A Novel Approach to Smart Higen Chatbot System by Using Data Mining Algorithm

Priya J. Sahane¹, Prof A. N. Nawathe², Prof S. U. Kadlag³
¹ME, Computer Engineering, Sangamner, Maharashtra, INDIA.
²³Assistant Professor, Sangamner, Maharashtra, INDIA.

Abstract— Data mining is important and useful tool as compare to several efficient tools for extracting data that are stored in a database. Frequent pattern mining is one of data mining algorithm, which will extract the items that occur more number of times. Data may be in the form of facts, numbers, or text. In current years, Organizations are accumulating vast and increasing amounts of data in different formats and different databases. In Frequent itemset mining, which focuses on finding a association among data. Change mining, which detects and report any considerable changes if occurs in the set of mined itemsets from different time periods. This project extends the Dynamic change mining problem, in the framework of frequent itemsets, by exploiting recurrent generalized itemsets to characterize information linked with infrequent patterns. To address this issue, I introduce two novel kinds of vibrant patterns, namely the HIGEN MINER (History Generalized pattern) and NON-REDUNDANT HIGEN MINERS. HIGEN MINER detects the mined frequent itemset, if any items become infrequent during next extraction those changes are focused and avoided by Apriori based support driven approach. NON-REDUNDANT HIGEN MINER, which stores items whose support value is minimum, these infrequent items may become frequent during next extraction. If it remains same, then these infrequent items are discarded. HIGEN Chatbot System is the system that helps to understand the human conversation.

A chatbot is a computer program designed to simulate an intelligent conversation with one or more human users. Primarily for engaging different types of application. The main aim of HIGEN chatbot is support to multiple domains the user into thinking that the program's output has been produced by a human. Programs playing this role are sometimes referred to as Artificial Intelligence Conversational Entities, talk bots or chatterboxes. The objective of the paper is to predict more accurately the presence of type of domain suppose user wants any type of questions of answers of any types of domains such as Medical related queries, Library related questions etc. In proposed work HIGEN chatbot gives the correct answers to particular questions using two variant kinds of algorithms Modified HIGEN and Non redundant HIGEN and techniques are used Naive Bayes and NLP are used to implement this work and comparison of these two system has been given in the present paper based on accuracy. As per the results, accuracy of HIGEN chatbot system and ALICE chatbot system is likely to have 98%, and 50% respectively.

The analysis shows that out of these two chatbot system is predict the accurate answer to the particular question with highest accuracy.

Keywords— HIGEN Miner, Non-Redundant HIGEN, Frequent itemset mining, NLP, Naive Bayes classification.

I. INTRODUCTION

Data mining can be defined as the extraction of useful knowledge from large amount of data repository. Compared with other data mining application fields, medical data mining shows a fundamental role and it has some unique characteristics. Data mining techniques are the result of a long process of research and product development. This development begins when business data was first stored on computers, continued with progress in data access, and more recently, generated new technologies that allow users to navigate through their data in real time. Data mining takes the evolutionary process beyond retrospective data access and navigation to prospective and proactive information delivery. Data mining is ready for application in the business area because it is supported by three technologies that are now sufficiently mature: Huge data collection, Data mining algorithms, Powerful multiprocessor computers. The medical data mining has the high potential in medical domain for extracting the hidden patterns in the datasets [4].

These patterns are used for clinical diagnosis and prediction. The medicinal data are widely spread, heterogeneous, voluminous in nature. The data should be combined and collected to provide a user oriented approach to novel and hidden patterns of the data. For ultimate analysis, normally, numerous tests generally involve the classification or clustering of large scale data. The test procedures are said to be necessary in order to reach the you have to entered correct question and answer and also check system give you correct answer of particular question of particular domain. This type of difficulty could be resolved with the aid of machine learning which could be used directly to obtain the end result with the aid of several artificial Intelligent algorithms which perform the role of classifiers.
Classification is one of the maximum techniques used in data mining. If a classification process is to be done[5]

In Previous years, Different types of chatbot systems are available such as HEBOTO, ALICE BOT, MEGAHAL, etc; this systems are used for chatting purpose but general chatting say it as Hi and Hello this fixed type of dataset is used in previous chatbot systems. All this chatbot is developed up to the 2002 year in this year work on chat systems was totally stopped. To develop all this chatbot systems using pattern matching algorithm matching process is done word by word matching and it required more time to give the answer to the human or user adding and deleting facility is not provided to this all chatbot systems some basic questions are stored and it not support any one particular domain it support only specific question and answer for in this systems developing New HIGEN Chatbot system which will be used to provide the accurate answers of any type of domain.

