Abstract— A class diagram models the static view of a system. The class diagrams are widely used during construction of executable code for software application as it is the only UML diagram which can be directly mapped with object oriented language. As class diagram contains duplicacy, the redundant source code generated increases the complexity of the program code. A solution is required to remove the redundancy from the class diagram, so that it model into efficient program code. To detect the similarity from the class diagram, it is converted into XML document. The clustering method is applied on the XML document. The clustering method applies on XML document group together the similar data. This method works on syntax and semantic of elements in the documents. The semantic clustering considers the position of elements. The similar elements are removed and a new redundant free class diagram is obtained. Now this class diagram models the object oriented language to get efficient program code.

Keywords— Data Guide, Data Modeling, Redundancy, Syntactic and Semantic Clustering, UML Class Diagram, XML Mining

I. INTRODUCTION

Class Diagram is the main aspects of modeling with programming language. The class diagram is used as source code for programming languages. The class diagram contains redundant elements then the code also contains the duplicacy which increases the complexity of application. To remove the similarities, class diagram is converted into XML document. The XML documents are used in large number of applications. The XML documents are represent by using the tree data structure which is complex structure. To simplify this complex structure, the mining task has been done. The various mining tasks [1] are

1. Classification
2. Clustering
3. Discovering of Frequent Tree Structure

The various challenges for document mining are XML document which is also known as syntactic clustering. Thus an algorithm is proposed in this topic for semantic clustering. The some important aspects used in this topic are briefly introduced here.

A. UML concepts

The UML stands for Unified Modeling Language. The UML is a language for visualizing, specifying, constructing, documenting the artifacts of software intensive system [2]. The UML is a popular object-oriented modeling language that can apply to application domains and implementation platforms as objects and methods. UML is effective for modeling large, complex software systems but it is simple to learn. UML is a language which is not only used for drawing diagrams but a complete language for capturing and express the knowledge about the subjects for the purpose of communication. The graphical notation is used to communicate in UML which is easy to understand. The UML is an expressive language that designs all the views which user wants to develop and create the system. The UML is independent of language and technology. The UML metamodel has been architected with design principles that are modularity, layering, partitioning, extensibility, reuse [2]. The UML language is categorized in three major elements that are UML’s basic building blocks, the rules which elicit how building blocks are put together, and some common mechanism that apply throughout the UML. The building block of UML is categorized as Things, Relationship, and Diagram. There are four kinds of things in the UML that are structural things, behavioral things, grouping things, annotational things. The various relationships in UML are dependency, association, generalization, realization. In UML the elements are shown by using graphical representation called Diagrams. The various diagrams
used in UML are Class Diagram, Object Diagram, UseCase Diagram, Component Diagram, Collaboration Diagram, State Diagram, Sequence Diagram, Deployment Diagram, and Activity Diagram. The UML has a number of rules for well formed model. The UML has semantic rules for Names, Scope, Visibility, Integrity, and Execution [2]. A well formed model is also Elided, Incomplete and Inconsistent. The common mechanism are applied throughout the language are Specifications, Adornments, Common Division, Extensibility Mechanism. The UML diagram used various notations [13] for their representation so that they are more suitable to use by the user.

B. Class Diagram

The class diagram is the main building block of object oriented modeling that describes a set of objects which share same properties which are their features, constraints, and semantics. It is used for general conceptual modeling of the systematic of the application, and for detailed modeling translating the models into programming code. Class diagrams can use for data modeling. A class diagram is represented with boxes which contain three sections that are The top part, The middle part, The bottom part. The top part tells the name of the class, The middle part list the attributes of the class, The bottom part gives the methods or operations the class.

![Class Diagram](image)

This Diagram shows the Class Diagram of student having class name STUDENT with attributes StudentName, StudentRollNo, and StudentDeptt and functions Login(), FeeDeposit(), Result(), Subjects().

C. XML

XML stands for eXtensible Markup Language. XML is used to represent large amount of data in expressive and flexible way. XML markup language consists of a set of rules to encode documents in both human-readable and machine-readable form. XML is called extensible as the large number of markup symbols used in XML and they define themselves. The design goals of XML emphasize simplicity, generality, and usability over the Internet. XML is simple and easy to use. XML is a flexible method to develop common information formats and then share the format and the data on the internet. XML is a carrier for information that is incorporated into a document. The XML document is used to store, search and display the data. The large number of applications of XML in Business, Scientific data and across Web develops the methods, techniques to manage and analyze XML data. The various advantages of using XML are that it is easy to handle by people and applications, content and presentation are independent to each other, easy parsing, uses hierarchical structure [12]. The different methods to categorize the problems in XML data management and knowledge discovery [6] are

- **XML Structure Mining**: The XML mining done on structure of data. The applications focused on structural data are pattern matching, change detection, similarity search and detection, and summarization.
- **XML Structure and Content Mining**: The XML mining done on both structure and content features of data. XML Structure and Content Mining is used as research works in semi structured data and text mining.
- **Semantics-aware XML Mining**: With the increase of XML-based application scenario, the mining is done not only on structures and contents of data but also on semantics of the data. The semantic features lead to one of the hardest challenges in data management and knowledge discovery.

