Ms. Anagha N. Chaudhari, Prof. Anuradha D.Thakare / International Journal of Engineering Research and Applications (IJERA) ISSN: 2248-9622 www.ijera.com
Vol. 2, Issue 6, November- December 2012, pp.872-877

Introducing GA based Computational Model using SVD Technique for Dimensionality Reduction

Ms. Anagha N. Chaudhari*, Prof. Anuradha D.Thakare**
*(Department of ME Computer Engineering,Pune University, Pune-44)
**(Department of Computer Engineering, Pune University, Pune-44)

ABSTRACT

This paper presents a survey of dimensionality reduction by SVD and how high dimensional data is transformed into a low dimensional vector space. SVD technique can be applied to many computations like Pseudo inverse, solving homogeneous linear equations, Low-rank matrix approximation, Nearest orthogonal matrix etc which is necessary for cluster analysis. This paper also tries to reflect the research profile of Evolutionary algorithm based approach for dimensionality reduction. Based on this survey, SVD based system for dimensionality reduction is proposed.

Keywords- Clustering, Dimensionality Reduction, Information Retrieval, Latent Semantic Indexing, Singular Value Decomposition.

1.INTRODUCTION

High-dimensionality indexing of feature spaces is critical for many data-intensive applications such as content-based retrieval of images or video from multimedia databases and similarity retrieval of patterns in data mining. Similarity-based retrieval has become an important tool for searching image and video databases, especially when the search is based on content of images and videos, and is performed using low-level features, such as texture, color histogram and shape. The application areas of this technology are quite diverse, as there has been a proliferation of databases containing photographic images. SVD became very useful in Information Retrieval (IR) to deal with linguistic ambiguity issues. IR works by producing the documents most associated with a set of keywords in a query. Keywords, however, necessarily contain much synonymy (several keywords refer to the same concept) and polysemy (the same keyword can refer to several concepts). the characteristics of polysemy and synonymy that exist in words of natural language have always been a challenge in the fields of IR and data mining. In many cases, humans have little difficulty in determining the intended meaning of an ambiguous word, while it is extremely difficult to replicate this process computationally[9]. A technique known Latent Semantic Indexing (LSI) addresses these problems by calculating the best rank-l approximation of the keyword-document matrix using its SVD. This produces a lower dimensional space of singular vectors that are called eigen-keywords and eigen-documents. Each eigen-keyword can be associated with several keywords as well as particular senses of keywords. In the synonymy example above, "cat" and "feline" would therefore be strongly correlated with the same eigen-keyterm.

Singular-value decomposition is used to decompose a large term by document matrix into 50 to 150 orthogonal factors from which the original matrix can be approximated by linear combination; both documents and terms are represented as vectors in a 50- to 150- dimensional space. Queries are represented as pseudo-documents vectors formed from weighted combinations of terms, and documents are ordered by their similarity to the query. Initial tests find this automatic method very promising [8].

Different measures for evaluating the performance of IR are: Precision is the fraction of the documents retrieved that are relevant to the user's information need. Recall is the fraction of the documents that are relevant to the query that are successfully retrieved and Fall-Out is the proportion of non-relevant documents that are retrieved, out of all non-relevant documents available.

2.RELATED WORK

In the year 2003, Peg Howland, Moongu Jeon, Haesun Park took MEDLINE database in their paper [11] and they described and compared four methods like Full Dimensional Space, Orthogonal Centroid, Centroid and LSA/GSVD. From MEDLINE database, different medical categories were considered like Heart Attack, Colon Cancer, Diabetes, Oral Cancer and Tooth Decay and each category was having 40 documents. So total dimensions with documents were 7159*200. The numbers of dimensions with documents were compared. The Full Dimensional Method includes 7159*200 dimensions whereas Orthogonal Centroid and Centroid methods consists of 5*200 dimensions.The effectiveness of LSA/GSVD is proved in the form of 4*200 dimensions. For each method smaller trace value (Sw) & larger trace value (Sb) is computed. Ratio of (Sw) and (Sb) is also calculated for each method, which highlights the importance of LSA/GSVD.

By Algorithm LSA/GSVD the dimension 7519 is dramatically reduced to 4, which is one less than the number of classes. The other methods reduce the dimension to the number of classes, which is 5. The ratio (Trace Sw/Trace Sb) is the approximate optimality measure. The optimality measure for other three methods were 0.09, 1.5 and 1.8 respectively whereas optimality measure of LSA/GSVD method is increased upto 79.We observe that the ratio is strikingly higher for the LSA/GSVD reduction than for the other methods, and that, the ratio produced by each of the three dimension reduction methods is greater than that of the full-dimensional data [11].