II. LITERATURE SURVEY

1. ALICE System

ALICE: the Artificial Linguistic Internet Computer Entity. Is a robot that you can make chatting with it. ALICE knowledge is stored in AIML files. AIML is an abbreviation of The Artificial Intelligent Mark up Language that is a derivative of Extensible Mark up Language (XML). The next sections describe necessary information about AIML elements, its categories, how they are used in ALICE and Pattern Matching Algorithm.

1.1 AIML Files

Each AIML file start with an <aiml> tag represent the AIML version being used, and it contains the AIML elements which consists of data objects called AIML objects. These objects are made up of units called topics and categories, which contain either, parsed or unparsed data. The topic is an optional top level element, has a name attribute and a set of categories related to that topic. Each category contains a pattern represent the user input and a template implies robot response[6]

1.2. Preparation for Pattern Matching

Before starting pattern matching procedures each input to the AIML interpreter must pass through two processes:
1- Normalization Process.
2- Producing input path from each sentence.

1.3.1 Normalization Process, involves 3 steps:

a. Substitution Normalizations:
are heuristics applied to an input that attempts to retain information in the input that would otherwise be lost during the sentence splitting or pattern fitting normalization. It can distinguish the dot notation if it is used as an abbreviation, end of sentences or just prefix of extension name and replace it by its appropriate meaning.

b. Sentence Splitting Normalization:

Split input into sentences using rules like “break sentences at periods”, after ? and !.

c. Pattern Fitting Normalization:

It involves two tasks: Removing punctuation from input to make it compatible with speech conversation. Converting input letters to upper case, this is necessary, because if origin user input was either lower or upper case it will be matched with patterns stored in capital letters but the opposite is not true[6].

1.3.2 Elizabeth

It is an adaptation of Joseph Weizerbaum’s Eliza program, in which the various selection, substitution, and phrase storage mechanisms have been enhanced and generalized to increase both flexibility and its potential adaptability.

2.1 Script File format

Knowledge is stored as a script in a text file, where each line in this text is started with a script command notation to distinguish between them, these notations are:

W, Q, V, I, K, N, O, M, & , AND / , that denotes in order, welcome message, quitting message, void input, input transformation, key word pattern, key word response pattern, output transformation, memorized phrase, action to be performed within a message, and a comment.

The script file may contain at most 4 parts as bellow:

**Part One:** Script command lines holding robot responses dealing with the cases of welcome, void and no key word messages. Welcome messages: begins with ‘W’ letter and one of them will be selected randomly by the system when it is started. Messages: begins with ‘V’ letter and one of them will be selected when user input is empty, the case when the user just press enter. No key word messages: begins with ‘N’ letter and one of them will be selected randomly by the system when there is no keyword pattern match occurs[6].
III. CHAT BOT SYSTEM

Chatbots are computer programs that interact with users using natural languages. This technology started in the 1960’s; the aim was to see if chatbot systems could fool users that they were real humans. However, chatbot systems are not only built to mimic human conversation, and entertain users. In this paper, we investigate other applications where chatbots could be useful such as education, information retrieval, business, and e-commerce. A range of chatbots with useful applications, including several based on the ALICE/AIML architecture, Chatterbots are often integrated into the dialog systems of, for example, automated online assistants, giving them the ability of, for example, small talking or engaging in casual conversations unrelated to the scopes of their primary expert systems. Large companies such as Lloyds Banking Group, Royal Bank of Scotland, Renault and Citroën are now using automated online assistants instead of call centres with humans to provide a first point of contact [6].

