iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 3 **Issue:** VII **Month of publication:** July 2015

DOI:

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)

Clustering Approach to Analyse Language trend in MNCs

Nutan Dahiya[#1] , Rachna Dhaka[*2]
[#] *Students, Computer Science & Engineering, Gateway Institute Of Engineering and Technology, Sonipat, Haryana(India)*
[*] *HOD, Computer Science & Engineering, Gateway Institute Of Engineering and Technology, Sonipat,Haryana(India)*

Abstract*—Data mining system discovers patterns and relationships hidden in data, and actually is a part of a larger process called "knowledge discovery" which describes the steps that must be taken to ensure meaningful results. The presented work will focus on implementing different data mining approaches on database which is simulated from different companies. The work has combined two major approaches to provide profiling of companies and language analysis via clustering and association rules. Clustering will be implemented to profiling companies according to the languages. After profiling, work will focuses on finding the trend of programming languages. The system will next identify the most dominant cluster among all the clusters formed.. A comparison between the results which will be obtained from using two different algorithms will be done. One algorithm used is Apriori algorithm and the other one is Simple K Means algorithm. The results obtained from these two algorithms will be identified and compared for the purpose of validation of trends of programming language. The work has been implemented in Weka Tool. This paper describes the Clustering approach very well.
Keywords: Language trend, Weka tool, Data mining, Clustering, Segments.

I. INTRODUCTION

Data mining (sometimes called data or knowledge discovery) is the process of analyzing data from different perspectives and summarizing it into useful information -information that can be used to increase revenue, cuts costs, or both. Data mining software allows users to analyze data from many different dimensions or angles, categorize it and summarize the relationships identified. Technically, data mining is the process of finding correlations or patterns among dozens of fields in large relational databases. Data mining is primarily used today by companies with a strong consumer focus - retail, financial, communication and marketing organizations. It enables these companies to determine relationships among "internal" factors such as price, product positioning, or staff skills, and "external" factors such as economic indicators, competition, and customer demographics. And, it enables them to determine the impact on sales, customer satisfaction, and corporate profits. Finally, it enables them to "drill down" into summary information to view detail transactional data. With data mining, student recruitment intelligence system find out point-of-interest records of students for employment to send targeted promotions based on an individual's knowledge. By mining they can find out the required information data to find out the relation between different companies and students according to their intelligence. This project is going to describe the activity related to **recruitment**, its various problems and their solutions. It uses MIS and ERP system for managing various administrator tasks like student data for recruitment.

Management information systems (MIS) are typically computer systems used for managing five primary components:
Hardware, Software, Data (information for decision making) , Procedures (design, development and documentation), People (individuals, groups, or organizations)

Academically, the term is commonly used to refer to the study of how individuals, groups, and organizations evaluate, design, implement, manage, and utilize systems to generate information to improve efficiency and effectiveness of decision making, including systems termed decision support systems, expert systems, and executive information systems. Most business schools (or colleges of business administration within universities) have an MIS department, alongside department accounting, finance, management, marketing.

Enterprise resource planning (ERP) is a cross-functional enterprise system driven by an integrated suite of software modules that supports the basic internal business processes of a company. ERP software integrates all facets of an operation, including product planning, development, manufacturing processes, sales and marketing.

ERP gives a company an integrated real-time view of its core business processes such as production, order processing, and inventory management, tied together by ERP applications software and a common database maintained by a database management system

Recruitment refers to the overall process of attracting, selecting and appointing suitable candidates to a one or more jobs

within an organization, either permanent or temporary. Recruitment of candidates is the function preceding the selection, which helps create a pool of prospective employees for the organisation so that the management can select the right candidate for the right job from this pool. The main objective of the recruitment process is to expedite the selection process. There are several recruitment groups that provide employment Opportunities and world class training to students of the Institute in-campus or off-campus in leading organizations/Industry.

All the reputed organizations including MNCs come to the Institute for Campus Recruitment. To achieve its goal, the Placement Cell works towards recognizing the core competencies of students.

A. Significance of the problem

Company recruits the new enrolled students on the basis of company's needs, student grade, technical language and protocols the companies defined for recruitment process. This study can also do a comparative analysis by using different algorithms to validate the result. Unlike the existing system this is an efficient approach (algorithm) which is easy to implement and does the following tasks.