In 2009, Habiba Drias, Ilyes khennak, Anis Boukhedra [12] explained in their paper, that all experiments were conducted on 30,000 documents with at most 20 terms. They suggested that in **Classical IR (CL-IR)** the Runtime increases as the number of documents increases. Next they proposed, **Genetic Algorithm (GA-IR)** where varied numbers of documents were used in terms to check the solution quality & runtime. Here Runtime decreases as the number of documents increases. GA-IR computes optimal solution for certain queries. Then the authors introduced the more effective algorithm as **Memetic Algorithm (MA-IR)** where varied numbers of documents were used in terms to check the solution quality & runtime. MA-IR computes always optimal solution.MA-IR outperforms GA-IR.

In 1994, Jose L. Rebeiro Filho, Philip C [13], explained in their paper, two approaches like Simulated Annealing Technique and Sequential & Parallel Genetic Algorithms. **Simulated Annealing Technique** uses a thermodynamic evolution process to search minimum energy states, whereas **Sequential & Parallel GA** is based on natural selection principles, these algorithms evolves throughout generations and targets large and very complex datasets.

3. SINGULAR VALUE DECOMPOSITION

Dimensionality reduction methods are usually based on a linear transformations followed by the selection of a subset of features, although nonlinear transformations are also possible. Techniques based on linear transformations, such as the Karhunen-Loeve (KL) transform, the Singular Value Decomposition (SVD) method, and Principal Component Analysis (PCA), have been widely used for dimensionality reduction and Data Compression. SVD is extraordinarily useful and has many applications such as data analysis, signal processing, pattern recognition, image compression, weather prediction, and Latent Semantic Analysis or LSA (also referred to as Latent Semantic Indexing or LSI). SVD is a successful technique arising from numerical linear algebra that is used in Latent Semantic Indexing (LSI).LSI can overcome the problems by using statistically derived conceptual indices instead of individual words and provide a dimension reduced space. Genetic algorithm can be used in combination with the reduced LSI and improve clustering efficiency and accuracy[1]. Formally, the singular value decomposition of an m×n real or complex matrix M is a factorization of the form shown below in Fig.1.

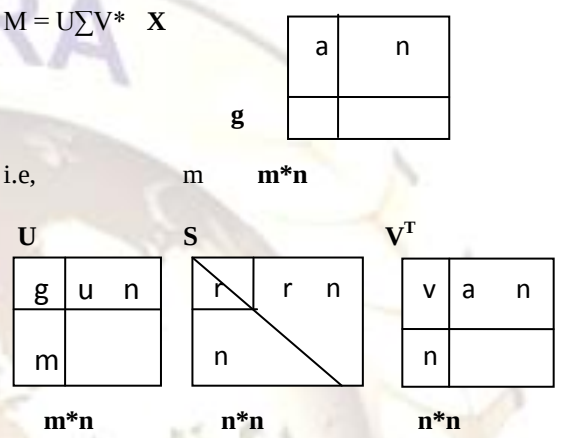

Figure 1: Singular Value Decomposition Formula Structure

where U is an m×m real or complex unitary matrix, Σ is an m×n rectangular diagonal matrix with nonnegative real numbers on the diagonal, and V* (the conjugate transpose of V) is an n×n real or complex unitary matrix. The diagonal entries Σi,i of Σ are known as the singular values of M. The m columns of U and the n columns of V are called the left singular vectors and right singular vectors of M, respectively.

The singular value decomposition and the eigen decomposition are closely related. Namely:

- The left singular vectors of M are eigenvectors of MM*
- The right singular vectors of M are eigenvectors of M*M

The non-zero singular values of M (found on the diagonal entries of Σ) are the square roots of the non-zero eigen values of both M*M and MM*.

Applications which employ the SVD include computing the pseudo inverse, least squares fitting of data, matrix approximation, and determining the rank, range and null space of a matrix. LSI attempts to project the documents of a collection into a lower dimensional space in order to improve retrieval performance. The lower dimensionality of the space is intuitively desirable; terms that are related should be brought closer together (the cluster hypothesis).

