Chatbot for College Related Queries

Mr. Sathis Kumar .T
Assistant Professor
Department of Computer Science & Engineering
Saranathan College of Engineering-620012, India

N. Vijay Kumar
UG Student
Department of Computer Science & Engineering
Saranathan College of Engineering-620012, India

R. R. Vinodh
UG Student
Department of Computer Science & Engineering
Saranathan College of Engineering-620012, India

U. Vinoth Kumar
UG Student
Department of Computer Science & Engineering
Saranathan College of Engineering-620012, India

T. Vivekananthan
UG Student
Department of Computer Science & Engineering
Saranathan College of Engineering-620012, India

Abstract

The project is to ask college related queries and get the responses through a chatbot an Artificial Conversational Entity. This System is a web application which provides answer to the query of the student. Students just have to query through the bot which is used for chatting. Students can chat using any format there is no specific format the user has to follow. This system helps the student to be updated about the college activities.

Keywords: Specific Requirements, Data Mining, Evaluation of System, Software Description

I. INTRODUCTION

The College bot project is built using artificial algorithms that analyses user's queries and understand user's message. This System is a web application which provides answer to the query of the student. Students just have to query through the bot which is used for chatting. Students can chat using any format there is no specific format the user has to follow. The System uses built in artificial intelligence to answer the query. The answers are appropriate what the user queries. The User can query any college related activities through the system. The user does not have to personally go to the college for enquiry. The System analyses the question and then answers to the user. The system answers to the query as if it is answered by the person. With the help of artificial intelligence, the system answers the query asked by the students. The system replies using an effective Graphical user interface which implies that as if a real person is talking to the user. The user just has to register himself to the system and has to login to the system. After login user can access to the various helping pages. Various helping pages has the bot through which the user can chat by asking queries related to college activities. The system replies to the user with the help of effective graphical user interface. The user can query about the college related activities through online with the help of this web application. The user can query college related activities such as date and timing of annual day, sports day, and other cultural activities. This system helps the student to be updated about the college activities.

II. SPECIFIC REQUIREMENTS

A. Data Mining

Data mining (the analysis step of the "Knowledge Discovery in Databases" process, or KDD), a field at the intersection of computer science and statistics, is the process that attempts to discover patterns in large data sets. It utilizes 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 preprocessing, model and inference considerations, interestingness metrics, complexity considerations, post-processing of discovered structures, visualization, and online updating. Generally, 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 is one of a number of analytical tools for analyzing data. It allows users to analyze data from many different dimensions or angles, categorize it, and summarize the relationships identified. Technically, amounts of data in different formats and different databases. This includes:

Operational or data mining is the process of finding correlations or patterns among dozens of fields in large relational databases.

1) Process of Data Mining

– Data Data are any facts, numbers, or text that can be processed by a computer. Today, organizations are accumulating vast and growing transactional data such as, sales, cost, inventory, payroll,
– Nonoperational data, such as industry sales, forecast data, and macro-economic data
– Meta data: data about the data itself, such as logical database design or data dictionary definitions

1) Information

The patterns, associations, or relationships among all this data can provide information. For example, analysis of retail point of sale transaction data can yield information on which mobile apses are selling and when.

2) Knowledge

Information can be converted into knowledge about historical patterns and future trends. For example, summary information on retail supermarket sales can be analyzed in light of promotional efforts to provide knowledge of consumer buying behavior. Thus, a manufacturer or retailer could determine which items are most susceptible to promotional efforts.

3) Data Warehouses

In computing, a data warehouse (DW or DWH) is a database used for reporting and data analysis. It is a central repository of data which is created by integrating data from multiple disparate sources. Data warehouses store current as well as historical data and are commonly used for creating trending reports for senior management reporting such as annual and quarterly comparisons. The data stored in the warehouse are uploaded from the operational systems (such as marketing, sales etc., shown in the figure to the right). The data may pass through an operational data store for additional operations before they are used in the DW for reporting. The typical ETL-based data warehouse uses staging, integration, and access layers to house its key functions. The staging layer or staging database stores raw data extracted from each of the disparate source data systems. The integration layer integrates the disparate data sets by transforming the data from the staging layer often storing this transformed data in an operational data store (ODS) database. The integrated data are then moved to yet another database, often called the data warehouse database, where the data is arranged into hierarchical groups often called dimensions and into facts and aggregate facts.

