Clustering Reporeapers Using Data Mining Techniques

Dr. Saranya.K G[1], Keerthana.G[2]
{kgs.cse@psgtech.ac.in[1], keerthanagovindaraju95@gmail.com[2]}

[1]Assistant Professor(S.Gr), [2]PG Student. Department of Computer Science and Engineering
PSG College of Technology[1]

Abstract. Artifact play a major role in software development. An artifact in software development is an actual by-product in which the works are documented and stored in a repository so it can be retrieved upon demand. Artifacts in software engineering involves specific development methods or processes. To make the software development outright, the collection of information is organized as artifact set. Build involves in converting source code into software artifact. The presence of build helps to provide tangible and observable results. Document clustering involves in organizing and managing texts in structured format. Document clustering provides feasible results increasing the accuracy and saves time when compared to manual text classification.

Keywords: Build, k-means, agglomerative clustering, minibatch k-means, elbow k-means..

1 Introduction

Software artifacts provide information about the data, which can be stored and can be retrieved whenever needed. Reporeapers are data representations which consists of software projects under different domains. Build files provide tangible and observable responses thereby reducing the complexity and time. To enhance reusability, the major objective of a domain analysis is to assess the feasibility of reusing a set of artifacts for the development of new software engineering projects. With respect to a particular domain analysis goal, “domains” are a type of software projects among which certain common artifacts may be reused or shared. Reuse of software artifacts comparison enables the data and model for cost, schedule and quality of different software projects. Domain analysis for the reuse of software engineering projects using machine learning involves in classifying the software projects using machine learning technique based on its domains.

Component retrieval is an important issue in supporting software reuse. Reporeapers is a large database containing valuable implicit regularities to be discovered. Domain analysis represents characteristics that may determine whether or not, one or several artifacts can be reused for a new software development. The characterization functions of domains are mapping of projects, described by characteristics into domains.

ICCAP 2021, December 07-08, Chennai, India

DOI 10.4108/eai.7-12-2021.2314573

Related Work

The number of open-source software projects has been growing exponentially. Reapers considers a number of different source artifacts in order to perform a thorough assessment about software engineering projects. Looking at the reporeapers for source artifacts beyond the source code gives much insight into a software project from both a technical and management point of view. It will also be useful for software and management reuse aspect.

Jens Nilsson et al [1] proposed a new approach to improve the robustness of software maintenance tools thereby showing no negative effect on the accuracy. The paper contributes with experimental results on:

1. Data driven dependency parsing of the programming languages C/C++, java and python
2. Transformation between dependency structure and phrase structure adopted to programming languages
3. Generic parser model selection and its effect on parsing accuracy.

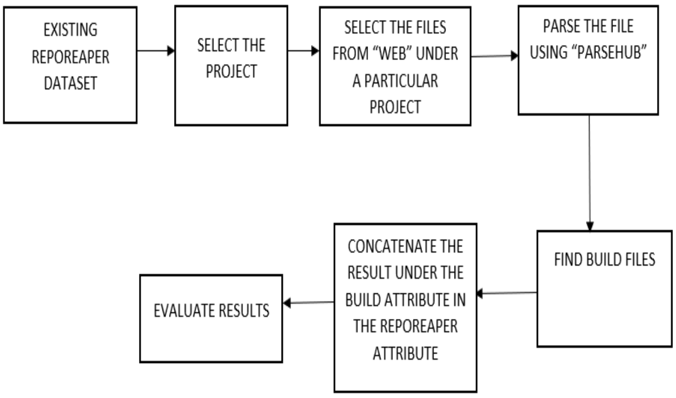
The general approach is divided into two phases training and production, to perform the two phases natural language processing must be applied to the needs of information extraction from the programming language code. Therefore, the source code is converted into syntax tree. The syntax tree is converted into dependency tree and the dependency tree is again converted into syntax tree. These procedures are applied by generating training data and training the generic parser with the training data.

Brian et al [2] proposes the concept of development for the C# programming language to the code Versioning System CVS to checkpoint phases, and a rigorous testing framework to guide the progression from one phase to the next. The work describes the framework in which the application can take place. The development of a parser for the C# programming language is in itself important to software engineering fields since parsers form the basis for many tools used by the software engineers. The vital importance of the paper is that it is simple, reliable parsers exist for commonly used programming languages, since parsers form the basis for many software engineering tools. The work contributes toward the development of a standard parser for C#, ensuring the development of tools for the language.