3.1 Computing the Singular Value Decomposition

1. Compute the QR factorization of the matrix A. (A = QR)
2. Reduce the upper triangular matrix R to a bidiagonal matrix B using orthogonal transformations.(R =U1BV1)
3. Reduce the bidiagonal matrix B to a diagonal matrix using an iterative method.(B =U2∑V2) [7].

The computation of the SVD can then be represented as

A = QR
= Q(U1BV1)
= Q(U1 (U2∑V2)V1)
= (QU1U2)∑(V2V1)
= U∑V

4. SINGULAR VALUE DECOMPOSITION FOR DIMENSIONALITY REDUCTION

- **Mathematical basis**: Set Theoretic Model -> represent documents as sets of words or phrases. Similarities are usually derived from set-theoretic operations on those sets. Eg: Standard Boolean model, Extended Boolean Model & Fuzzy Retrieval Model. Algebraic Models -> represent documents and queries usually as vectors, matrices, or tuples. The similarity of the query vector and document vector is represented as a scalar value. Eg:Vector Space Model & LSI (SVD).
- **Probabilistic Models** : Treat the process of document retrieval as a probabilistic inference. Similarities are computed as probabilities that a document is relevant for a given query. Eg: Binary Independence Model & language Models.
- **Feature-based retrieval models**: view documents as vectors of values of feature functions (or just features) and seek the best way to combine these features into a single relevance score, typically by learning to rank methods.

5. SVD FOR CLUSTERING

As in the architecture given below, after providing the input documents when singular value decomposition is applied to the input then this technique decomposes the large term by document matrix into a set of K orthogonal factors. Then the less important dimensions corresponding to "noise" due to word choice variability are ignored. A reduced rank approximation to the original matrix is constructed by dropping these noisy dimensions. Then after applying the fitness function to the reduced population, we use Genetic Operators for n generations to form good quality clusters and then most relevant documents are retrieved.

Data mining is component of the knowledge discovery in databases process concerned with the algorithmic means by which patterns are extracted and enumerated from data. This knowledge discovery process has several steps. One of the important steps is to clustering the data.

The simplest definition is shared among all and includes one fundamental concept: the grouping together of similar data items into clusters. The greater the similarity within a group and greater the difference between groups, the better the clustering.

Clustering Singular Value Decomposition (CSVD), for indexing data by reducing the dimensionality of the space. This method consists of three steps: partitioning the data set using a clustering technique, computing independently the SVD of vectors in each cluster to produce a vector space of transformed features with reduced dimensionality, and constructing an index for the transformed spaced. From the information retrieval viewpoint, experiments demonstrate that CSVD achieves better recall and precision than simple SVD for the same number of retained dimensions [3].

Clustering represents an unsupervised classification technique which is defined as group n objects into m clusters without any prior knowledge. A novel methodology, Clustering Singular Value Decomposition (CSVD), is proposed for approximate indexing of numeric tables with high dimensionality.

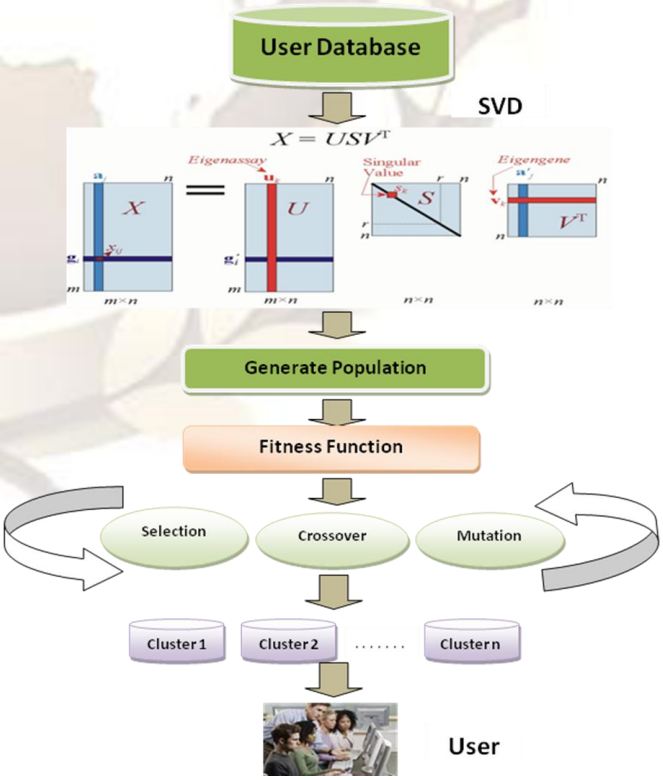

Figure2: General System Architecture

This method achieves additional dimensionality reduction than those based on SVD for a fixed Normalized Mean Squared Error (NMSE) by exploiting the local structure of heterogeneous datasets. In the proposed method, the feature space is first partitioned into multiple subspaces and a dimension reduction technique such as SVD is then applied to each subspace. The effectiveness of the proposed CSVD method is validated with datasets consisting of feature vectors extracted from three benchmark databases containing feature vectors extracted from remotely sensed images [3].