A data warehouse constructed from integrated data source systems does not require ETL, staging databases, or operational data store databases. The integrated data source systems may be considered to be a part of a distributed operational data store layer. Data federation methods or data virtualization methods may be used to access the distributed integrated source data systems to consolidate and aggregate data directly into the data warehouse database tables. Unlike the ETL-based data warehouse, the integrated source data systems and the data warehouse are all integrated since there is no transformation of dimensional or reference data. This integrated data warehouse architecture supports the drill down from the aggregate data of the data warehouse to the transactional data of the integrated source data systems.

Data warehouses can be subdivided into data marts. Data marts store subsets of data from a warehouse. This definition of the data warehouse focuses on data storage. The main source of the data is cleaned, transformed, cataloged and made available for use by managers and other business professionals for data mining, online analytical processing, market research and decision support However, the means to retrieve and analyze data, to extract, transform and load data, and to manage the data dictionary are also considered essential components of a data warehousing system. Many references to data warehousing use this broader context. Thus, an expanded definition for data warehousing includes business intelligence tools, tools to extract, transform and load data into the repository, and tools to manage and retrieve metadata.

Dramatic advances in data capture, processing power, data transmission, and storage capabilities are enabling organizations to integrate their various databases into data warehouses. Data warehousing is defined as a process of centralized data management and retrieval. Data warehousing, like data mining, is a relatively new term although the concept itself has been around for years. Data warehousing represents an ideal vision of maintaining a central repository of all organizational data. Centralization of data is needed to maximize user access and analysis. Dramatic technological advances are making this vision a reality for many companies. And, equally dramatic advances in data analysis software are allowing users to access this data freely. The data analysis software is what supports data mining. It enables these companies to determine relationships among "internal" factors such as price, mobile apps 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.

2) Levels of data mining

1) Data mining elements

Extract, transform, and load transaction data onto the data warehouse system. Store and manage the data in a multidimensional database system. Provide data access to business analysts and information technology professionals. Analyze the data by application software. Present the data in a useful format, such as a graph or table.

3) Different Levels of Analysis

1) Artificial neural networks

Non-linear predictive models that learn through training and resemble biological neural networks in structure. Genetic algorithms: Optimization techniques that use processes such as genetic combination, mutation, and natural selection in a design based on the concepts of natural evolution.

2) Decision Trees

Tree-shaped structures that represent sets of decisions. These decisions generate rules for the classification of a dataset. Specific decision tree methods include Classification and Regression Trees (CART) and Chi Square Automatic Interaction Detection (CHAID). CART and CHAID are decision tree techniques used for classification of a dataset. They provide a set of rules that you can apply to a new (unclassified) dataset to predict which records will have a given outcome. CART segments a dataset by creating 2-way splits while CHAID segments using chi square tests to create multi-way splits. CART typically requires less data preparation than CHAID.

3) Nearest Neighbor Method

A technique that classifies each record in a dataset based on a combination of the classes of the k record(s) most similar to it in a historical dataset (where k 1). Sometimes called the k-nearest neighbor technique.

Rule induction: The extraction of useful if-then rules from data based on statistical significance.

Data visualization: The visual interpretation of complex relationships in multidimensional data. Graphics tools are used to illustrate data relationships.

B. Clustering

Clustering is a data mining technique that makes meaningful or useful cluster of objects that have similar characteristic using automatic technique. Different from classification, clustering technique also defines the classes and put objects in them, while in classification objects are assigned into predefined classes. To make the concept clearer, can take library as an example. In a library, mobile appss have a wide range of topics available. The challenge is how to keep those mobile apps in a way that readers can take several mobile apps in a specific topic without hassle.

III. EVALUATION OF SYSTEMS

A. Existing System

In our college exists only the manual way of asking the queries to the appropriate staffs which will be an inconvenient way for students since they could not clarify their doubts at the time they need. Retrieval-based models (easier) use a repository of predefined responses and some kind of heuristic to pick an appropriate response based on the input and context. The heuristic could be as simple as a rule-based expression match, or as complex as an ensemble of Machine Learning classifiers. These systems don't generate any new text, they just pick a response from a fixed set. Retrieval-based methods don't make grammatical mistakes. However, they may be unable to handle unseen cases for which no appropriate predefined response exists. For the same reasons, these models can't refer back to contextual entity information like names mentioned earlier in the conversation.

1) Disadvantages

- It only response for predefined keywords.
- Difficult to update staff details and provide time consuming process.

B. Proposed System

This Chatbot will automate the existing manual responding system thereby making the existing system simpler. This system will be designed in such a way that it will answer the queries based on the training dataset and also learn the new queries and answers to them. Generative models (harder) don't rely on pre-defined responses. They generate new responses from scratch. Generative models are typically based on Machine Translation techniques, but instead of identify the synthetic similarity for entered Keyword.

