ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)

Vol 8, Issue 11, November 2021

A Review Approach for Information System Recovery using Artificial Intelligence Algorithm And Grouping of Documents

[1] Pankaj Ameta, [2] Pankaj Kumar Vaishnav, [3] Prashant Sharma

[1][2][3] Pacific University, India

Email: [1] pankajameta1989@gmail.com, [2] panvas23@pacific-it.ac.in, [3] prashant.sharma@pacific-it.ac

***Abstract*----The need for expertise in Information Retrieval Systems pushed researchers to Analyze intelligent systems that seek to incorporate and use such knowledge in order to optimize the system. In this paper, it is shown an evolutionary system (EVS), and the Results obtained with the construction of a system of this nature. In this paper a contribution in the field of Information Retrieval (IR), Proposing the development of a new system using evolutionary techniques, implement A system for unsupervised learning type, to group documents in an information Retrieval System (IRS) where Their groups and number of are unknown a priori by the System. The results prove the feasibility of building a large-scale application of this type in Order to integrate it into a knowledge management system that needs to handle Controlled document collections.**

***Keywords*---- Information Retrieval, EVS, IRS**

I. INTRODUCTION

In a few years, the world-wide network technology - World Wide Web - has become a universal tool for all kinds of cultural, professional and commercial activities. The process of digitalization and the transformation of documents that is being carried out are two clear examples of the revolution of the information, which has allowed its access to an unlimited number of users. In this context, Information Retrieval (IR) can be defined as the problem of selecting information from a storage mechanism in response to queries made by the user [1]. It is very important to distinguish it from Data Recovery, and to mention the differences between these concepts, as indicated by C. J. van RIJSBERGEN [2].

Information retrieval systems (IRS) are a class of information systems that deal with databases composed of documents, and process user queries allowing them to access the relevant information in an appropriate time interval. The main of these limitations is its binary recovery criterion, which is very sharp and strict so it can be considered a data recovery system rather than information. Because of this, other paradigms have been designed in order to extend this mode of recovery and overcome its problems, such as the vector model [3] the probabilistic model and the diffuse IR model.

In recent years, there has been a growing interest in the application of Artificial Intelligence (IA) techniques [4], [5], and Data Mining [6], [7] to the field of IR in order to address the shortcomings of traditional IRS. An example of this is the Automatic Learning paradigm [8] based on the design of systems with capacity to acquire knowledge by itself and the advanced techniques of Data Mining [9]: Neural Networks Neural Networks), Evolutionary Algorithms (AE), Grouping Methods, Fuzzy Logic, Inductive Reasoning among others, in order to extract different types of knowledge in the information being processed. Evolutionary Algorithms (AE) [10] are not specifically automatic learning algorithms, but they offer a powerful and domain independent search methodology that can be applied to a large number of learning tasks [11]. For these reasons, the application of EAs to RI has increased in recent years.

In this paper, the grouping of documents is studied, meaning the process of finding groups within a collection of documents, based on criteria of similarity or distance between them, without requiring a priori knowledge of other characteristics. The development of an Evolutionary System (EVS) is presented. Using evolutionary techniques, it implements a learning system of the unsupervised type, in order to group documents, where groups are unknown a priori by the system. One of the contributions of the proposed system is based on the representation of individuals in the system, typical of genetic programming (PG). The proposed method may be presented as an alternative to traditional clustering methods, such as hierarchical and / or partitioned algorithms based on known algorithms such as K-Means.

ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)

Vol 8, Issue 11, November 2021

II. INTRODUCTION TO INFORMATION RETRIEVAL SYSTEMS

To added the distribution of information through the so-called "information highways", and the ever-lower cost of storage media. All this places us within a developing environment of electronic information that can be accessed by automatic means. Another aspect that we have to consider is the diversification of the media, which brings with it a greater amount of non-standard information, image, sound, text, etc.