Like the classical LSI, there are some parameters that may affect the performance of a clustered SVD retrieval system. The most important parameters are the rank of the truncated SVD matrices and the number of clusters chosen in the partial clustered SVD retrieval [10].

6. PROPOSED SYSTEM

Employing the SVD based document representation, LSI can overcome the problems by using statistically derived conceptual indices instead of individual words and provide a dimension reduced space.

In this paper the proposed system is not only describing the singular value decomposition but also Hybrid SVD (Modified SVD) and Genetic algorithm based SVD.The most well-known and widely used algorithm for computing the Singular Value Decomposition (SVD) A --- U ~V T of an m x n rectangular matrix A is the Golub-Reinsch algorithm (GR-SVD) [2].

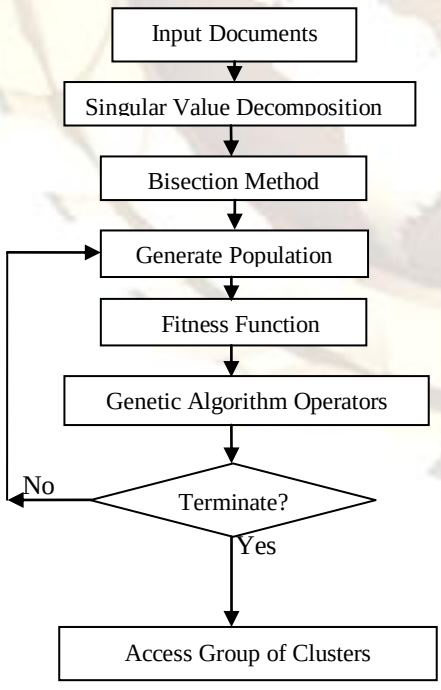

Figure 3: Proposed System Architecture

In this paper, an improved version of the original GR-SVD algorithm is presented. The new algorithm works best for matrices with m >> n, but is more efficient even when m is only slightly greater than n (usually when m ~ 2n) and in some cases can achieve as much as 50 percent savings. If the matrix U ~s explicitly desired, then n^2 extra storage locations are required, but otherwise no extra storage is needed.

This algorithm consists of two phases. In the first phase one constructs two finite sequences of Householder transformations like P^k; k = 1 to n and Q^k; k = 1 to n-2. P (I) zeros out the sub diagonal elements in column i and Q (J) zeros out the appropriate elements in row j. The system is shown in the above Fig 3.

6.1 The Modified SVD (MOD-SVD)

Our original motivation for this algorithm is to find an improvement of GR-SVD when rn >> n. In that case some improvement is possible:

Each of the transformations $P^{(i)}$ and $Q^{(i)}$ has to be applied to a sub matrix of size (m-i+1) * (n-i+1). Now, since most entries of this sub matrix are ultimately going to be zeros, it is intuitive that if it can somehow be arranged that the $Q^{(i)}$ does not have to be applied to sub diagonal part of this sub matrix, then we will be saving a great amount of work when m>>n. This can indeed be done by first transforming A into upper triangular form by Householder transformations on the left:

$$L^T[A] \rightarrow \begin{bmatrix} \triangle \\ 0 \end{bmatrix} \equiv \begin{bmatrix} R \\ 0 \end{bmatrix}$$

Where R is n*n upper triangular and L is diagonal and then proceed to bidiagonalize R.

The important difference is here we are working with much smaller matrix R than A, and so it is conceivable that the work required to bidiagonalize R is much less than that originally done by the right transformations when m >> n [2].