2) Advantages

- It provide the results based on the labeled and unlabeled data.
- It reduce the manual work
- It take less time complexity

IV. DESCRIPTION

There are several modules are used in this project.

A. Server Interface

In this module the admin can add the staff and events information to server. The server will be create on mango db. Server GUI is created using Python coding. Server interface has various functions such add, delete and update.

B. User Interface

In this module, we can create the interface for parent. GUI is created using Python. Android application is used to view the details about staff details and events details.

C. Search Query

A search query is a query that a student enters into a chatbot to satisfy his or her information needs. Web search queries are distinctive in that they are often plain text or hypertext with optional search-directives (such as "and"/"or" with "-" to exclude). They vary greatly from standard query languages, which are governed by strict syntax rules as command languages with keyword or positional parameters.

A search query, the actual word or string of words that a search engine user types into the search box, is the real-world application of a keyword – it may be misspelled, out of order or have other words tacked on to it, or conversely it might be identical to the keyword.

D. Similarity Prediction

In this module is used to , we have proposed the prototype of Chatbot, together with Synthetic similarity graph query matching with an existing queries. This algorithm used to improve the search results. So we create the repository for quick access. Implement bag of terms concept using Synthetic Similarity approach to extract the relevant and exact results for query terms.

E. Optimal Results

In this module we provide the results based on Student search. Text Processing is done using NLP. Then, Acquired keywords are matched against the knowledge base to retrieve the appropriate response.

V. SOFTWARE DESCRIPTION

A. Python

Python is an interpreted high-level programming language for general-purpose programming. Created by Guido van Rossum and first released in 1991, Python has a design philosophy that emphasizes code readability, and a syntax that allows programmers to express concepts in fewer lines of code, notably using significant whitespace. It provides constructs that enable clear programming on both small and large scales. Python features a dynamic type system and automatic memory management. It supports multiple programming paradigms, including object-oriented, imperative, functional and procedural, and has a large and comprehensive standard library. Python interpreters are available for many operating systems. CPython, the reference implementation of Python, is open source software and has a community-based development model, as do nearly all of its variant implementations. CPython is managed by the non-profit Python Software Foundation.

Python is a multi-paradigm programming language. Object-oriented programming and structured programming are fully supported, and many of its features support functional programming and aspect-oriented programming (including by meta programming and meta objects (magic methods)). Many other paradigms are supported via extensions, including design by contract and logic programming. Python uses dynamic typing, and a combination of reference counting and a cycle-detecting garbage collector for memory management. It also features dynamic name resolution (late binding), which binds method and variable names during program execution.

B. MongoDB

MongoDB is a free and open-source cross-platform document-oriented database program. Classified as a NoSQL database program, MongoDB uses JSON-like documents with schemas. MongoDB is developed by MongoDB Inc., and is published under a combination of the GNU Affero General Public License and the Apache License.

VI. CONCLUSION

The proposed system would be a stepping stone in having in place an intelligent query handling program. An intelligent question answering system has been developed using the Naïve Bayesian concept. The system is capable of answering the query of the student in an interactive way using the chat agent that is used. Although there is still scope for improvement, the system performs fairly well in identifying syntactically similar question and to a certain extent semantics is also considered. Also because we make use of a filtering process the search space is reduced and so the system becomes more efficient algorithmically.

REFERENCES

[1] Adrian Horzyk, Stanis law Magierski, and Grzegorz Miklaszewski “An Intelligent Internet Shop-Assistant Recognizing a Customer Personality for Improving Man-Machine Interactions” in Recent Advances in Intelligent Information Systems. ISBN 978-83-60434-59-8, pages 13–26

[2] Cai, C. H., Fu, A. W., Cheng, C. H. and Kwong, W. W. “Mining Association Rules with Weighted Items.” in Proceedings of International Database Engineering and Applications Symposium,

[3] Salto Martinez Rodrigo "Development and Implementation of a Chat Bot in a Social Network" Information Technology: New Generations (ITNG), 2012 Ninth International Conference on 16-18 April 2012
[4] S. J. du Preez, M. Lall, S. Sinha "An intelligent web-based voice chat bot" EUROCON 2009, EUROCON '09. IEEE Date of Conference: 18-23 May 2009
[5] s. J. Du preez1, student member, ieee, m. Lall2, s. Sinha3, mieee, msaiee "an intelligent web-based voice chat bot" enterprise application development, tshwane university of technology (tut), staatsartillerie road, pretoria west, 0001, south africa
[6] Ramachandra. V. Pujeri1, G.M. Karthik " Constraint based frequent pattern mining for generalized query templates from web log" 1KGiSL Institute of Technology, Coimbatore, Tamil Nadu, INDIA