Information Retrieval Systems (IRS) are a class of information systems that deal with databases composed of documents and process user queries allowing them to access relevant information in an appropriate time interval (see Figure 1) . These systems were originally developed in the 1940s with the idea of assisting managers of scientific documentation.

An IRS allows the retrieval of information, previously stored, by means of a series of queries (queries) to the documents contained in the database. These questions are formal statements of expression of information needs and are often expressed through a query language.

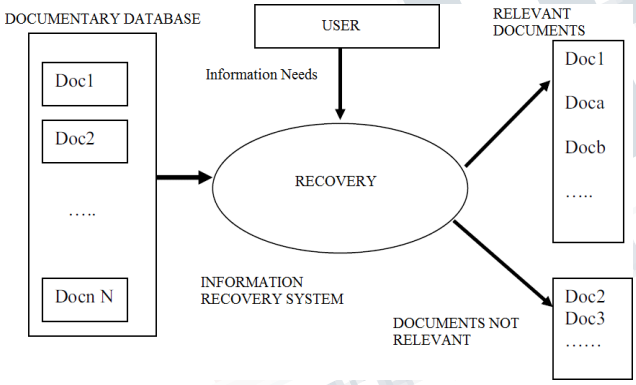

Fig. 1. Information Retrieval Process

- ***Components of an Information Retrieval System***

An IRS is composed of three main components: the documentary database, the query subsystem and the matching or evaluation mechanism (see figure 2.).

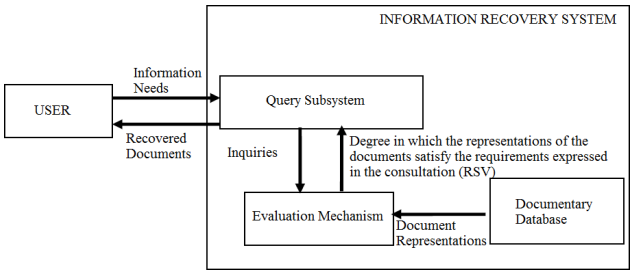

Fig. 2. Generic Composition of an Information Retrieval System

3.1 The documentary database

A document contains data of a usually textual form, although the technological evolution has favored the profusion of multimedia documents, adding to the text photographs, graphic illustrations, animated videos, audios, etc. These documents are not entered directly into the IRS, but will be represented by elements called descriptors. The reason for using these descriptors is to give a greater efficiency to the recovery process, allowing the search time in it to be much smaller.

3.2 The query subsystem

This subsystem consists of the interface that allows the user to formulate their queries to the IRS and by a parser that takes the query written by the user and breaks it down into its component parts. To perform this task, it includes a query language that collects all the rules to generate queries the methodology for selecting the relevant documents.

3.3 The evaluation mechanism

At this point, we have a representation of the content of the documents in our documentary base and also a representation of the queries we want to perform of the query subsystem. What remains to be resolved is the selection of documents that are considered relevant, among the documents that form the documentary base, according to the criteria of our consultation.

3.4 Classification of Information Recovery Systems

There are several models of IR, each has its advantages and disadvantages, we will comment on several of the existing models and analyze the components that form them.

3.5 The boolean model

This model uses Boolean algebra theory. This deals with propositions, associated by means of logical operations AND, OR, NO, SI ..... THEN, and that, therefore, allows

ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)

Vol 8, Issue 11, November 2021

calculations of algebraic type.

3.6 The vector space model

Salton was the first to propose IRS based on vector space structures in the late 1960s within the framework of the SMART project [12]. Since documents can be represented as vectors of terms, documents can be placed in a vector space of m dimensions, with as many dimensions as components have the vector.

3.7 The probabilistic model

This model improves IRS performance through the use of information from the statistical distribution of terms in documents. The frequency of occurrence of a term in a document or set of documents could be considered a relevant data when establishing a query to the documentary database.

