Design and Development of Algorithm for Software Components Retrieval Using Clustering and Support Vector Machine

Amarjeet Kaur
Department of Computer Science and Engineering
Chandigarh University

Iqbaldeep Kaur
Associate Professor
Department of Computer Science and Engineering
Chandigarh University

Abstract- Component Based Software Development is important area in software development. In this paper, we describe various algorithms and techniques for efficiently retrieval of components from the component repository. We discuss XNOR similarity function, clustering algorithms like k-mean, K-medoid, K-mode and supervised leaning algorithm like support vector machine. This algorithm takes input as software components an output is a set of highly cohesive pattern group

Keywords– Clustering, Retrieval, K-mean, K-medoids, K-mode, Support Vector Machine.

I. INTRODUCTION

In the last decades, for retrieval of software components, various models come under existence. The first approach of information retrieval based on computer based technology came into existence in 1940s.Information retrieval system grew with the increase of speed of processor and storage capacity. On that time various processes, techniques were used for efficiently retrieval of information with in less time.

Evolution of information retrieval

1940: In 1940 Univac computer was described by Holmstrom. Univac computer was capable of searching for text references associated with a subject code. The code and text were stored in steel tape. Univac computer could process at the rate of 120 words per minute. [23]

1950: In 1950 information retrieval as a research discipline was starting to emerge with two important developments: how to index documents and how to retrieve them. At this time information was retrieve with the help of ranked retrieval method[23].

1960: In 1960 the information was retrieve with the help of ranked retrieval method. The documents and queries were viewed as vectors within an N dimensional space. At this time similarity between document and query vector was also measured.[23]

1970: One of the key developments of this period was that Luhn's term frequency (tf) weights (based on the occurrence of words within a document), were complemented with Spärck Jones's work on word occurrence across the documents of a collection. Her paper on inverse document frequency (idf) introduced the idea that the frequency of occurrence of a word in a document collection was inversely proportional to its significance in retrieval.[23]

1980-1990: At this time the advances on the basic vector space model were also developed and probably the most well-known is Latent Semantic Indexing (LSI), where the dimensionality of the vector space of a document collection was reduced though singular-value decomposition [23].

1990-1999: In 1990,Small numbers of websites were in existence. As time passing, in mid of 1993 the number of websites were increased and also increase the number of pages in websites. In 1999 at that the interaction between the commercial and research related information retrieval communities were strong. Computer users also increased from hundreds or thousands to many millions of international users.

2000 to present: With growth of people's needs, technologies must be upgraded according to the need of people .There are various technologies and methodologies which are used to retrieve the information from repository but there are some problems that different techniques does not efficiently retrieve the components according to user requirements. So the work is continue on the efficiently retrieval of information.

1.1 Information retrieval

Information Retrieval (IR) is the process by which a collection of data is represented, stored, and searched for the purpose of knowledge discovery as a response to a user request (query)[22].The main goal of information retrieval process is to retrieve the relevant documents in optimal time. IRSs usually implement following processes i) user enters the query ii) user query match with the collection of documents with the help of matching algorithm iii) ranking of relevant documents iv) retrieval results .

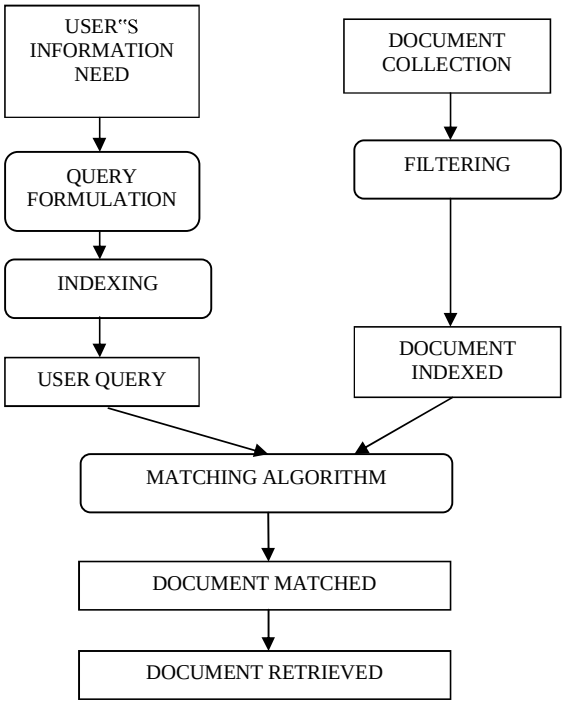

Fig 1 A general framework of IR System [22]

