniques based on a user's relevance feedback have been used to discover a user's real information need. Similarly, document descriptions have been modified. Another very important factor is the ranking/matching function. It is this ranking function that to focus most of the discussion on. A ranking function is used to order documents in terms of their predicted relevance to a particular query. It is very difficult to design such a ranking function that can be successful for every query, user, or document collection (which we will call contexts). In this search, it is argue in favor of a method that systematically adapts a ranking function and tailors it to different users' needs (i.e. in different contexts). In particular, an inductive learning technique, for the adaptation purpose and compare our results against two well-known retrieval systems. (Virginia Polytech,2011)and (Detelin Luchev paper, Applying genetic algorithm in query improvement,2007)

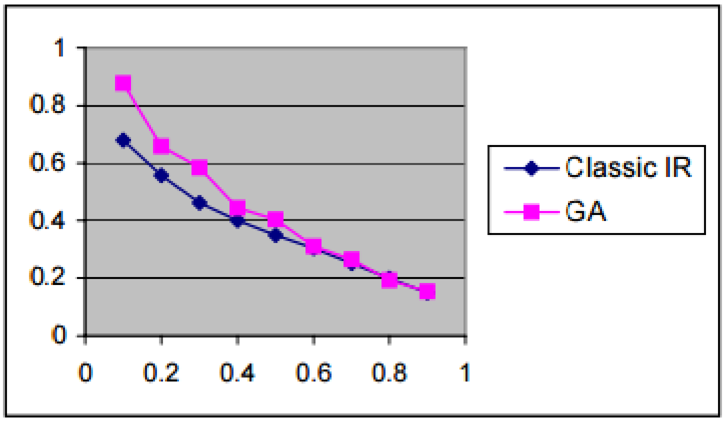


Figure 1. represent the relationship between average of the recall precision for 100 queries of CISI

Fuzzy theory, as a framework describing formally the concepts of vagueness, imprecision, uncertainty and inconsistency provide interesting extensions to the area of information retrieval. Imprecision and vagueness are present in natural language and take part in real-world human communication. User friendly and flexible advanced IRS should be able to offer user interface for non experienced users allowing natural deployment of these concepts in user system interaction for more effective information retrieval. IR models exploiting fuzzy techniques can overcome some of the limitations pointed out in first part of this article. They support different grades of document-query relevance, cut inaccuracies and oversimplifications happening during document indexing and introduce the concepts of vagueness and imprecision in query language.**(** Nicholas J. Belkin and W. Bruce Croft,1998)**.**

2 Genetic Programming. Genetic Programming by is known as unique situation or ex-tension to GA. Encoded individuals (chromosomes) have hierarchical framework, infinite dimension plus they are frequently modeled as tree houses. Therefore could be modeled numerical treatments, reasonable words and sometimes even entire computer applications (i.e. Lisp programs). Genetic coding is just synthetic development of research requests and a local device for acting.(John Koza)

3. Evolutionary Techniques and Fuzzy Logic Principles in IRS. Fuzzy concepts affect most phases of IR process. They are deployed during document indexing, query formulation and search request evaluation. Information retrieval is seen as fuzzy multi-criteria decision making in the presence of vagueness. In general, document is interpreted as a fuzzy set of document descriptors and queries as a composite of soft search constraints to be applied on documents. Document-query evaluation process is based on fuzzy ranking of the documents in documentary collection according to the level of their conformity to the soft search criteria specified via user queries. The document-query matching has to deal with the uncertainty arising from the nature of fuzzy decision making and from the fact that user information needs can be recognized, interpreted and understood only partially. Moreover, the document content is described only in a rough, imperfect way. In the fuzzy enabled IR frameworks, soft search criteria could be specified using linguistic variables. User search queries can contain elements declaring level of partial importance of the search statement elements. Linguistic variables such as ”probably” or ”it is possible that”, can be used to declare the partial preference about the truth of the stated information. The interpretation of linguistic variables is then among the key phases of query evaluation process. Term relevance is considered as a gradual (vague) concept. The decision process performed by the query evaluation mechanism computes the degree of satisfaction of the query by the representation of each document. This degree, called Retrieval Status Value (RSV), is considered as an estimate of the relevance of the document with respect to the query. RSV