The probabilistic model uses probability theory to construct the search function and to establish its mode of use [13] [14]. The information used to compose the search function is obtained from knowledge of the distribution of the indexing terms throughout the collection of documents or a subset of it.

III. EVOLUTIONARY SYSTEM

We tried to implement an unsupervised system that allows the evolutionary grouping of the documents of an IRS. Our focus applies a fitness function that combines the concepts of distance and of similarity between documents. The proposed system will be composed of the Following elements and characteristics:

- Individuals
- Population size.
- Number of generations.
- Production operators.

4.1 Factors to consider in the proposed evolutionary model

There are six factors to consider in our system that are important for its implementation, the quality of these depends on the system to better fulfill its purpose. These are:

- The set of terminals.
- The set of primitive functions.
- The measure of adaptation or fitness
- The parameters to control execution
- Elitism.
- The criteria for completing an execution.
- Criteria for selecting initial documents

4.2 Functional specification of the evolutionary system

4.2.1 Requirements Catalog

The catalog of requirements for the developed system, whose purpose is to specify the functional needs of the system, has components: The catalog of functional requirements of the system, the catalog of safety requirements, and the catalog of control requirements.

4.2.2 Decomposition into subsystems

The system is basically decomposed into four subsystems (plus one of access control): A document processing subsystem, from which all proposed methodological techniques of document processing are carried out, in order to obtain the characteristic vectors of all Documents. A subsystem to create the initial population (Generation 0), from which Creates the set of individuals that make up the initial population, using the Chromosome model proposed, and using an approach that Randomly place all the documents of the IRS, always having Individuals without repeating in the tree structure. A subsystem to process the Evolutionary System, from which the following generations of individuals from the system are created, starting from the initial population, applying the fitness operator proposed in this paper, and the parameters that describe and control the system as well as the proposed production operators (crossing and mutation) that create a new individual from the current generation.

A result management subsystem, from which the results obtained with our system are compared to the supervised clustering algorithm (Kmeans), taking into account the number of groups that are processed and interacting with the processing subsystem of the evolutionary system, to generate the results. In Figure 3 we show each of the developed subsystems previously described.

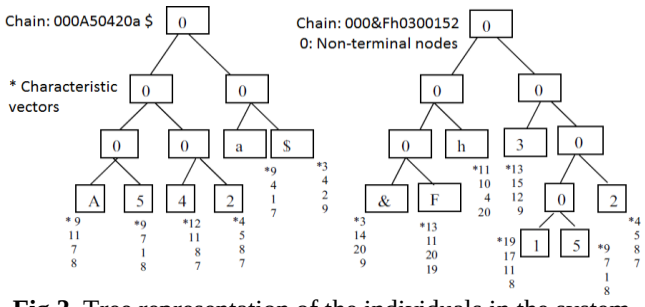

Fig.3. Tree representation of the individuals in the system

ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)

Vol 8, Issue 11, November 2021

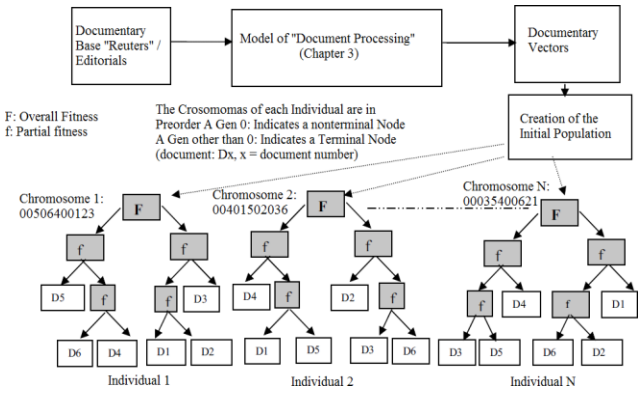

Fig. 4 Initial population of individuals in the Evolutionary System

We consider that the representation of individuals is one of the original contributions, which unlike the other proposals made on GA which, as far as we know, use only fixed- length chains. In this way, our internal representation is based on chains whose length depends on the number of documents to be processed. These chains are transformed by means of production rules into unbalanced trees representing individuals of different shapes and sizes that we can evolve, letting it be the evolution itself that decides which is the best of the configurations.

