Survey of Algorithm for Software Components Reusability Using Clustering and Neural Network

Indu Verma
Department of Computer Science and Engineering
Chandigarh University

Iqbaldeep kaur
Associate Professor
Department of Computer Science and Engineering
ChandigarhUniversity

Abstract- Information Retrieval Systems are the systems that are combination of the tools and techniques which provide relevant information to the user regarding his query which is useful to the user. Information system's goal is to provide information that matches to his input query. Information system may work on the ranking system that is user can give ranks to the information regarding particular query so that next time top ranked information will be provided first.

Since, we have many ways for information retrieval such as keyword, semantic, rank based search and many more. However, the use of clustering when combined with methods of information retrieval has proved to give promising results. So, we have concluded to work in this field of clustering information using learning algorithm "Survey of Algorithm for Software Components Reusability Using Clustering and Neural Network".

With the development of the component technology and the expansion of component library, representing and retrieving components has become a major issue to reuse the components. Many papers have been published providing various techniques but few of the paper have given a systematic review of these techniques. This study discusses the current state of the reusability techniques for the retrieval of components. The result of the review is classified on the type of approach and the type of approaches to validate the approaches. This paper we focus on Clustering Algorithm because this is unsupervised machine learning method.

Keywords – Clustering, Retrieval, K-mean, K-mode, Neural Network.

I. INTRODUCTION

Information is the meaningful data. After processing data we get information which informs, i.e. from data we get information and knowledge. Information is conveyed either as message or the content of a message or through direct or indirect observation of something. Information can be encoded into various forms for transmission and interpretation. For example, information may be encoded into signs, text, images, signals etc. The process of retrieving or getting something back from somewhere is called retrieval. The process of accessing information from memory and the act of getting back information is also called retrieval.

Retrieval could refer to in computer science, Information retrieval, Data retrieval, Knowledge retrieval, Text retrieval.

Information retrieval system (Fig1) is the system of obtaining information from Meta data according to user need. Information retrieval is the activity of obtaining information resources relevant to an information need from a collection of information resources.

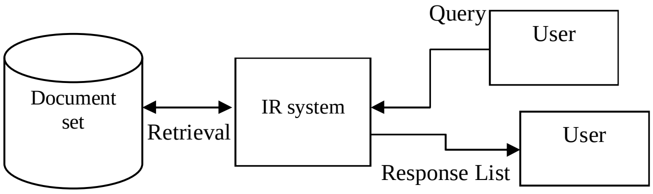

Fig1. Information Retrieval System

A. Evolution of Information Retrieval

Prior to the broad public day-to-day use of search engines, IR systems were found in commercial and intelligence applications as early as the 1960s. The earliest computer-based searching systems were built in the late 1940s and were inspired by pioneering innovation in the first half of the 20th century [16]. As with many computer technologies, the capabilities of retrieval systems grew with increases in processor speed and storage capacity. The development of such systems also reflects a rapid progression away from manual library-based approaches of acquiring, indexing, and searching information to increasingly automated methods.

B. Software Engineering, CBSE and Information Retrieval

In view of software engineering, all the components within the same cluster have high cohesion and low coupling. Software component clusters can be treated as highly cohesive groups with low coupling which is the desired feature. Clustering is not any one specific algorithm that we can stick firm to, but it must be viewed as the general task to be solved. Clustering algorithms may unsupervised or supervised. In unsupervised clustering the partitions are viewed as the unlabelled patterns or components. Supervised clustering algorithms label the patterns which can be used to classify the components for decision making. Hence the partitions obtained by clustering process may be labeled or unlabeled. Document clustering or text clustering is one of the main themes in text mining. It refers to the process of grouping documents with similar contents or topics into clusters to improve both availability and reliability of text mining applications such as information retrieval, classifying text, summarizing document sets, etc.

C. Classification of Information Retrieval

Classification is the process of analyzing a particular input and assigning it to category.

a) *Classification approaches:*

Table-1 Approaches

Manual Classification	**Rule-based Classification**	**Statistical Classification**
Quality is high.	**It can encode complex decision strategies.**	**It is robust and adaptive.**
It is expensive and slow.	**It needs domain Expertise.**	**It need training**

Numerous Supervised learning algorithms for information retrieval are present. Some of them are:

- ✓ Naïve Bayes
- ✓ Decision Trees
- ✓ Neural Network
- ✓ K-nearest neighbor
- ✓ Support vector machine

Classification is a crucial part of information retrieval systems.