1) Profiling the students according to the Companies in which they placed.
2) It can analyze the language(s) according to which company can recruit.
3) To target the names of companies earlier according to the grades and technical language the students knows.
4) To find out how many students were recruited by which company.
5) To explore the relation of recruitment group with high profiled MNC's.

It also validates the result comes from two different approaches.

II. RESEARCH BACKGROUND

In year 2013, Dorina Kabakchieva, Predicting Student Performance by Using Data Mining Methods for Classification. Data mining methods are often implemented at advanced universities today for analyzing available data and extracting information and knowledge to support decision-making.

In year 2014, Ajay kumar, Swati Singhal, Praveen Dhankher and Anju Gulia, a study was done which works in three phases. First phase, provide profiling of students according to their grades via clustering then, a number of different segments are formed on the basis of type of grades provided to the student by the institution. The students according to their grades are grouped into different clusters according to the company rating. Second phase considers validating the student data using J-48 algorithms. In third phase, a comparison is made between different algorithms (J-48, LMT and Bayes Net algorithms) to improve prediction accuracy of student recruitment datasets.

In year 2015, Praveen Rani, Dr. Rajan Vohra, presents a comprehensive statistical experiment to identify the number of students those are ready for placements and students those are not fulfilling the basic criteria for placement from a large database of all computer engineering students of a college containing their academic record. The design of experiments software named Weka Tool is used for making three clusters of whole database which will categorise the students according to their qualifications.

III. RESEARCH METHODOLOGY

For solving the problem some research techniques and methodologies are used for obtaining the desired result.

A. First of all literatures and research papers were reviewed for getting more information about the problem and knowing which type of work was done by others on this topic and by which method.
B. Then tools required for solving the problem were identified and the best tool was selected from all. Organize filed visits to Engineering colleges that conduct placement drive for reputed companies
C. Determine nature and definition of research problem and work flow of the problem for getting accurate and desired result.
D. Organize the database with useful attributes and populate it then perform data analysis using suitable tool e.g., WEKA in order to generate the result.

IV. CONCEPTUAL FRAMEWORK

Data mining (the analysis step of the "Knowledge Discovery in Databases" process, or KDD), an interdisciplinary subfield of computer science, is the computational process of discovering patterns in large data sets involving methods at the intersection of artificial intelligence, machine learning, statistics, and database systems. The overall goal of the data mining process is to

extract information from a data set and transform it into an understandable structure for further use. Aside from the raw analysis step, it involves database and data management aspects, data pre-processing, model and inference considerations, interestingness metrics, complexity considerations, post-processing of discovered structures, visualization, and online updating.
Cluster is a group of objects that belong to the same class. In other words the similar object are grouped in one cluster and dissimilar are grouped in other cluster. Clustering is the process of making group of abstract objects into classes of similar objects. Cluster analysis or clustering is the task of grouping a set of objects in such a way that objects in the same group (called a cluster) are more similar (in some sense or another) to each other than to those in other groups (clusters). It is a main task of exploratory data mining, and a common technique for statistical data analysis, used in many fields, including machine learning, pattern recognition, image analysis, information retrieval and bioinformatics. K-Means is one of the simplest unsupervised non-hierarchical learning methods among all partitioning based clustering methods. It classifies a given set of *n* data objects in *k* clusters, where *k* is the number of desired clusters and it is required in advance.

A. Profiling Of Companies

In this problem, all the companies are grouped according to the no. of recruitments they held. The result will in form of different groups or we can say different clusters that provide information about the number of students and group of companies that recruit these students. The whole data of recruitment process is refined and feed to Weka tool for clustering. Then companies are segregated into different groups. A few dataset is taken from an recruitment group and then whole database is simulated to analyze the result.

In fig.1 c1, c2, c(n) are various clusters refers different groups of student and companies according to grades and protocols of different companies.

B. Language Trend Analysis Via Clustering

Dataset of different segments (which are found by profiling), will collect and feed to Weka to find out the dominant language in which particular company recruit the most. Language Trend will be discovered by using clustering method. It will find out the dominant language for each company. From here we come to know about the type of language, dominant language and particular language preferred by particular company during recruitment process.