= 1 corresponds to maximum relevance and RSV = 0 denotes no relevance. The values within the range (0, 1) correspond to particular level of document relevance between the two extremes 0 and 1. Possibility theory together with the concept of linguistic variable defined within fuzzy set theory provides a unifying formal framework to formalize the processing of imperfect information. Inaccurate information is inevitably present in information retrieval systems and textual databases applications. The automatically created document representation based on a selection of index terms is invariably incomplete and far worse than document representations created manually by human experts who utilize their subjective theme knowledge when performing the indexing task. Automated text indexing deals with imprecision since the terms are not all fully significant to characterize the document content and their statistical distribution does not reflect their relevance to the information included in the document necessarily. Their significance depends also on the context in which they appear and on the unique personality of the inquirer. During query formulation, users might have only a vague idea of the information they are looking for therefore face difficulties when formulating their information needs by the means of query language of particular IR system. A flexible IRS should be designed to provide detailed and rich representation of documents, sensibly interpret and evaluate soft queries and hence offer efficient information retrieval service in the conditions of vagueness and imprecision. In the following, Extended Boolean IR model as the representative of fuzzy IR models will be discussed in details. Some other recent fuzzy IR models will be briefly presented.

4. Statistical regularity concept of search information in documentary databases. Statistical regularity concept is to be study For the establishment of statistical regularities in the study processes of search information in documentary databases. it is a concept in statistics and in the theory of probability which shows that random events exhibit regularity when repeated many times or that enough adequately similar random events. It is an umbrella term that covers the law of large figures, theorems and all central limit theorems.

Saying a series of judgments may create not equivalent, although comparable, outcomes for every sequence: the typical change, the typical along with other distributional qualities is likely to be round the same for every series of tests.

Observations of this occurrence provided the original inspiration for the notion of what is now known as frequency probability.

5. Search engine optimization: SEO , Search Engine Optimization. is the process of getting traffic from the free, organic, editorial or natural search results on search engines. All major search engines such as Google, Bing and Yahoo have primary search results, where web pages and other content such as videos or local listings are shown and ranked based on what the search engine considers most relevant to users. Payment isn't involved, as it is with paid search ads. (Search engine websit,2016).

5.1. Methods. Methods such as , Getting indexed, Preventing crawling, Increasing prominence, marketing strategy, International markets is to be considered here.

5.2. indexing Methods. For Indexing methods creation in the process of information searching on the basis of data obtained as a result of statistical analysis performed queries to documentary databases we need to consider the following points. Search engine indexing, Indexing, Index design factors, Merge factors, Storage techniques, Index size, Lookup speed, Maintenance, Fault tolerance and Index data structures.

Internet search engine architectures differ in ways of catalog storage to meet up the different style elements as well as in the manner indexing is conducted.

Suffix tree Figuratively organized just like a pine, facilitates linear time research. By keeping the suffixes of phrases, constructed. The suffix tree is just a kind of trie. Hashing, that will be essential for internet search engine indexing is supported by attempts. Employed for clustering and trying to find designs in DNA sequences. A significant disadvantage is the fact that keeping a within the pine might need room beyond that necessary to shop the term itself. Another manifestation is just a suffix selection, that will be thought to need less digital storage and facilitates data-compression like the

BWT algorithm. These index are like , Inverted index , citation index, Ngram index, document term matrix,

There may be within the style of SE's a significant problem the administration of sequential processing procedures. There are lots of possibilities for race conditions and coherent faults. For instance, there is a brand new doc put into the corpus and also the catalog should be updated, however the catalog simultaneously must proceed answering search requests. This can be a crash between two duties that are competing. Contemplate that writers are suppliers of information, along with a webcrawler may be the customer of the information, getting the written text and keeping it in a cache (or corpus). The index may be the info made by the corpus' customer, and also the ugly index may be the customer of data made by the catalog that is forward. This really is generally known as a maker-customer design. The indexer may be the maker of data that is searchable and customers would be the people who have to research. The process is increased whenever using processing and dispersed storage. Within an energy to size with bigger levels of listed info, the structure of the internet search engine might include distributed processing, where the research engine includes many devices working together. This causes it to be harder to keep a completely synchronized, dispersed, similar structure and escalates the possibilities.(MICHAEL S. LEW, RAMESH JAIN University of California at Irvine, USA)

5.3 Index merging. The list is stuffed using repair or a combine. There is a repair similar to a combine but first removes the items of the list. The structure might be made to help small indexing, in which a combine then parses each record into phrases and recognizes the doc or files to become included or updated. For precision, recently listed files, usually surviving in digital storage, using the catalog cache living on a single computer hard disk drives are conflated by a combine.