1.1.1 Information retrieval models

Information retrieval model has three fundamental models. These models specify the details of the document representation, the query representation and the retrieval functionality.

A. Boolean model

In Boolean model, it allows for the use of different operators of Boolean algebra like AND, OR, and NOT, for query formulation. The main disadvantage of Boolean model is that it does not provide ranking to the retrieve documents. [22]

B. Vector Space Model

The main feature of vector space model is that it provides ranking to the retrieve documents. In vector space model, number of documents and user query is represents as vector and the angle between the two vectors are calculated by using the similarity cosine function. [22] This model also introduce two terms tf (term frequency and idf(inverse term frequency).tf is frequency of occurrence of the terms in the document or query texts and idf is inverse of the

number of documents that contain a query or document term. According to Vector space space model i) preprocessing of documents and user query ii) transformation by using tf(term frequency) and idf (inverse term frequency) .

C. Probabilistic model

The main feature of the probabilistic model is that it provides ranking to the retrieve documents by their probability of relevance given a query [9]. Documents and queries are represented by binary vectors ~d and ~q, each vector element indicating whether a document attribute or term occurs in the document or query, or not. Instead of probabilities, the probabilistic model uses odds O(R), where O(R) = P(R)/1 − P(R), R means "document is relevant" and ¯R means "document is not relevant" [22].

1.2 Component Based Software Engineering

Component Based Software Engineering is the idea of reusability of software components. Component Based Software Engineering is widely used because in object oriented programming the objects are too complicated and does not allow plug play. Component Based Software Engineering allow plug and play and also provides various guidelines, methods and models for development of software. Basically in Component Based Software Engineering it is a collection of software components known as repository. It selects the software components from the repository and assembles them with other component of the software.

In Component Based Software Engineering, the component is also known as module of software, a piece of code, software documents etc. During software development various challenges exist like complexity, delivery time, budget etc. Component Based Software Engineering takes less time to implement software because user selects the components from the component repository.

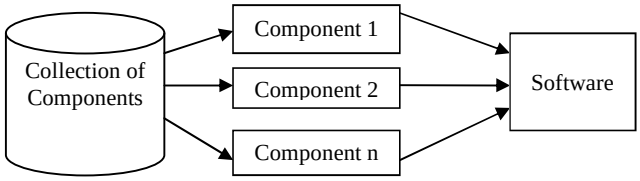

Fig2 Component Based Software Development [1]

1.4 Clustering

There are different methods for information retrieval. But they does not provide relevant documents in optimal time. Clustering is also information retrieval technique which provides better results as compare to other retrieval techniques.Clustering is mainly the process of making the group of similar type of objects. The benefit of grouping over categorization is that, it is flexible for modification as well as assists distinct features that illustrate dissimilar groups. Clustering can be supervised, semi supervised and unsupervised.

1.4.1 Clustering methods

A. Partitioning method

Assume a database of n objects is, the division method create k partition of data. Each partition will characterize a cluster and k=n. It means that it will categorize the information into k groups that assume the following necessities.

i) Every cluster has at least one object.
ii) Each object must belong to exactly one group.
iii) The number of divisions (say k,) the partitioning method will construct an primary partitioning.It uses the iterative rearrangement technique to advance the divisioning by moving objects from one cluster to other.

Example: - 1) K-medoid
2) K-means

B. Hierarchal Method
It constructs the hierarchical breakdown of given set of data objects. It categorizes this method on basis of how the hierarchal decomposition is created. It is top down approach.