6.2 Bisection Method

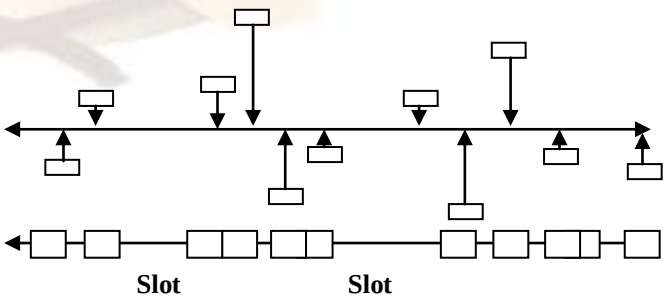

Figure 4: Slots preserved when a dataset is projected onto a singular vector

As given in the above Fig 4, first the truncated SVD using the appropriate rank k, must be found for any dataset and then the gaps between entries of the left singular vectors are calculated. If a gap between two entries is large enough, a division is placed between the corresponding rows of the original matrix. Again, clustering the columns of a matrix is similar. The only difference being that the algorithm uses gaps in the right singular vectors to determine where to divide the columns [6]. Genetic algorithm (GA) belongs to the search techniques that mimic the principle of natural selection and heredity. It performs search in complex, large and multimode landscapes, and provides near-optimal solutions for objective or fitness function [9].

6.3 Algorithm

LSI is also used to perform automated document categorization. Dynamic clustering based on the conceptual content of documents can also be accomplished using LSI.

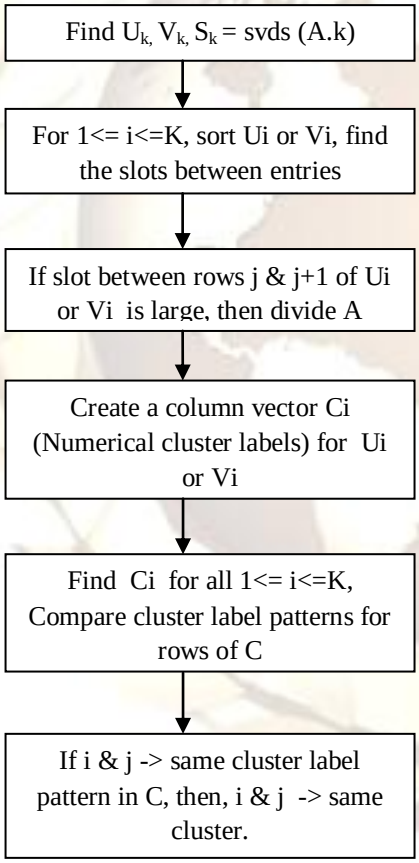

Figure 5: SVD Bisection method algorithm

7. ADVANTAGES AND DISADVANTAGES

Because it uses a strictly mathematical approach, LSI is inherently independent of language. LSI is not restricted to working only with words. It can also process arbitrary character strings. Early challenges to LSI focused on scalability and performance. Another challenge to LSI has been the alleged difficulty in determining the optimal number of dimensions to use for performing the SVD. As using SVD dimensionality technique, we can easily reduce the keywords of huge documents through the concept of LSI (Polysemy & Synonymy) and also by using the Bisection Method [6], we can reduce the number of documents in the datasets as we already have the reduced keywords sets with us. After applying these two concepts, when Genetic Algorithm (Global Optimization Algorithm), is applied on the reduced dataset, the process undergoes with number of generations so as to give us better quality & relevant (crisp) clusters at the end. Hence, the huge & multi-dimensional dataset can be reduced with the above mentioned approach and it results into fast and efficient searching mechanism with relevant & most required results.

8.FUTURE SCOPE

SSVD seeks a low-rank, checkerboard structured matrix approximation to data matrices. The desired checkerboard structure is achieved by forcing both the left- and right-singular vectors to be sparse, that is, having many zero entries. The sparsity implies selection of important rows and columns when forming a low rank approximation to the data matrix. Because the selection is performed both on the rows and columns, our SSVD procedure can take into account potential row–column interactions and thus provides a new tool for bi clustering. The effectiveness of SSVD has been demonstrated through simulation studies and real data analysis. There are a few potential directions for future research. First, SSVD is developed as an unsupervised learning method. It is interesting to evaluate its usage as a dimension reduction tool prior to use of classification methods [4],[5]. How to determine the best rank of the truncated SVD matrices is an unsolved classical problem in LSI. The number of clusters chosen for a particular query is determined by the initial query on the centroid vector. A suitable threshold value could be determined based on certain heuristics and experiments, which could be an interesting topic for future study [10].