Monika Gupta et al [3] proposes a process of collecting similar documents into group, where similarity is some function on document. The procedure of document clustering involves certain steps such as pre-processing, where the stop words, since they are frequent and irrelevant. Next the words are stemmed. The second step involves feature extraction, which produce the set of features. This removes the noise and reduce the dimensionality of feature space. The third step involves document representation, which approaches to use the vector space model for document representation. Final step involves Document clustering, where the target documents are grouped into different clusters. The overall goal of text mining to extract information from the large dataset and provide it into an useful and understandable format.

Shalini Verma et al [4] approaches to solve software engineering problems skills, instruction and methods applied. Software engineering is a set of problem-solving skills, instructions and methods applied upon a variety of domains to create useful systems that is used to solve problems. Clustering defines the classes and objects which are similar are put into one group and the clusters which are not similar are put into the other group. The paper presents various techniques for clustering which include: Partitioning clustering, Hierarchal clustering, Well-shaped cluster, Density clustering, Centroid based clustering, K-means clustering. The clustering technique in which large dataset are divided into small dataset in which objects which are similar are clustered in a single group. The various clustering technique is discussed and accuracy is checked. Duo Liu et al [5] proposes a method for Software analysis, which is a complex technique for software maintenance. Clustering technique has been widely used in software engineering project. The various clustering techniques applied in the paper are, Hierarchical clustering, Self-organized map, Traditional Fcm Clustering, Possibilistic clustering, Adaptive SVM.

Proposed Methodolgy


EXISTING REPOREAPER DATASET
SELECT THE PROJECT
SELECT THE FILES FROM "WEB" UNDER A PARTICULAR PROJECT
PARSE THE FILE USING "PARSEHUB"
FIND BUILD FILES
CONCATENATE THE RESULT UNDER THE BUILD ATTRIBUTE IN THE REPOREAPER ATTRIBUTE
EVALUATE RESULTS

Software projects has a wealth of information including process, code and the people which help in the development of software product. Retrospective analysis helps to know about the growth and evolution of the software product. The goal of the reporeaper is to identify the practices that an engineered software practice would come up with the motive of developing a generalized framework to identify the real-world projects. A reference implementation of the evaluation framework known as the reaper, is available as an open source project.

A publicly available data set is got from GitHub repositories.An engineered software project holds software engineering practices in one or more of its attributes such as project management, testing and documentation. To achieve this the essential software engineering practice must be identified. The evaluation framework helps to achieve this goal.

Reporeaper consists of seven attributes into a JAR file. The build file properties control what should go into the plug-in distribution. Build.xml resides in the base directory of the project. There is no constraint on the location of the file and name of the file. All the valid build files require project element and one target element.

Advantages

- As the software becomes more complex, build files allow to establish a standard way of building the program
- Good code comes from building on top of good code, so when a build file is present it is easy to reuse the existing code

Clustering

Clustering is the process of grouping similar objects into a same group than compared to the other objects. Cluster analysis is done by implementing algorithms. Rather than implementing algorithm cluster analysis involves in detail understanding of the data. Therefore, clustering is considered to be a multi-objective optimization problem. The selection of algorithm for clustering depends on the structure of the data and the expected outcome from the data.

K-Means

K-means is one of the unsupervised learning calculations, that is utilized to solve problems related to clustering.

The algorithm works to increase the objective functions in squared error function

The steps involved in the algorithm:

1. First the k points are placed into the space represented by the objects That are being clustered

These points denote the initial group centroids

2. Each object is allocated to the group nearest centroid.

3. When all the objects have been assigned respectively into the groups, the K centroids position is recalculated.

4. The steps 2 and 3 are repeated until the centroids are no longer moving.

This procedure produces a separation of the objects into the group from which the metric to be minimized can be calculated.

Elbow K-Means

The “elbow” method aids the data scientist to procure the best number of clusters by fitting the model with range of values for K.

Agglomerative Clustering

Agglomerative bunching utilizes a base up methodology, wherein every data point begins in its own group. These groups are then joined eagerly, by taking the two most comparable groups together and consolidating them.

Mini Batch K-Means

MiniBatchKMeans works comparatively to KMeans, with one essentialness distinction: the batch_size parameter. batch_size controls the quantity of arbitrarily chosen perceptions in each group. The bigger the measure of the clump, the more computationally expensive the preparation procedure.

Advantages

- It is simple and easy to use for the users
- Easy for searches based on the topics
- When the information or the database is too large, it is difficult to find relevant information. Domain clustering is used to find relevant information.

Evaluation Metrics

Accuracy

Precision is the most instinctive execution measure and it is essentially a proportion of effectively anticipated perception to the aggregate. Exactness is a measure when symmetric datasets where estimations of false positive and false negatives are relatively same.