Example: - 1) Agglomerative algorithm
2) Divisive algorithm

C. Density Based Method
it is based on the concept of compactness. The design is to persist upward the given cluster as the compactness in the region go above some threshold i.e. for each data within a given cluster, a minimum number of points has to be included by the radius of a given cluster

Example: - 1) DBSCAN
2) OPTICS

D. Grid Based Method
In this the objects together from a lattice. The object space is quantized into fixed number of cells that from a network arrangement. The key benefit of this method is rapid processing time. It is reliant only on the number of cells in each dimension
.
Example: - 1) STING
2) WaveCluster

E. Model Based Method
In this method a model is hypothesize for each cluster and find the best fit to the given model. It establishes the cluster by clustering the density function. This reveals spatial allocation of the data points. This method also provides a technique of repeatedly determine number of clusters based on standard information, taking noise into description and give up vigorous clustering techniques
.
Example: - 1) Conceptual Clustering
2) Expectation-Maximization

F. Constraint Based Method
This technique is carried out by merging of user or application oriented constraints. The constraint refers to the user prospect of preferred clustering consequences. The constraint provides the method of communication with the clustering method.

II. RELATED WORK

A research on Information retrieval is going on, brief literature review of work done in the field of retrieval of software components with help of supervised and unsupervised learning is discussed here:
Amit Kumar [21] The author has discussed that for retrieval of software components keyword technique is used and also proposed two algorithm for how to select the software components and how to insert the software components in the repository by using keyword technique.

Shweta Yadav et.al. [24] In this paper author discussed although many efforts have been done to the retrieval of software components most of the techniques does retrieve accurate component according to the user requirement .So it define algorithm which is combination of two most popular retrieval techniques i.e. based on keyword based approach and semantic component retrieval techniques.
C.V. Guru Rao et.al [27] Design and define an algorithm for clustering the documents. Authors have discussed a clustering of components on the basis of XNOR similarity function. The input to the clustering algorithm can be a set of software entities or software requirement documents or any set of software components. And the output is set of clusters. According to the algorithm i) preprocessing of documents ii) Find out the frequency of each word in each document iii) Cluster the documents on the basis of XNOR similarity function.