There are a number of restrictions that limit the creation of such individuals. Thus, for example, two equal individuals should not be created in the initial set. The rules of production that guarantee that this condition is fulfilled oblige the grammar of creation of the nodes of each individual to be made by making a tour of the tree in "Preorder".

Another novelty that is added to the system is the possibility of establishing the so-called "threshold of complexity", assigned by the user, in order that each individual can be structurally limited so that the system does not consume too much time with it. This threshold of complexity is applied to limit the depth of the tree.

The initial population will be formed by the set of individuals, generated at the beginning in the so-called generation 0, such as, for example, it can be seen in figure 4 that shows the process, the structure that each individual and the values which characterize each document, which are collected from the pre-processing of the documents discussed above.

IV. EVOLUTIONARY BEHAVIOR MODEL

In the diagram we show the behavior of the system when dealing with a task or a process that we manage, and describe all the functionality of the application. The use cases are represented by the ellipses and the actors are represented by the human figures. This way, in our system the actors could manage documents, manage new individuals of the population, manage generations of the evolutionary algorithm, manage the evolution of the adaptation function, manage the results of the system, and disconnect from the application. The genetic system use diagram is shown in figure 5.

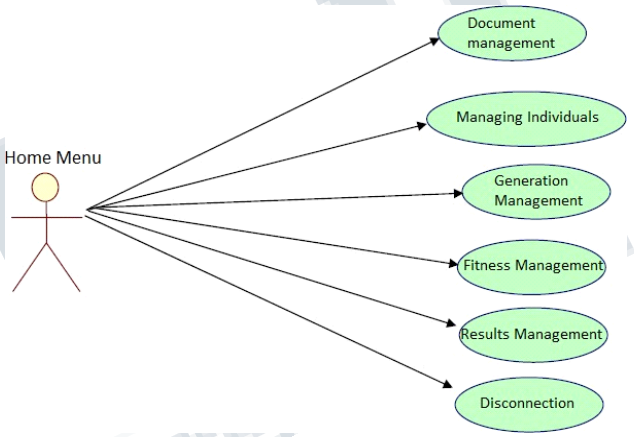

Fig.5. Genetic System Use Case Diagram

V. RESULT AND ANALYSIS

For the actual tests we used documents from the Reuters collection 21578 [15], taking the distributions that most scatter data present (distribution 21, distribution 2, distribution 20, distribution 8), and documents from the collection of newspaper editorials the world.

The reason for using Reuters 21578 is that this collection constitutes one of the de facto standards within the domain of automatic document categorization, and as such is used by many authors of the subject as a "touchstone". This collection has been described, processed, and commented is composed of 21578 documents, distributed in 22 files, where each document has 5 different categorization fields called Stock Market Value, Organization, Person, Place and Topic.

In addition, we have used the English collection of the editorials of the newspaper from the years 2015 and 2017, which has been manually classified in the document processing stage.

ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)

Vol 8, Issue 11, November 2021

Thus, in the actual tests, we used the Reuters 21, Reuter 2, Reuter 20 and Reuters 8 distribution because they are the ones that contain the highest dispersion rate among all distributions. Samples of documents from the collections (different amounts of documents) were analyzed, and the algorithm's behavior was analyzed in order to maximize the number of hits, the average number of hits between the different tests, and finally to draw graphs with the evolution of the system.

For the collection of newspaper publishers, we apply the same criteria, using the main categories of the Eurovoc thesaurus. To do this, it was necessary to perform a manual classification in each of the editorials that compose the collection (1402 publishers), according to the method described, in order to compare the results of the algorithm with the actual data.