IV. DATASET

As per review of existing system in ALICE chatbot system were used fixed dataset i.e. Questions and Answers. But in proposed system variable dataset and it can be adopt any type of domains such as library dataset, medical dataset, agricculture dataset.

V. PROPOSED SYSTEM

The design of the system requires the complete understanding of the problem domain. The proposed system has been developed with an aim to efficiently used chatbot system. The objective of the proposed system in that adding and deleting facility is provided.
This proposed system can handle 16 GB data where we can store different types of questions and answers. This can adopt any type of domain such as Medical dataset, Library dataset, Agriculture dataset, Educational dataset, etc.

Chatter bots are often integrated into dialog-oriented for various practical purposes such as offline help, data service, or information acquisition. Chatter bots use sophisticated natural language processing systems, but many simply scan for keywords within the input and pull a reply with the most matching keywords, or the most similar wording pattern, from a textual database.

In all work of chatter bots, all the data such as keywords, questions and responses are previously stored. At the time of responding these matching are performed and responded. So over a period of time, the things are the same to respond. So a chatter bot will give you the same answer after five years as now.

Medical queries is an important application of classification for example; search of new question and answers based on their symptoms by using the classification rules about diseases from known cases. Decision tree is a way of implementing the classification.

2) Higen Algorithm

The HIGEN MINER algorithm is used to search frequent itemset in particular time interval. HIGEN is used to measure the frequent item set in given data set. Here multiple generation of itemset over different time duration is avoided. Generally data’s are distributed all over web and stored in various format or may contain different or missing contents. Data cleaning, Data extraction, Data loading which removes noisy, irrelevant data and not just about removing bad data or missing values, but also discovering hidden relationships in the data, finding where the data came from that are most accurate, and influential which parts are the majority appropriate for use in analysis.

Multiple taxonomy evaluations over the same pattern, the generalizing procedure of each itemset is belated after its support estimate in all time stamped itemsets and hierarchically applied on infrequent indiscriminate itemsets of escalating generalization level. To reduce the extraction time, HIGEN generation is performed quickly, without the need of a post processing step. Also, a customized adaptation of the HIGEN MINER algorithm is anticipated to address NON-REDUNDANT HIGEN extraction.

3) Non-Redundant Higen Miners

In the set of mined itemsets, the frequent items are placed in HIGEN MINER and infrequent items are placed under NON-REDUNDENT HIGEN MINER. The infrequent items will be observed over specific time period to take changes into account if any.

---

**VI. DATA MINING TECHNIQUE FOR HIGEN CHATBOT SYSTEM**

1) Classification Technique

Classification is a concept of finding a model which finds the class of unknown objects. It basically maps the data items into one of the some predefined classes. Classification model generates a set of rules based on the features of the data in the training dataset. Further these rules can be used for classification of future unknown data items. Classification is the one of the most important data mining technique.
FIG.5 ADMIN LOGIN WINDOW

FIG.6 HIGEN CHATBOT SYTSEM FOR ADDING Q & A

FIG.7 TO ADD NEW Q & A OF PARTICULAR DOMAIN

FIG.8 TO TEST OLD Q & A

FIG.9 GRAPHICAL REPRESENTATION OF HIGEN CHAT BOT SYSTEM AND ALICE CHAT BOT SYSTEM

VII. CONCLUSION

A Novel Approach To Smart Higen Chatbot System By Using Data Mining Algorithm Using Data Mining Techniques System Is Developed Using HIGEN Chat bot system. The Objective Of The Proposed Work Is To Predict More Accurately The Presence Of Questions and Answers With Reduced searching time using ALML interpreter and unsing two novel kind of dynamic pattern HIGEN Miner and Non Redundant HIGEN for different types of domain.
VIII. FUTURE SCOPE

In our future work, chatterbot will connect to users online chat accounts to get self-trained and automatically generate History Blocks. These Hi-Blocks will help to decide the next response to the query. Various classification algorithms need to be tested for better results. One of our work will also include privacy preservation of chat of the users so that the data captured by these will not reveal users identity and his personal chats. Extension of this work can be the utilization of voice recognition for direct audio interaction. Addition of multi-languages is also an important task yet to be done.

REFERENCES