C.V. Guru Rao et.al [28] Design and define an algorithm for clustering the documents. Authors have discussed a clustering of components on the basis of XOR similarity function. The input to the clustering algorithm can be a set of software entities or software requirement documents or any set of software components. And the output is set of clusters. According to the algorithm i) preprocessing of documents ii) Find out the frequency of each word in each document iii) Cluster the documents on the basis of XOR similarity function.
C.V. Guru Rao et.al [25] Design and define an algorithm for clustering the documents. Authors have discussed a clustering of components on the basis of BASIC similarity function. The input to the clustering algorithm can be a set of software entities or software requirement documents or any set of software components. And the output is set of clusters. According to the algorithm i) preprocessing of documents ii) Find out the frequency of each word in each document iii) match each frequent word with the predefined terms on the basis of basic similarity function iv) no of clusters
C.V. Guru Rao et.al [26] In this paper author proposed a novel similarity measure by modifying the Gaussian function. The similarity measure designed is used to cluster the text documents.
.Xiaofei Zhoua [29] this paper explores that k-mean algorithm is used for text categorization. Authors have discussed for text categorization firstly doing tokenization, stop words, stemming of documents. Then every term of the document match with selected features and then apply the k-mean algorithm.
Mingyu Yao [18] this paper explores the idea to cluster the Chinese text using k mean algorithm. According to the algorithm in this paper firstly doing i) Preprocessing of text ii) Transformation of the text using Term frequency (TF) and Inverse document frequency (IDF) iii) Getting the weighted text apply the k-mean algorithm. K mean algorithm has three steps. i) Determine the centroid coordinate ii) Determine the distance of each object to the centroids iii) Group the object based on minimum distance (find the closest centroid).
K. A. Abdul Nazeer [12] Author has discussed about k mean and modified k-mean.it define K- mean algorithm cluster the documents but it does not provide accurate results because of randomly chosen centroid initially so according to modified k mean i)find out the initially centroid ii) apply k mean algorithm.
Shalini S Singh [16] Author has discussed the comparative analysis of k-mean and k-mediod algorithm. This paper concludes that both clustering methods are suitable for spherical shaped clusters in small to medium sized data sets. K-means and k-medoids – both the methods find out clusters from the given database. Both the methods require to specify *k*, no of desired clusters
Joshua Zhexue Huang et.al[11] Author has discussed k-means is the mostly used algorithm for clustering data because of its efficiency in clustering very large data. However, the standard k-means clustering process cannot be applied to categorical data due to the Euclidean distance function and use of means to represent cluster centers. So k-mode algorithm is used to cluster categorical data. Steps of k-mode algorithm i) Randomly select k unique objects as the initial cluster centers (modes)ii) Calculate the distances between each object and the cluster mode iii) assign the object to the cluster whose center has the shortest distance to the object iv) repeat this step until all objects are assigned to clusters v) Select a new mode for each cluster and compare it with the previous mode. vi) If different, go back to Step 2; otherwise, stop.
Zhexue Huang[5] Author has discussed the comparison between the k-mean algorithm and k-mode algorithm. This paper describe the process of k-mode clustering that is applied to categorical data.
Rishi Syal[19] Author has discussed k-mean, k-mode and modified k-mode algorithm. It also described the drawbacks of k-mean and k-mode algorithm that is overcome by modified k-mode clustering algorithm. According to modified k-mode algorithm i) Partitioning the Dataset into blocks ii) Sub Clustering each block using modified K-Mode Algorithm.
Osama Abu Abbas[10] Author has discussed the comparison of different clustering algorithms like k-mean, hierarchal algorithm, E-M algorithm ,SOM algorithm. The conclusion of this paper is i) all algorithm have some ambiguity in some data when clustered ii)k-mean and EM algorithm is less accurate as compare to other algorithms iii)k-mean and EM algorithms shows better results when using huge dataset iv) hierarchal and SOM algorithim shows better results when using small dataset.
Monika Arora [20] Author has discussed about support vector machine algorithm and also discuss SVM with RBF kernel, SVM with Linear kernel, SVM Polynomial kernel. The conclusion of this paper is that RBF kernel gives better results.
Thomas Finley [8] Author has discussed supervised clustering method SVMcluster based on an SVM framework for learning structured outputs. The algorithm accepts a series of "training clusters," a series of sets of items and clusterings over that set. The method learns a similarity measure between item pairs to cluster future sets of items in the same fashion as the training clusters.
Sebastián Maldonado [17] Author has discussed two algorithms i) kernel-penalized SVM (KP-SVM),it is an embedded method that simultaneously selects relevant features during classifier construction by penalizing each

feature's use in the dual formulation of support vector machines (SVM). This algorithm is used to optimize the shape of an anisotropic RBF Kernel eliminating features that have low relevance for the classifier. ii) KP-SVM employs an explicit stopping condition, avoiding the elimination of features that would negatively affect the classifier's performance.

Harris Drucker [6] In this paper author compared support vector machine to Rocchio, ide regular and ide dec-hi in information retrieval of text documents using relevancy feedback. This paper conclude that if inverse document frequency weighting is not used because one is unwilling to pay the time penalty needed to obtain these features, then SVM are better whether using term frequency(TF) or binary weighting.SVM performance is better than ide dec-hi if TF-IDF weighting is used.