6.1 Experiment environment

Within the experimentation environment there must be a user that provides the documents that are to be grouped. The role of the user contributing documents will be represented by the document samples of each documentary collection represented by their feature vectors.

We will analyze other aspects related to the experimental methodology. Due to the simulation nature of the evolutionary system, its operation is pseudo random. This results in the need to perform several executions, with different seeds as input for the random number generator, until reaching the optimum solution.

In this way, the experiments with the evolutionary algorithm were realized realizing five executions on each one of the samples taken from the experimental collections. The output of the experiment is the best fitness obtained and the speed of convergence or generation where the best fitness is found. Finally, as a measure of the quality of the algorithm, it is possible to take care of the best solution obtained and the robustness, that is, the average quality of the algorithm (average of the values of the different solutions obtained). In the previous experiments we have done using the Reuters collection we have verified that our AG works correctly with different samples of documents, and that the average fitness behavior in several runs of the algorithm is correct, as we can see in figure 6. Note in the figure the way the average fitness returns with each of the samples of processed documents.

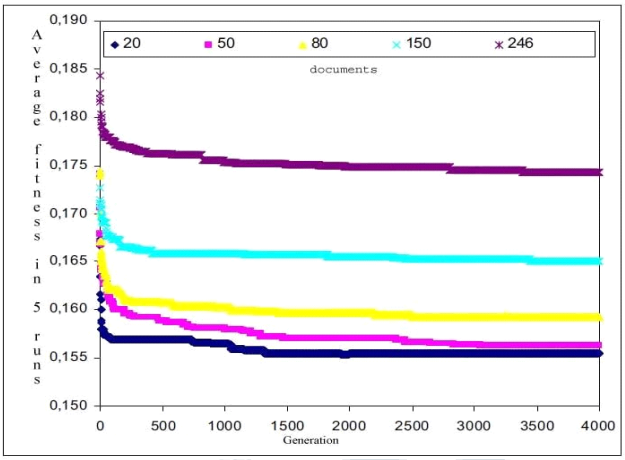

Fig.6 Evolution of Fitness of the AG, for different numbers of documents

In the experiments within our work environment, samples of randomly collected documents were used from "very few, few, many and many documents", with the requirement that they belong only to two of the categories of the Reuters distribution or Editorials in English. Each of the samples was processed with the 5 different seeds, and each of the results was compared with the "Kmeans" method.

In order to show the obtained results, we have defined the success factor as follows:

$$a_i - a0$$

Where ai corresponds to the number of average hits obtained by the AG, and a0 is the minimum value of average hits obtained by said AG. Thus, in Figure 7, we show the success factor versus the values of α. And we can see that the best values in each of the samples are presented when we apply a value of α close to the value of 0.85. These results occur because the value of the best fitness takes the minimum value by using the value of **α**, from which the contribution of the metric of the inverse of similarity does not compensate that of the metric of distance, which causes that the value Of fitness does not tend to stabilize, as we see in Figure 8, which shows the behavior of the best fitness by varying the values of **α** with different samples of documents.

ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)

Vol 8, Issue 11, November 2021

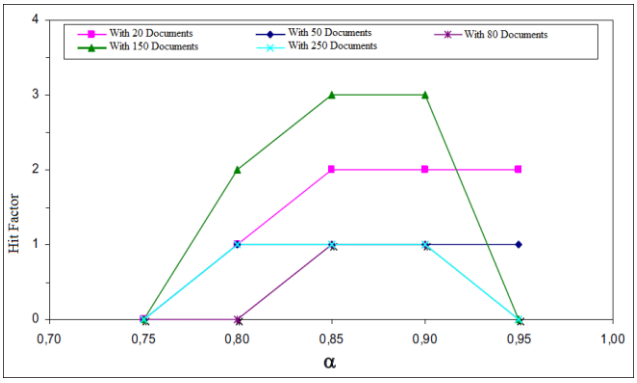

Fig.7 Dependence of the number of hits with the value of α in the AG, taking different samples of documents from the distribution 21 of the Reuters collection