The same database is feed to Weka and form four clusters. These cluster contain information aboutcompanies as well as the dominant language in each company. Dominant can be defined as maximum number of candidate placed by company on basis of their technical language. The company having maximum entries of same language is dominant language for that company.

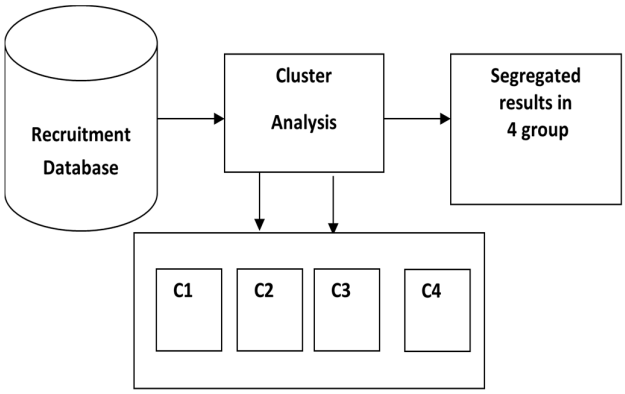


Fig. 1 profiling the companies

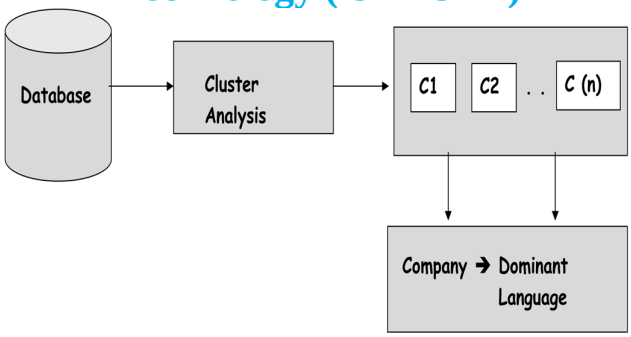


Fig. 2. Language trend analysis via clustering

V. RESULTS AND DISCUSSION

The data analysis is processed using WEKA data mining tool for exploratory data analysis A database of 448 records/entries are created and simulated for the purpose of solving first problem. After loaded the database in Weka all other attributes are removed and keep only two attributes named language and company.

A. Profiling Of Companies

A database of 448 records/entries are created and simulated for the purpose of solving first problem. In this part, the whole data of recruitment is refined and input to Weka tool for clustering. Then companies are segregated into different groups. In Fig 3 database snapshot is shown.

In this we only consider language and company for getting best results.

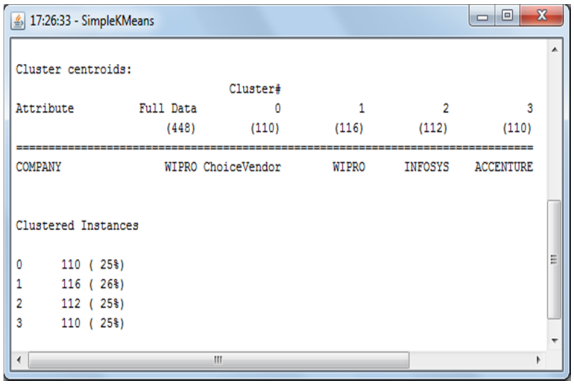


Fig. 3 Profiling of Companies- clusters formation

Here are the above said results represent in tabular form.

Cluster Number	Company	Count of Placed Candidates
Cluster 0	ChoiceVendor	110
Cluster 1	Wipro	116
Cluster 2	Infosys	112
Cluster 3	Accenture	110

Table 1 Profiling of Companies- clusters formation

B. Discovering Language Trend Via Clustering

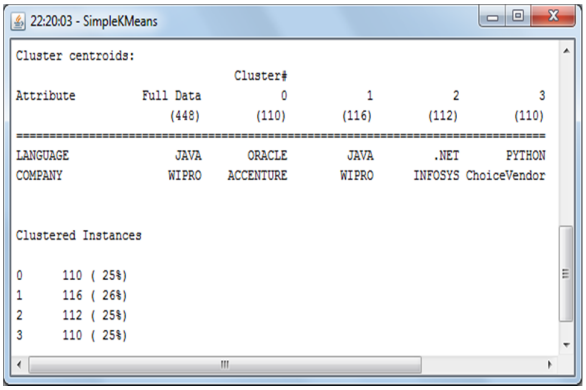


Fig. 4 Cluster View

Fig. 6.5 show the relation between Language and Company. Clusters are formed on basis of 4 different companies and the dominant languages in all clusters or we can say companies.