III. ALGORITHM FOR INFORMATION RETRIEVAL PROCESS

A. Clustering Using XNOR Similarity Function

XNOR similarity function is vector based algorithm. This algorithm is used to cluster the software documents on the basis of XNOR similarity function. It takes input as software documents and gives set of cluster as output. According to the XNOR similarity function we find out the similarity between the two documents if the feature is not present in one of them then it will give Z(worst case) .and if the feature is not present in both documents then it will also provides the zero. Else the feature is present in both documents then it will give the value one.[28]

XNOR Similarity Function Algorithm

Begin
Initialize D , SW , STMW,F,W,i
(Each SW € STMW)
For (D=1,D>Ld , D++)
{
Remove stop words and stemming words from each document.
Find unique words (Uw) in each document and count of the same.
Find frequent itemsets (CUw)of each document Form a word set W consisting of each word in frequent item sets of each document.
Form Dependency Boolean Matrix with each row and
column corresponding to each Document and each word
respectively
For each document (R =1, R > Ld , R++)
{
{ For each itemsets(CUw=1,CUw>Ld,CUw++)
}
For each word in word set do
}
If (word wk in Word set W is in document Di)
{
Set D[Di, wk] = 1
Else
Set D[Di, wk] = 0
End if
End for
End for
}
Find the Feature vector (Fv) (belongs)similarity matrix by evaluating similarity value for each document pair applying
Apply XNOR.Hybrid XNOR Function defined in table 1 to obtain the matrix with feature vectors for each document pair.
Replace the corresponding cells of matrix by count of number of zeroes in tri state feature vector.
At each step, find the cell with maximum value and document pairs containing this value in the matrix. Perform Clustering (Uniform) Group

such document pairs to form clusters. Also if document pair (X,Y) is in one cluster and document pair (Y, Z) is in another cluster, form a new cluster containing (X, Y, Z) as its elements.
Repeat
until no documents exist or we reach the stage of first
minimum value leaving zero entry.
Output the set of clusters obtained.
Label the clusters by considering candidate entries.
End of algorithm

Fig3: Algorithm for Clustering

Table2: XNOR Similarity Measure

A	B	S(A,B)
0	**0**	**Z**
0	**1**	**1**
1	**0**	**1**
1	**1**	**0**

Advantages

i) To normalize the documents
ii) By using this algorithm the output is highly cohesive software documents

Disadvantage

i) XNOR similarity function sometimes find out the most frequent words which are not that much important as the word present in the document which is not frequent.
ii) It does not define how many clusters should be there.

B. K-mean Algorithm

The k- mean is of the simplest unsupervised leaning algorithms that solve the well-known clustering problem. The procedure follows a simple and easy way to classify a given data set through a certain number of clusters (assume k clusters) fixed a priori. The main idea is to define k centroids, one for each cluster. These centroids should be placed in a wiser way since different location tends to produce different results. So, the better choice is to place them as much as possible far away from each other. The next step is to take each point belonging to a given data set and associate it to the nearest centroid. When no point is pending, the first step is completed and an early grouping is done, At this point, it is needed to recalculate k new centroids as bar centers of the clusters resulting from the previous step. After these k new centroids, a new building has to be done between the same dataset points and the nearest new centroid. A loop has been generated, because of this loop it may notice that the k centroids change their location step by step until no more changes are done.

K-means Algorithm

```
var w = initialCentroids(t, K);
   var N = t.length;
   while (!stoppingCriteria)
 {
      var w = [ ][ ];
       for (var n = 1; n <= N; n++) {
          v = arg min (v0) dist(w[v0], t[n]);
          w[v].push(n);
   }
```

Fig4. K mean algorithm

Advantages

i) It gives better results as compare to hierarchal Clustering method when the number of variables is huge and k is small.
ii) It is easy to implement.
iii) It is easy to understand.

Disadvantage

i) K mean clustering algorithm does not handle data sets with categorical attributes.[12]
ii) It does not take guarantee of unique clustering because we choose random centroid initially
iii) The prediction of K is a difficult task
iv) It does not automate the process.

C. K-medoid Algorithm

It is also known as Partitioning Around Medoid, K-medoid algorithm is similar to the k-mean. In k-mean it uses centroid to represent the cluster and it is sensitive to outliers. This means, a data object with an extremely large value may disrupt the distribution of data. K-medoids method overcomes this problem by using medoids to represent the cluster rather than centroid. A medoid is the most centrally located data object in a cluster.[16] So according to k mediod algorithm.