Supervised learning is used to automatically estimate classifiers from data:

- ✓ Based on training set.
- ✓ Evaluated on separate test set.

- ✓ Input represented as feature vectors.
- ✓ Example : Neural Network

D. Clustering

Clustering is a technique for searching dense subgroup, subsets from multidimensional or Meta data. Clustering in information retrieval has been used many types of purposes like file grouping, indexing and many more. Documents in same cluster behave similarly with respect to relevance to information need. In one cluster or group there should be same type of documents, as documents in one cluster share many things like their properties, attributes, features. With clustering effectiveness of information retrieval is improved. Clustering is unsupervised learning technique. Various types of clustering algorithm are:

i. Connectivity based clustering (Hierarchical Algorithms)
  - Agglomerative algorithm
  - Divisive algorithm

ii. Centroid based clustering(Partitioning algorithms)
  - K-mediods
  - K-means

iii. Distribution bsed clustering

iv. Density Based clustering

II. Related Work

In this paper the author Chintakindi Srinivasa [1] discussed the idea to cluster the software components. We define a similarity function and use the same for the process of clustering the software components to find degree of similarity between two document sets or software components. A similarity matrix is obtained for a given set of documents or components by applying similarity function.

In this paper the author Vangipuram Radhakrishna [2] discussed a generalized approach for clustering a given set of documents or software components by defining a similarity function called hybrid XOR function to find degree of similarity between two document sets or software components. A similarity matrix is obtained for a given set of documents or components by applying hybrid XOR function. We define and design the algorithm for component or document clustering which has the input as similarity matrix and output being set of clusters. The output is a set of highly cohesive pattern groups or components.

Tthe author Chintakindi Srinivas of the paper [3] proposed a novel similarity measure by modifying the Gaussian function. The similarity measure designed is used to cluster the text documents and may be extended to cluster software components and program codes. The similarity measure is efficient as it covers the two sides of the term-axes.

The author Shalini S Singh [12] compared k means and k mediod unsupervised learning techniques. Both the techniques depend upon initial selection of points. K mediod is somebut better technique than k means as it is not sensitive to noisy data but have high computational cost.

In this paper the author Kavitha Karun A [7] discussed clustering algorithm with their limitations and their solutions. A comparison between k mean variants and their features is discussed.

In this paper the author Khan S Sharoz, et.al [21] presented an approach to compute the initial modes for K-modes clustering algorithm for clustering categorical data using Evidence Accumulation. The procedure is motivated by the observation that some data objects do not change their class membership even when subjected to different random initial conditions (modes). We utilized the idea of Evidence Accumulation for combining the results of multiple K-mode clusterings. The resultant modes of each of these runs were stored in a Mode-Pool. The most diverse set of modes were extracted from the Mode Pool as the initial modes for the K-mode algorithm. The computed modes were majorly being representative of those patterns that are less susceptible to random selection of initial modes.

In this paper the author K.-L. Du [13] provided survey and introduction to neural network based clustering. Various aspects are discussed and their examples are explained.

III. ALGORITHM FOR INFORMATION RETRIEVAL PROCESS

A. Clustering Using Similarity Function

A method for information retrieval is to cluster the document containing information. A similarity function is defined in table 3 and that similarity function is used for the clustering of documents. The degree of similarity between two or more documents is found out using similarity function [1]. A similarity matrix is obtained by applying similarity function. Clusters are formed but not fixed. As shown in figure 3 below this can be used for unsupervised technique.

```
Algorithm_Clustering
Begin
Initialize Variable
Stemming word (Sw)
Stop word (Ws)
Document (D)
Target Document (Dn)
For ($D_i$=1, $D_i$<= Dn, $D_i$++)
{
Remove Ws&Sw
Where (Ws € Di and i = 1……n)
For each $D_i$
{
Count Uw  ( Where Uw€ Di)
For each  Uw
{
Create Drset
(Drset€ SVD)
}
}
}
End
For (Wk=1, Wk<=Dn, Wk++)
{
Create DBM
(DBM→Dependency Boolean Matrix)
}
For
{
($D_i$=1, $D_i$<= Dn, $D_i$++ )
CreateDBM (Dependency Boolean Matrix
(Where $D_r$and $D_c$ € DBM)
IfDi=1,
Set D [Di, Wk] =1
Else
Set D [Di, Wk] =0
End
}

Find the feature vector similarity matrix by evaluated similarity value for each component pair applying similarity measure defined in table 3 to obtain the matrix with feature vectors for each component or document pair.

Replace the corresponding cells of matrix by count of number of zeroes in tri state feature vector.
Step 6: At each step, find the cell with maximum value and document pairs containing this value in the matrix.
```

```
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.
Step 7: Repeat Step6 until no components or documents exist or we reach the stage of first minimum value leaving zero entry.
Step 8: Output the set of clusters obtained.
Step 9: Label the clusters by considering candidate entries.
End of algorithm
```