The information retrieved from here is as follows:

1) Cluster 0 represents recruited candidates having selection in Accenture Company on basis of Oracle language.
2) Cluster 1 represents recruited candidates having selection in Wipro Company on basis of Java language.
3) Cluster 2 represents recruited candidates having selection in Infosys Company on basis of .Net language.
4) Cluster 3 represents recruited candidates having selection in ChoiceVendor Company on basis of Python language.

Language Wise Distribution of The Companies

Following table represents all Languages in each Company

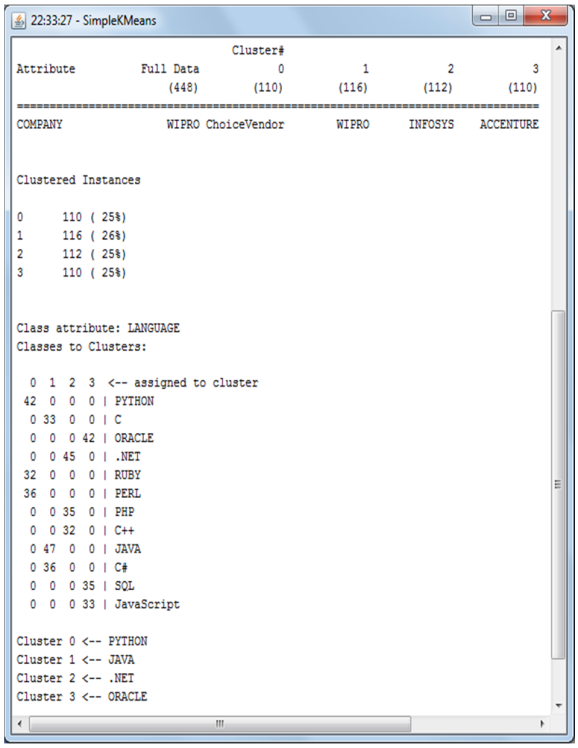


Table 2 Language trend via Clustering

Cluster Number	Company	Language
Cluster 0	ChoiceVendor	PYTHON
Cluster 1	Wipro	JAVA
Cluster 2	Infosys	.NET
Cluster 3	Accenture	ORACLE

Fig. 5 Language wise distribution of the Companies

VI. CONCLUSION

In this fast paced world technology has become an essential part of any organization. Every company tries to use best and updated technology in combination of software and hardware to gain an edge over its competitors. There are different set of programs written in different languages used by organizations all over the world. This thesis is made for the purpose to find out these programming languages used by companies while designing software. There are many programming languages ie. .net, java, oracle etc. available to design programs and selection among these is done by developers by giving preference to security, design, accessibility. The main motive of this thesis is to figure out which IT company uses which language to develop its own programs which will be used by other companies in outside world for daily operations. This also helps students in getting their dream job in their desired companies by knowing those companies' requirements.