K-medoid Algorithm

Input: X = {x1,x2,...xn} (dataset to be clustered or matrix of dissimilarity
1. k (number of clusters)
2. metric (kind of metric to use on dissimilarity matrix)
3. diss (flag indicating that X is the matrix of dissimilarity or not)
Output: W = {w1,w2,...,wk} (vector of clusters medoids)
L = {l(x) | x = 1,2,...,n} (set of cluster labels of X)
foreach mi € W do
mi ← xj € X; (e.g. random selection)
end if diss ≠ true
Dissimilarity ← CalculateDissimilarityMatrix(X, metric);
elseDissimilarity ← X;
end repeat
foreach xi € X **do**
l(xi) ← argminDissimilarity(xi, Dissimilarity, W);
end
changed ← false;
foreach mi € W **do**
Mtmp ← SelectBestClusterMedoids(X, Dissimilarity, L);
end
if Wtmp ≠ W
W ← Wtmp;
changed ← true;
end

Fig5. K medoid algorithm

Advantages

K-medoids clustering algorithm is more robust as compare to k-mean in the presence of outlier.

Disadvantage

i) Need to specify k, the total number of clusters in advance.
ii) K-medoids clustering does not automate the process.
iii) K-medoids works efficiently for small data sets but does not scale well for large data sets.[16]

D. K-mode Algorithm

The k-mode clustering technique is the improved standard of k- mean process. K-mean does not handle data sets with categorical attributes. So to handle categorical data k-mode clustering is used. The procedure of k –mode is similar to the k-mean.but in k-mode the mean is replaced with mode. So according to k mode algorithm [5]

i) Select randomly *k* objects as the initial cluster centers (modes).
ii) Calculate the distances between each object and the cluster mode; assign the object to the cluster whose center has the shortest distance to the object
iii) Repeat this step until all objects are assigned to clusters. Select a new mode for each cluster and compare it with the previous mode. If different, go back to Step 2; otherwise, stop

K-mode algorithim

k-modes(k,W)
 	
Choose k data points as the initial modes (cluster centers)
 	
repeat
 	
for each point **m** € W do
 	
compute the distance from **m** to each mode;
 	
assign **m** to the closest mode
 	
endfor
 	
re-compute the mode using the current cluster memberships
 	
until the stopping criterion is met

Fig6. K mode algorithm

Advantages

i) it is used to handle data sets with categorical attributes.
ii) It is used to more classify data as compare to k mean. Another advantage of the k-modes algorithm is that the modes give characteristic descriptions of clusters. These descriptions are very important to the user in interpreting clustering results.[19]

.

Disadvantage

i) K-mode does not take guarantee of unique clustering because we choose random centroid initially.
ii) k-mode clustering does not automate the process.

D. Support Vector Machine

Support Vector Machine is also known as Support Vector Networks. SVM is supervised leaning algorithms that solve the well-known clustering problem. The procedure follows a simple and easy way to classify the documents. It is used for boundary data analysis. This technique is used to achieve the higher accuracy in the process of retrieval. According to diagram the goal of SVM modeling is to find the optimal hyperplane that separates clusters of vector in such a way that cases with one category of the target variable are on one side of the plane and cases with the other category are on the other size of the plane. The vectors near the hyperplane are the support vectors.