Accuracy = TP+TN/TP+FP+FN+TN

Precision

Precision is the proportion accurately of correctly predicted positive observations to the total predicted positive observations. Low false positive rate indicates high precision.

Precision = TP/TP+FP

Recall

Recall is calculated by taking the accurately predicted positive samples divided by all the samples corresponding to the actual class.

Recall = TP/TP+FN

F1 Score

F1 score is calculated by taking the weighted average of two other metrics namely Precision and Recall. The advantage of using F1 score is handling of imbalanced class distribution.

F1 Score = 2*(Recall * Precision) / (Recall + Precision)

EVALUATION MEASURE	100	200	300	500
PRECISION	0.95	0.92	0.92	0.922
RECALL	0.95	0.93	0.93	0.93
F-MEASURE	0.94	0.91	0.91	0.91

Experimental result for Build analysis

Approach	**Precision**	**Recall**	**Accuracy**
K-means	0.62	0.65	0.63
Agglomerative clustering	0.84	0.84	0.84
Elbow k-means	0.41	0.78	0.41
Mini Batch K-means	0.60	0.79	0.61

Experimental result for clustering

Conclusion

Build is used to find observable and tangible result. Build in the project is analysed through the ParseHub tool. ParseHub tool retrieves all the projects with the build files. Then these files are compared with the original repository to represent in binary format.

Domain analysis, the first step includes in collecting the domain names from the GitHub topics. The description of each project is retrieved using ParseHub tool. The descriptions are pre-processed to remove stop words. Stemming and tokenization is done to the description. K-means, elbow k-means, agglomerative clustering, mini batch k-means algorithm is applied to the description to find the group of similar descriptions

References

[1] Jens Nilsson, Welf Lowe, Johan Hall, Joakim Nivre, “Parsing Formal Languages using Natural Language Parsing Techniques”, October 2009, 11th International Conference on Parsing Technologies (IWPT)

[2] Brian A James F. Power, Maynooth and John T.Waldon, “Applying Software Engineering Techniques to Parser Design: The Development of a C# Parser”, 2002 , South African Institute of Computer Scientists and Information Technologists

[3] Jan Kurs Mircea Lungu, Oscar Nierstrasz, “Top-Down Parsing with Parsing Contexts A Simple Approach to Context-Sensitive Parsing”, 2014/8/13, ISWT’14

[4] Monika Gupta, Kanwal Garg, “A Review on Document Clustering”, Volume 6, Issue 5, May 2016, International Journal of Advanced Research in Computer Science and Software Engineering

[5] Shalini Verma, Abhinav Mishra, “Various Clustering Techniques in Software Engineering -A Review”, Vol. 4, Issue. 7, July 2015, International Journal of Computer Science and Mobile Computing

[6] Duo Liu, Chung-Horng Lung, Samuel A. Ajila, “Adaptive Clustering Techniques for Software Components and Architecture”, 2015, IEEE 39th Annual International Computers, Software & Applications Conference

[7] Neepa Shah, Dr. Sunita Mahajan, “Distributed Document Clustering Using K-Means”, Volume 4, Issue 11, November 2014 ISSN: 2277 128X, International Journal of Advanced Research in Computer Science and Software Engineering

[8] Chintakindi Srinivasa, Vangipuram Radhakrishnab,Dr.C.V.Guru Rao, “Clustering and Classification of Software Component for Efficient Component Retrieval and Building Component Reuse Libraries”, 2014, 2nd International Conference on Information Technology and Quantitative Management, ITQM

[9] D. S. Vijayan, A. Mohan, J. J. Daniel, V. Gokulnath, B. Saravanan, and P. D. Kumar, “Experimental Investigation on the Ecofriendly External Wrapping of Glass Fiber Reinforced Polymer in Concrete Columns,” vol. 2021, 2021.

[10] Ramanpreet Kaur and Amandeep Kaur, “Text Document Clustering and Classification using K-Means Algorithm and Neural Networks”, October 2016 , Indian Journal of Science and Technology.

[11] Vivek Kumar Singh , Nisha Tiwari, Shekhar Garg, “Document Clustering using K-means, Heuristic K-means and Fuzzy C-means”, 29 December 2011, International Conference on Computational Intelligence and Communication Networks

[12] Rakesh Chandra Balabantaray, Chandrali Sarma, Monica Jha, “Document Clustering using K-Means and K-Medoids”, 2011 IEEE, International Conference on Computational Intelligence and Communication Systems

[13] Jaya Zade, Dr. G. R. Bamnote , Prof. P. K. Agrawal, “Text Document Clustering Using K-means Algorithm with Its Analysis and Implementation”, Vol-3, Issue-2, 2017, Imperial Journal of Interdisciplinary Research (IJIR).