Fig2: Algorithm for Clustering

Table2: Similarity Measure

E1	E2	Sim(E1,E2)
Absent	Absent	Neglect
Absent	Present	0
Present	Absent	0
Present	Present	1

B. K-mean Algorithm

It is a form of unsupervised learning algorithm for solving clustering problem in any field. It follows simple procedure by fixing the number of cluster priory to make it easy to classify given data set. Mainly there should be k centers defined, one for each cluster. These centers should be placed manipulative way so that they produce same result as different location causes different result. So, better choice to place these centers far away from each other. Next step is to take each point belonging to a given data set and associate it to the nearest center. First step will complete when there is no point pending. Now re-calculate k new centroids as barycenter of the clusters resulting from previous step. After this binding has to be done between the same data set points and the nearest new center. The value of k keeps on changing until convergence in results occurs. The Objective Function is:

$$J(V) = \sum_{i=1}^{c} \sum_{j=1}^{ci} (\|xi - vj\|)2$$

where,

'$||x_i - v_j||$' is the Euclidean distance between x_i and $v_{j.}$

'c_i' is the number of data points in i^{th} cluster.

'c' is the number of cluster centers.

The algorithm discussed below [7]

Algorithm_K-means

Let X = {$x_1,x_2,x_3,\ldots\ldots..,x_n$} be the set of data points and V = {$v_1,v_2,\ldots\ldots.,v_c$} be the set of centers.

1) Randomly select *‘c’* cluster centers.

2) Calculate the distance between each data point and cluster centers.

3) Assign the data point to the cluster center whose distance from the cluster center is minimum of all the cluster centers..

4) Recalculate the new cluster center using:

$$vi = (\frac{1}{ci})\sum_{j=i}^{ci} xi$$

where, *‘c_i’* represents the number of data points in i^{th} cluster.

5) Recalculate the distance between each data point and new obtained cluster centers.

6) If no data point was reassigned then stop, otherwise repeat from step 3).

Fig3. K mean algorithm

Strengths:

- Easy to understand, fast and robust.
- When data set are distinct gives best result.

Weaknesses:

- Value of k should be predefined.
- It can be really slow since in each step the distance between each point to each cluster has to be calculated, which can be really expensive in the presence of a large dataset.
- This method is really sensitive to the provided initial clusters, so this problem has been addressed with some degree of success using k mode method.

C. K-mode Algorithm

The drawback of K-means clustering algorithm is that it cannot cluster categorical data because of the dissimilarity measure it uses. The K-modes algorithm for clustering is based on K-means paradigm but removes the numeric data limitation whilst preserving its efficiency [7]. The extension of K-mean algorithm is K-modes algorithm to cluster categorical data by removing the limitation imposed by K-means through following modifications:

- It uses a simple matching dissimilarity measure or the hamming distance for categorical data objects.
- It replaces means of clusters by their modes.

The dissimilarity measure between two categorical objects X and Y described by m categorical attributes can be defined by the total mismatches of the corresponding attribute categories of the two objects. The objects are said to be more similar if the number of mismatches is less. This measure is often referred to as simple matching [19].

$$d1(X,Y) = \sum_{k=1}^{m} \delta(xj, yj) \qquad (1)$$

Where

$$\delta\,(xj\,,\,yj) = \begin{cases} 0 & ;\, xj = yj \\ 1 & ;xj \neq yj \end{cases} \qquad (2)$$

Pseudo code for k-modes algorithm is listed as follows [18].

```
Algorithm_K-modes
1. Select k initial modes, one for each cluster.

2. Allocate an object to the cluster whose mode is the nearest to it according to (1). Update the mode of the cluster after each allocation according to Theorem 1[20].
3. After all objects have been allocated to clusters, retest the dissimilarity of objects against the current modes. If an object is found such that its nearest mode belongs to another cluster rather than its current one, reallocate the object to that cluster and update the modes of both clusters.
4. Repeat 3 until no object has changed clusters after a full cycle test of the whole data set.
```

Fig4. K mode algorithm

Strengths:

- It is good for categorical data.