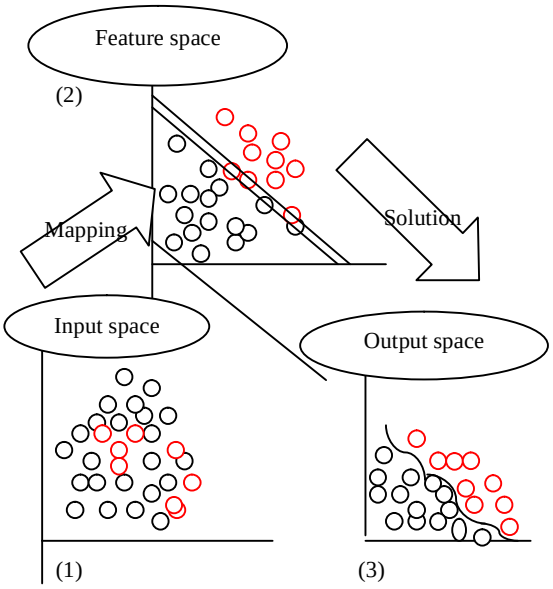

Fig7: Support Vector Machine Algorithm

Working of support vector machine:-

i) Firstly separating data/document into training from the testing tests.[20]
ii) Each instance or records of the training set contains a target value and several attributes (feature or observed variable). [20]
iii) The target of SVM is helping to produce a model or goal, which is based on training data and also predicts the target value of the test data.
iv) This training set enables the dataset the label pairs and classify it into the group for categorization.[20]
v) The support vector machine (SVM) is applied to achieve the optimized solution for the training set of data.
vi) The training vectors are mapped into a higher dimensional space by the use of function.[20]

3.5.1 Support Vector Machine using Linear kernel

SVM uses a linear separating hyper plane. This function uses the maximal margin in this higher dimensional space. There is the penalty parameter which uses for the error term. Because of using the kernel values, it usually depends on the product of the inner feature vectors. It is called the kernel function. According to linear kernel ,the training data is said to be linearly separable, when data can be separated at two hyper planes of the margins in a way that there are no points between them and then try to maximize their distance. This is the simplest kernel and shows good performance for linearly separable data.[20]

3.5.2 Support Vector Machine using Polynomial kernel

the polynomial kernel is a non linear kernel, used for large set of attributes values and polynomial kernels ,where the kernel values may go to infinity. That are linearly dependent on n dimensions, the kernel function of order n can be used to transform them into linearly independent vectors on those n dimensions. Once they are transformed into the dimension space, they become linearly separable.[20]

3.5.3Support Vector Machine using RBF kernel

The RBF radial basis function is most popular in choosing of kernel types in Support Vector Machines (SVM). This is mainly because it is localized and has finite responses across the entire range of the real x-axis. This kernel is basically suited best to deal with data which have a class-conditional probability distribution function approaching the Gaussian distribution. It maps such data into a different space where the data becomes linearly separable.[20]

3.5.4Support Vector Machine using RBF kernel

Sigmoid kernel is not as efficient for classification as are the other three. Indeed, one of the fundamental requirements on a valid kernel is that it must satisfy Mercer's theorem, and that requires that the kernel be positive definite. In cases where the kernel is not positive definite, the results may be drastically wrong, so much so that the SVM performs worse than chance. [20]

Advantages

i) Support vector machine handling the cases where the documents are not completely clustered.
ii) It is used to make the process robust.
iii) SVM is helps to automate the process.

Disadvantages

The main drawback of support vector machine is if the number of features is much greater than the number of samples, the method is likely to give poor performances.[20]

IV. CHALLENGES

In software component retrieval, component based development is a challenge in itself. Another big issue is component identification, means to select accurate component in optimal time and there is also a need to make the process automatic and robust. So, there must be need of an algorithm that helps in achieve the accuracy and efficiently retrieval of software components, better than already discussed algorithms.

V. CONCLUSION

This paper provides the brief review of software component retrieval and to make the process better, automated and less complex, it needs to be integrating with support vector machine. We also evaluate some of algorithms like XNOR, K-mean, K-medoids, K-mode, and also explain the support vector machine with different kernels like, RBF kernel, Polynomial kernel and linear kernel algorithms to check the optimization and accuracy of retrieval of software components. Support vector machine using RBF kernel solves the problem of clustering and provides better results as compare to unsupervised clustering algorithms.

VI. ACKNOWLEDGEMENT

Author is especially thankful to Dr. Amit Verma , Associate Professor & Head, Department of Computer Science & Engineering, Chandigarh university, Gharuan, Mohali, India for his valuable suggestions about this topic.