5.4 Meta-tag indexing. Particular files frequently include embedded meta-information for example keywords writer, explanation, and vocabulary. For HTML websites, the meta-tag includes keywords that are also contained in the catalog. Earlier Web search engine technology might just index the keywords within the meta-tags for that catalog that is forward; the entire doc wouldn't be parsed. In those days total-text indexing wasn't too proven, or was computing devices in a position to support technology. Assistance was originally incorporated by the look of the HTML markup vocabulary for that very purpose without needing tokenization of being precisely and quickly listed for meta-tags.

In Desktop search, several options include meta-tags to supply a means for writers to help modify the way the search engine will catalog content from numerous documents that's not apparent in the document information. Desktop search is more under the control of the user, while Internet search engines must focus more on the full text index.

Conclusion. This survey deals with the fundamentals of the information retrieval and genetic algorithm to solve The research areas in web search and various issues . It also deals with big information search which are promising research areas.

The IR methods, despite their sophisticated functions, need enhancement and modification to be able to accomplish greater efficiency and supply increased acceptable solutions to inquirer. Planning to attain greater efficiency, methods and more versatile versions are required. Fuzzy-set construction continues to be demonstrated as appropriate formalism for managing and acting imprecision and vagueness, the new subjects of data access.

Numerous studies considering numerous programs of fuzzy-set engineering performed and have now been started, some current described in this essay. Fuzzy methods in most stages of IR's implementation so raises consumer satisfaction and has taken enhancement of IR outcomes. Major methods are a great device to remove low-specific information from information. Their own capability develop to estimation and enhance may be used to design Web research person. Implicit information, like the press-flow, created throughout the web-browsing actions might be used to keep an eye on each single user's choices. Simultaneous deployment of fuzzy set techniques for better document modeling and genetic algorithms for query optimization brings a significant contribution to the ultimate goal of web search: bringing knowledge to man.

Search Engine utilizing a method and system for efficient storage and retrieval of data. Furthermore the system comprises a record file, an index file, a duplicate segment file and access to a network of computers. The index files contains locations of data items, pointers to other index files, or an empty designation. The index files are arrays that contain locations corresponding to a predetermined range of characters with which the data items may be formed. Data items are stored according to the character strings of each data item. The first portion of a data object is indexed according to the indexing method of the present invention while a second portion of the data object is indexed according to another known database technology.

References

[1]. IEE , MultiAgent system, 2009.

[2]. Applying genetic algorithm in query improvement problem. Int. Journal on Information Technologies and Knowledge 1(12), 309–316 (2007)

[3]. Dept. of Accounting & Inf. Syst., Virginia Polytech. Inst. & State Univ.,2011

[4]. Nicholas J. Belkin and W. Bruce Croft. Information filtering and information retrieval: two sides of the same coin? Communications of the ACM, 35(12):pp. 29–38, December 1992.

[5]. Search engine website, 2016.

[6]. MICHAEL S. LEW Leiden University, The Netherlands NICU SEBE University of Amsterdam, The Netherlands CHABANE DJERABA LIFL, France and RAMESH JAIN University of California at Irvine, USA,2016)

[7]. Beel, Jöran and Gipp, Bela and Wilde, Erik (2010). "Academic Search Engine Optimization (ASEO): Optimizing Scholarly Literature for Google Scholar and Co."(PDF). Journal of Scholarly Publishing. pp. 176–190. Retrieved April 18, 2010.

[8]. E. Greengrass. Information retrieval: A survey. DOD Technical Report TRR52-008-001, 2001.

[9]. Gloria Bordogna and Gabriella Pasi. Modeling vagueness in information retrieval. pages 207–241, 2001.