Weaknesses:

- Results attained by this algorithm totally depend on the choice of initial random cluster center, which can produce non-repeatable clustering results. Hence there will be creation of improper clusters.
- And to make the above the above process automated neural network will be helpful.

Above discussed partitioning algorithm k-means, k-mode could not help in making retrieval process automated and back propagated.

D. Neural Network

A type of artificial intelligence that tries to simulate the way a human brain works. A neural network works by creating connections between processing elements (neurons), rather than using a digital model, in which all computations manipulate zeros and ones. The organization and weights of the connections determine the output. Some benefits of neural network:

- Adaptive learning: An ability to acquire skill to perform different activities based on the data provided for training.
- Self-Organisation: An ANN can create its own functional body that retains the information acquired during learning time.
- Real Time Operation: Data processing may be carried out in parallel by ANN, and various hardware devices are being designed and manufactured specially to exploit this capability in the best manner possible. Types of neural network:

  - ✓ Feedback ANN
  - ✓ Feed Forward ANN
  - ✓ Classification-Prediction ANN

a) ***Feedback ANN*** -In this type of ANN, the output is fed back into the network to improvise the input to achieve the best possible results internally. The feedback network feeds information back into itself and is pertinent in solving optimization problems. These are mainly used by the internal system error corrections.

b) ***Feed Forward ANN*** - A feed-forward network is a simple neural network consisting of an input layer, an output layer and one or more layers of neurons. The power of the network can be noticed based on group behavior of the connected neurons and the output is decided through continuous assessment of its output by scrutinizing. The virtue of this network is that it becomes proficient in estimating and analyzing input patterns.

c) ***Classification-Prediction ANN*** –It is the subset of feed-forward ANN and the classification-prediction ANN is applied to retrieve data from Meta data. The network is trained to identify particular patterns and classify them into specific groups and then further classify them into "novel patterns" which are new to the network.

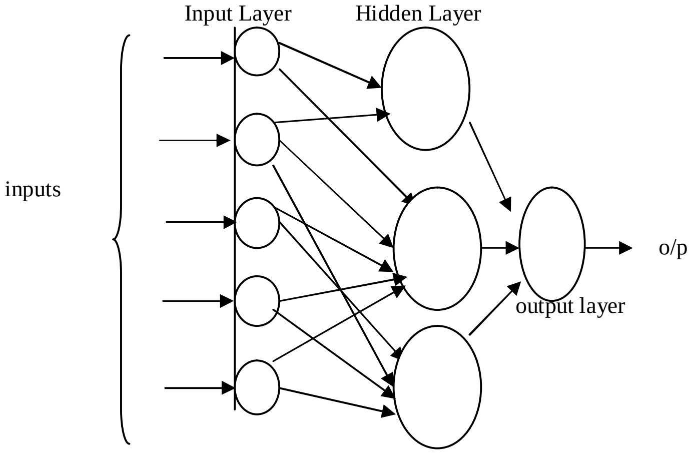

Fig5. Neural Network [13]

Conceptually Back Propagation Algorithm works as discussed below:

```
Algorithm_Back Propagation Neural Network

Input: ProblemSize, InputPatterns,iterations$_{max}$ , learn$_{rate}$
Network ← ConstructNetworkLayers()

Network$_{weight}$ ← InitializeWeights(Network, ProblemSize)

For (i=1 To iterations$_{max}$)

  Pattern$_i$ ← SelectInputPattern(InputPatterns)

  Output ← ForwardPropagate(pattern$_i$ , Network)

  BackwardPropagateError(Pattern$_i$,Output$_i$ , Network)

  UpdateWeights(Pattern$_i$, Output$_i$ , Network,learn$_{rate}$ )

End

Return (Network)
```

Fig6. Algorithm for Back Propagation

IV. CHALLENGES

The most eminent challenge faced in the retrieval process is the correct formulation of the query according to any basic retrieval technique used to represent the assets. As only an appropriate query will result into a correct match and the relevance of the asset fetched would also be high. K –mean and K – mode algorithms may not provide proper results in clustering. Neural network makes retrieval process efficient and effective due to its learning criteria.