REFERENCES

[1] Royce, W. 1970. Managing the development of large software systems. IEEE.
[2] Belkin, J. N., Croft B. W. 1992. Information filtering and information retrieval: Two sides of the same coin? .ACM.
[3] Ibba, R.1993. Structure-based Clustering of Components for Software Reuse. IEEE.
[4] Szyperski, C. 1996.Components vs. Objects vs. Component Objects. WCOP'96 workshop report. Special issue in object-oriented programming, pp 127-130.
[5] Huang, Z. 1998.Extensions to the *k*-Means Algorithm for Clustering Large Data Sets with Categorical Values. ACT, 283–304.
[6] Drucker,H.2001.Support Vector Machine: relevance feedback and information retrieval. Elsevier.
[7] Rodriguez, I.2003.A Framework for Selecting Components Automatically. Elsevier Science, vol.
[8] Finley, T., Joachims, T.2005.Supervised Clustering with Support Vector Machines. International Conference on Machine Learning, Bonn, Germany.
[9] Khan, S.S.2007. Computation of Initial Modes for K-modes Clustering Algorithm using Evidence Accumulation. IJCAI.
[10] Abbas, A.O.2008.Comparisons between data clustering algorithms. The International Arab Journal of information Technology Applications, Volume 5 –No.
[11] Huang, Z.J.2009.Clustering Categorical Data with k-Modes. IJSCE.
[12] Nazeer, A.A.K, Sebastian, P.M.2009.Improving the Accuracy and Efficiency of the k-means Clustering Algorithm.WCE.
[13] Chen, X., Yin, W., Tu2, P.2009. Weighted *k*-Means Algorithm Based Text Clustering. IEEE.

[14] Kaur, A., Singh, K.2010.Component Selection for Component based Software Engineering. International Journal of Computer Applications, Volume 2 –No.1.
[15] Binkley,D.2011.Information Retrieval Applications in Software Development. Encyclopedia of Software Engineering.
[16] Singh,S.S., N C Chauhan,C.N.2011.K- Means v/s K-Medoids: A Comparative Study National Conference on recent trends in Engineering and technology.
[17] Maldonado,S.,Weber,R.,Basak,J.2011.Simultaneously feature selection and classification using kernel penalized support vector machine Elsevier.
[18] Yao,M.2012.Chinese text clustering algorithm based k-means. Physics Procedia 33 301 – 307.
[19] Syal,R. 2012.Innovative Modified K-Mode Clustering Algorithm IJERA Volume 2, Issue 4 pp. 390.
[20] Arora, M.2012, Efficient and Intelligent Information Retrieval using Support Vector machine.IJSCE Volume-1, Issue-6.
[21] Kumar, A.2013.Software Reuse Libraries Based Proposed Classification for Efficient Retrieval of Components. IJARCSSE Volume 3, Issue 6.
[22] Sharma, M.Patel, R.2013. A Survey on Information Retrieval Models, Techniques and Applications.IJETAE Volume 3, Issue 11,
[23] Sanderson,M., Croft,W.B.2013.The History of Information Retrieval Research.
[24] Yadav,S.,Kaur ,K.2013.Design of Rank Based Reusable Component Retrieval Algorithm. IJARCSSE, Volume 3, Issue 11.
[25] Srinivasa,C., Radhakrishna,V.2014. Clustering and Classification of Software Component for Efficient Component Retrieval and Building Component Reuse Libraries . Elsevier.
[26] Srinivasa,C., Radhakrishna,V.2014. A Modified Gaussian Similarity Measure for Clustering Software Components and Documents. Proceedings of the ACM SIGMOD Conference on Management of data.
[27] Srinivasa,C., Radhakrishna,V.2014. Clustering software components for program restructuring and component reuse using hybrid XOR similarity function.Procedia Technology.
[28] Srinivasa,C., Radhakrishna,V.2014. Clustering software components for program restructuring and component reuse using hybrid XNOR similarity function. Procedia Technology.
[29] Zhoua,X.2014.Text Categorization Based on Clustering Feature Selection. Procedia Computer Science 31(2014) 398 – 405 .