9. CONCLUSION

SVD technique is mainly used for reducing the huge and high dimensional datasets as a dimensionality reduction technique.SVD is very useful tool for image compression along with many computations like Pseudo inverse, solving homogeneous linear equations, low-rank matrix approximation, Nearest orthogonal matrix, etc.
Our proposed system is concentrating on first reducing the datasets using SVD technique i.e, keywords based reduction (LSI method) and then

proceeding further for reducing the number of documents using bisection method.

Once we have with us reduced keywords and documents, we get the population ready for applying fitness functions to identify the best fittest chromosomes for applying genetic operators for n number of generations, so as to get good quality and relevant clusters.

It is expected that fast and efficient clustering will be achieved by using proposed SVD based system.

REFERENCES

[1] Wei Song, Soon Cheol Park, "An Efficient Method of Genetic Algorithm for Text Clustering Based on Singular Value Decomposition", Division of Electronics and Information Engineering, Chonbuk National University, Korea,(7th International Conference on Computer and Information Technology,0-7695-2983-6/07), *DOI 10.1109/CIT.2007.197 ©2007*.

[2] Tony F. Chan, "An Improved Algorithm for Computing the Singular Value Decomposition", Yale University, (ACM,Trans. Math. Softw. 8, 1 (Mar.1982)), 84-88.

[3] Vittorio Castelli, Alexander Thomasian and Chung Sheng Li, "CSVD: Clustering and Singular Value Decomposition for Approximate Similarity Searches in High Dimensional Spaces", IBM Research Division and CS&E Dept. at Univ. of Connecticut, June 12, 2000.

[4] A. Civril, M. Magdon Ismail, "Column subset selection via sparse approximation of SVD", Meliksa University, Computer Engineering Department, Rensselaer Poly-technic Institute, Computer Science Department, 0304-3975/$ © 2011 Elsevier B.V.

[5] Mihee Lee, Haipeng Shen, Jianhua Z. Huang, J. S. Marron, "Biclustering via Sparse Singular Value Decomposition", Department of Statistics and Operations Research, University of North Carolina, Department of Statistics, Texas A&M University, *DOI: 10.1111/j.1541-0420.2010.01392.x, ©2010*, (The International Biometric Society).

[6] Emmeline P. Douglas "Clustering Datasets With Singular Value Decomposition", Department of Mathematics, The Graduate School of the College of Charleston, (UMI number: 1461189 Copyright 2009 by Douglas, Emmeline P).

[7] Sivasankaran RajaManickam,"Efficient Algorithms for Sparse Singular Value Decomposition",The Graduate School of University of Florida,©2009 Sivasankaran Rajamanickam.

[8] George W. Furnas & Scott Deerwester, "Information Retrieval Using a Singular Value Decomposition Model of Latent Semantic Structure",University of Bellcore & Chicago, ©1988 (ACM O-8979 A-274-8 88 0600 0465 $ 130).

[9] Wei Song, Cheng Hua Li, Soon Cheol Park, "Genetic Algorithm for Text Clustering using Ontology and Evaluating the Validity of various Semantic Similarity Measures ", Department of Electronics and Information Engineering, Chonbuk National University, Jeonju, Jeonbuk 561-756,Republic of Korea, (©2008 Elsevier Ltd. All rights reserved).

[10] Jing Gao, Jun Zhang, "Clustered SVD Strategies in Latent Semantic Indexing", Laboratory for High Performance Scientific Computing and Computer Simulation, Department of Computer Science, University of Kentucky, 773 Anderson Hall,Lexington,KY405060046,USA,*doi:10 .1016/j.pm.2004.10.005.* (©2004 Elsevier Ltd).

[11] Peg Howland, Moongu Jeon, Haesun Park, "Structure Preserving Dimension Reduction for Clustered Text data based on the Generalized SVD", SIAM J.MATRIX ANAL. APPL. *Vol. 25, No. 1, pp.165–179*, 2003 Society for Industrial and Applied Mathematics.

[12] Habiba Drias, Ilyes khennak, Anis Boukhedra, "A Hybrid Genetic Algorithm for Large Scale Iformation Retrieval", Department of Computer Science, Algeria, (978-1-4244-4738-1/09/$25.00 ©2009 IEEE).

[13] Jose L. Rebeiro Filho, Philip C., "Genetic Algorithm- Programming Environments", Treleaven University, London, UK, (0018-9162/94/$4.00 ©1994 IEEE).