V. CONCLUSION

This paper aims at throwing a light on Clustering Method for the components based on facets description. These retrieval methods for large scale component library for retrieval will on one hand meet various needs to retrieve; on the other hand, these can ensure the efficiency of the retrieval. In order to further improve the retrieval algorithm precision and recall rates, and to extend the term dictionary; how to expand retrieval condition in case of non-modify the program; How to design good retrieval interface and efficient retrieval platform for component reuse are next steps of retrieving the key issues.

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] ChintakindiSrinivasa "Clustering and Classification of Software Component for Efficient Component Retrieval and Building Component Reuse Libraries", Elsevier, 2014.

[2] ChintakindiSrinivasa "Clustering software components for program restructuring and component reuse using hybrid XOR similarity function", Procedia Technology, 2014.

[3] ChintakindiSrinivas, "A Modified Gaussian Similarity Measure for Clustering Software Components and Documents", Proceedings of the ACM SIGMOD Conference on Management of data, 2014.

[4] Bhatia V,Dhaliwal D, "Multiple Search Clssification of Repository," International Journal of Advanced Research Computer Science and Software Engineering, vol. 4, 3 March 2014.

[5] Singh A, " Development of Component Repository," International Journal of Advanced Research in Computer Science and Software Engineering, vol. 4, Issue 4, April 2014.

[6] Gupta S, "Reusable Software Component Retrieval System," International Journal of Application or Innovation in Engineering and Management, vol. 2, January 2013.

[7] K Karun Kavitha, Isaac Elizabeth "Cogitative Analysis on K-Means Clustering Algorithm and its Variants", International Journal of Advanced Research in Computer and Communication Engineering, Vol. 2, Issue 4, April 2013.

[8] Tomar P, "New Algorithm for Component Selection to Develop Component-Based Software with X Model," International Journal of Advanced Research in Computer Science and Software Engineering, vol. 1, 8 August 2013.

[9] Yadav S, "Design of Rank Based Reusable Component Retrieval Algorithm, "International Journal of Advanced Research in Computer Science and Software Engineering, vol. 3, 11,November 2013.

[10] Maqbool.O and Babri. H.A.," The Weighted Combined Algorithm: A Linkage Algorithm for Software Clustering" Computer Science Department,Lahore University of Management Sciences, DHA Lahore, Pakistan

[11] David Binkley, "Information Retrieval Applications in Software Development", Encyclopedia of Software Engineering,2011.

[12] Shalini S Singh, N C Chauhan "K- Means v/s K-Mediods: A Comparative Study", National Conference on recent trends in Engineeirng and technology, May 2011.

[13] K.-L. Du, "Clustering: A neural network approach", Elsevier, 2010.

[14] Kaur I, "Analytical Study of Component Based Software Engineering," World Academy of Science Engineering and Technology, vol. 3, 27 February 2009.

[15] Guru Rao C.V, "An Integrated Classification Scheme for Efficient Retrieval of Components," Journal of Computer Science, vol. 4, 2008.

[16] Dave M, Joshi R, Bhatia R "Retrieval of Most Relevant Component Using Genetic Algoritms," Research Gate, vol. 1, 26 June 2006.

[17] Rodriguez I, "A Framework for Selecting Components Automatically," Elsevier Science, vol.2003.

[18] Crnkovic Y, "Component Based Software Engineering-New Challenges in Software Development," Journal of Computing and Information Technology, vol. 4, 2003.

[19] Z. Huang, "Extensions to the k-means Algorithm for Clustering Large Data Sets with Categorical Values", Data Mining and Knowledge Discovery, 1998 – Springer.

[20] C. Szyperski, "Components vs. Objects vs. Component Objects," WCOP'96 workshop report. Special issue in object-oriented programming, pp 127-130, 1996.

[21] Khan Shehroz S, Dr. Kant Shri "Computation of Initial Modes for K-modes Clustering Algorithm using Evidence Accumulation", IJCAI.