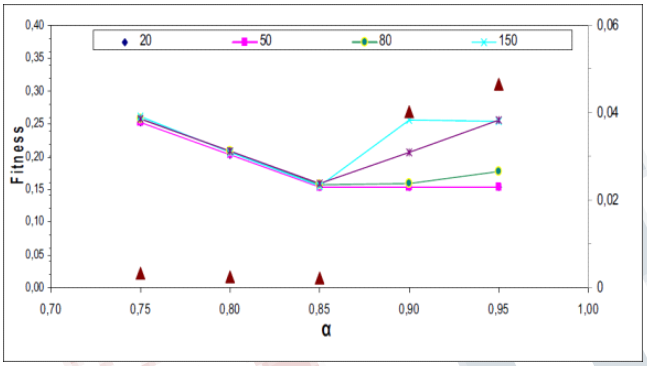

Fig.8 Better Fitness vs values for different samples of documents from the Reuters collection: Distribution 21. It can be seen that there is an increase in fitness dispersion for values higher than 0.85, due to the greater contribution of the Distance Euclidean that makes more insensible to the Fitness to find the groupings.

VI. CONCLUSION

In general our algorithm provides good results with both documentary collections. We can affirm that at the end of the evolution, and using all the parameters studied and configured for the algorithm, we offer a high effectiveness in the grouping of both documentary collections, being therefore the algorithm stable and robust independently of the sample or documentary collection used. In this paper the solution has been proposed, making use of an evolutionary system for the problem of the grouping of documents. A system has been constructed, which makes use of a genetic algorithm that allows the grouping of documents in an unsupervised way.

REFERENCES

[1] W. B. Frakes, R. Baeza and Yates, Information Retrieval: Data Structures & Algorithms, Prentice Halls, 2004.

[2] C. J. v. RIJSBERGEN , INFORMATION RETRIEVAL, London: Butterworths, 1979.

[3] S. F. Dierk, "The SMART retrieval system: Experiments in automatic document processing," in *IEEE Transactions on Professional Communication*, Princeton, N.J, 1972.

[4] Winston and P. Henry, Artificial intelligence, IBEROAMERICANA: ADDISON WESLEY, 1994.

[5] R. J. Stuart and N. Peter, Artificial Intelligence : A Modern Approach, Prentice Hall, 1996.

[6] B. W. Michael, Survey of Text Mining, Verlag New York: Springer, 2004.

[7] B. W. Michael and C. Malu, Survey of Text Mining II, Verlag London: Springer, 2008.

[8] T. M. Mitchell, Machine Learning, McGraw-Hill Education, 1997.

[9] C. Felix, V. Alfredo, N. Angela and F. Mujica, "Applying Data Mining Techniques to e-Learning Problem," in *Evolution of Teaching and Learning Paradigms in Intelligent Environment*, Springer, 2007, pp. 183-222.

[10] A. E. Eiben, R. Hinterding and Z. Michalewicz, "Parameter control in evolutionary algorithms," *IEEE Transactions on Evolutionary Computation*, vol. 3, no. 2, pp. 124- 141, 1999.

[11] J. J. Grefenstette, Genetic Algorithms for Machine Learning, Netherlands: Kluwer Academic Publishers, 1995.

[12] S. F. Dierk, "The SMART retrieval system: Experiments in automatic document processing," *IEEE Transactions on Professional Communication*, vol. 15, no. 1, pp. 17-17, 1972.

[13] N. Fuhr, "Probabilistic Models in Information Retrieval," *Computer Journal*, vol. 35, no. 3, pp. 243-55, 1992.

[14] A. Bookstein, "Outline of a General Probabilistic Retrieval Model," *Journal of Documentation*, vol. 39, no. 2, pp. 63-72, 1983.

[15] D. Lewis, "text categorization test collection," Reuters, 1997.