D. Semantic and Syntactic Matching

The two techniques of element level granularity are Semantic and Syntactic matching [5]. The element level semantic matching [5] is categorized as

- **Weak Semantic element level technique**
- **Strong Semantic element level technique**

The weak semantic matching depends upon their labels and it is independent of their position in the graph. This is also known as syntactic matching technique. These technique works upon labels of the node, data type, strings, and similar sound words. Besides that the strong semantic element level technique is known as semantic matching technique which deals with the semantics of labels. The semantic matching depends on concepts of node as well as their position in the graph. The semantic matching technique based on the tools which provides the semantic information.

E. Clustering and Classification

The fundamental tasks in XML Mining are Clustering and Classification. **Classification** is the process in which data is grouped according to predefined
methods that are already defined by user. **Clustering** is the process of grouping a set of objects in such a way that objects in the same group are more similar to each other than to those objects which are in other groups. Clustering is an unsupervised classification. The Classification and Clustering algorithm both are used to segment the data. The Classification algorithm segments the data into previously defined groups but the clustering algorithm segments the data into new defined groups. This clustering algorithm uses to explore the natural groups. The aim of clustering algorithm is to discover high-quality clusters having the inter-cluster similarity is low and the intra-cluster similarity is high. The clustering algorithm is categorized according to the technique it uses to cluster the data objects. The various types of clustering algorithms [7] are

1. Hierarchical Clustering Algorithm
2. Partitioning Clustering Algorithm
3. Locality-based Clustering Algorithm
4. Grid-based Clustering Algorithm

II. **SURVEY**

The XML documents are clustered based on their structure only or both the structure and the content [1]. The approach simplifies the complexity of tree representation by providing the flexible representation of XML document based on subsets of their paths. Yoon, Raghavam, Chakilam, Kerschberg says that an XML documents consists of sequence of ePaths [14] which is associated with the element contents [14]. These paths are generated by some criteria such as length of the paths, start from root, or end at the leaf. This method is used as standard method for vocabulary reduction. A complete path is a path which starts from the root of tree and ends at the leaf of tree. The length of the path is the numbers of nodes in the path. The two methods are used for clustering based on path similarities of XML data are SL (same label) and SP (same path) [3]. The SL clustering clusters the nodes according to same tag name. These data nodes are stored in a cluster as their order in the document. Each cluster has a label for data nodes stored in it that is cluster identifier. The numbers of created clusters are same as the numbers of labels in the document. This is a simple and easy method to implement. The SP clustering clusters the nodes having the same absolute path [11]. The data nodes are stored in the cluster according to their order as in SL method. Each cluster has an absolute path for data nodes stored in it that is cluster identifier. This method is implemented using path index such as DataGuide [3]. DataGuide manages the same absolute path as one set. The numbers of created clusters are same as number of DataGuide nodes. Giunchiglia and Shvaiko introduce new type of clustering PSim clustering which is the combination of both SP and SL clustering [5]. The data nodes stored in a SL cluster with the same label can be divided again into one or more SP clusters as their absolute path. The PSim clustering has the features of both SL and SP clustering method. Manning and Kahana provides the lists of words and subsequently recalls the studied items in the order they come to mind that list is known as the free recall [4]. It reflects how the items are stored and retrieved from memory. The recency and primary are two effects of recall. These effects refer to the well-established tendency to show more recall of items from the ends, and to a lesser recall from the beginning of the studied list. The temporal clustering is the tendencies to successively recall items that occupied neighboring positions in the studied lists. The striking effect of semantic clustering is also required with the order of recalls by the study positions of the items. The semantic matching is performed by using two levels that are element level and structure level [5]. The element level semantic matching compute mapping between labels at nodes and the structure level compute mapping between subgraphs. The element level semantic techniques is categorized as weak semantic and strong semantic. The weak element level semantic matching technique is known as syntactic matching and strong element level matching technique is known as semantic matching. The syntactic matching depends only on the labels of the nodes where as semantic matching depends upon labels of nodes as well as with the position of node in the graph.

III. **OBJECTIVE**

The XML data is used in large numbers of applications, thus it is very much necessary to maintain that data.

- The algorithm is used to eliminate the redundancy from the XML document which is prepared from Class Diagram.
- The algorithm emphasis on both the syntax and semantic of XML documents for similarity.
- It provides the group of similar data that are called clusters having the similar properties.

IV. **METHODOLOGY**

The algorithm is proposed to remove redundancy from class diagram through clustering.

- The Class Diagram is converted to XML document.
- The tree data structure is prepared from the XML document.
- The similar data in it are grouped together.
- The new grouped data is arranged as subparts called cluster of data.
- Now these clusters are compared to find the repetition of data in the cluster. The repetition in the
A cluster is detected through the semantic and structure of data.
- Then similar elements in the cluster are removed.
- The clusters are compared with each other to remove the similar elements among them.
- Then all the remaining data in the clusters are grouped together to get the redundant free document.
- Now final class diagram is obtained which is free from redundancy.

V. FUTURE TRENDS

Semantic Clustering works on two aspects one is position of the label and other one is the meaning of the label. Here the work emphasis on the position of label in the class diagram and covers some aspects of the meaning. The more work can be done on the meaning of the label to get redundant free class diagram.

REFERENCES