VII. FUTURE WORK

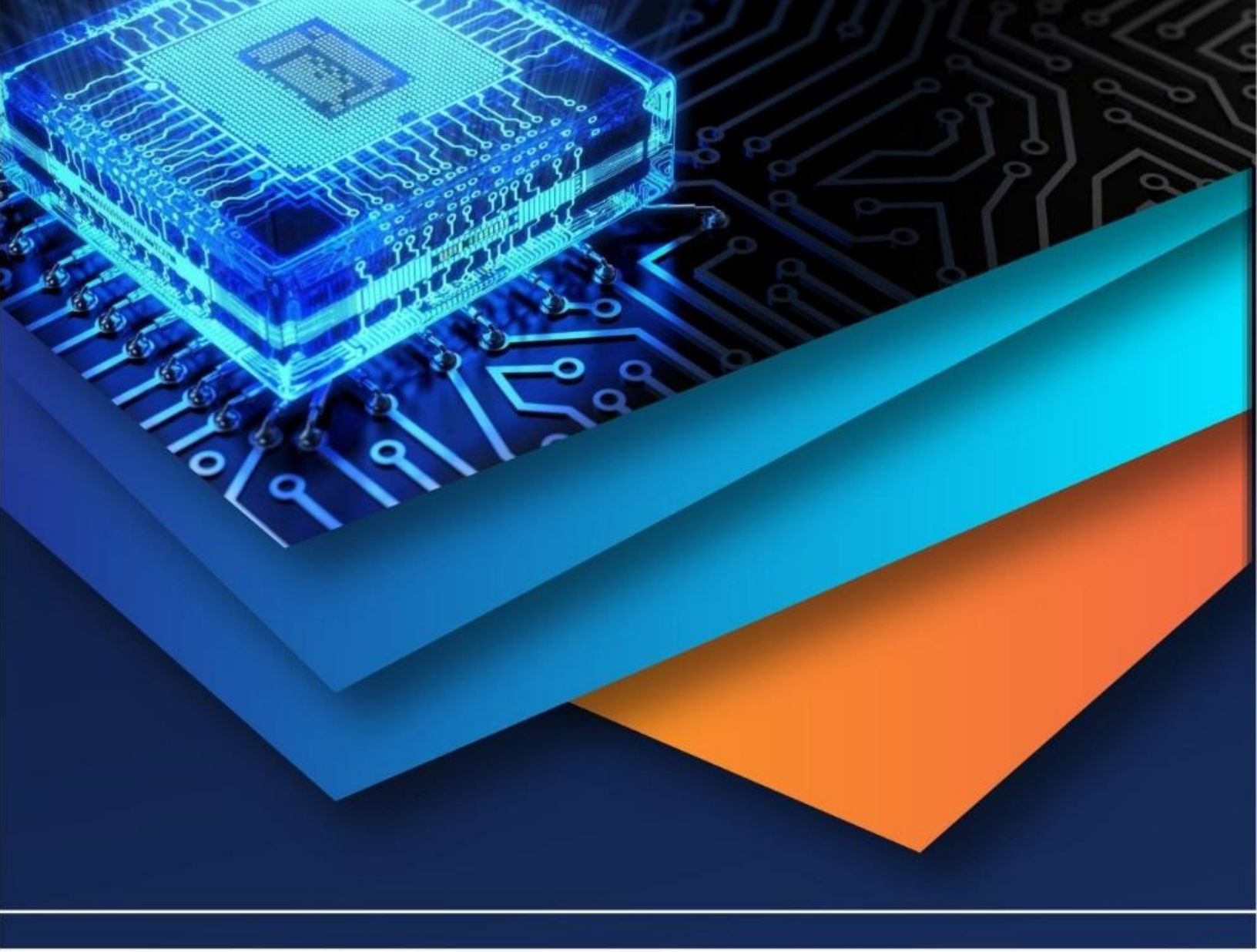
The language analysis can also be done by applying a different approach. Furthermore their results can also be compared to validate trends discovered by two or more different approaches, concepts, algorithms etc. So this research has a good scope to move further and the another paper of this research gives a clear glimpse of that.

VIII. ACKNOWLEDGEMENT

Author would like to thank to their head Ms. Rachna, HOD Computer Science & Engineering, Gateway Institute of Engineering and Technology, Sonipat, Haryana for her continuous and valuable support during this whole project.

REFERENCES

[1].http://mckinseyonsociety.com/education-to-employment/report/
[2].http://www.naukrihub.com/recruitment/importance-of-recruitment.html
[3].http://www.uni-weimar.de/medien/webis/teaching/lecturenotes/machine- learning/unit-en-decision-trees-algorithms.pdf
[4].http://www.cs.princeton.edu/courses/archive/spr07/cos424/papers/mitchell- dectrees.pdf
[5]. http://www.ise.bgu.ac.il/faculty/liorr/hbchap9.pdf
[6]. http://www.d.umn.edu/~padhy005/Chapter5.html
[7]. http://www.cs.ccsu.edu/~markov/ccsu_courses/DataMining-7.html
[8]. http://www.cs.umd.edu/~samir/498/10Algorithms-08.pdf
[9]. http://wiki.pentaho.com/display/DATAMINING/Classifiers
[10].http://www.anderson.ucla.edu/faculty/jason.frand/teacher/technologies/palace/datamining.html
[11]. http:// www.ijcsit.com/.../ijcsit2015060371.pdf
[12]. http:// www.ijcsms.com/journals/volume%2520.pdf



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY