# AI Based Bot Virtual Guide

**Ms.Nitisha Tungar[1], Ms.Nutan Avhad[2], Ms.Pranoti Gayakhe[3] , Ms.Rutuja Musmade[4] ,**

**Mr.U.R. Patole[5]**

[1]*Computer Engineering, SVIT, Nashik,*
[2]*Computer Engineering, SVIT, Nashik,*
[3]*Computer Engineering, SVIT, Nashik,*
[4]*Computer Engineering, SVIT, Nashik,*
[5]*Assistant Professor, Department of Computer Engineering, SVIT, Chincholi, Nashik, Maharashtra, India*

---

**Abstract** - This paper proposes a general solution for the School timetabling problem. As all staff is busy and the end time lecture conduct is severe problem for college. So to automatic Virtual Guide is been implemented which will extract web content based on recent topic been teached. An enormous amount of learning material is needed for the elearning content management system to be effective. This has led to the difficulty of locating suitable learning materials for a particular learning topic, creating the need for automatic exploration of good content within the learning context. We aim to tackle this need by proposing a novel approach to find out good materials from www for eLearning content management system. This work presents domain ontology concepts based query method for searching documents from web and proposes concept and term based ranking system for obtaining the ranked seed documents which is then used by a concept-focused crawling system. The set of crawled documents so obtained would be obtained an appropriate set of content material for building an e-learning content management system. The filtered data crawled will be provided with speech output.

***Key Words***: DOM Parser, Web Crawler, text to speech, speech to text.

## 1.INTRODUCTION

This work proposes that Information Retrieval (IR) techniques and technologies could be specifically designed to traverse the WWW and centrally collect educational resources, categorized by topic area. IR systems are generally concerned with receiving a users information need in textual form and finding relevant documents which satisfy that need from a specific collection of documents [3]. Most existing content retrieval techniques rely on indexing keywords. Unfortunately, keywords or index terms alone cannot adequately capture the document contents, resulting in poor retrieval performance [7]. Typically, the information need is expressed as a combination of keywords and a set of constraints. However, here we use learning terms associated with topic under consideration extracted from the domain ontology. These topics and learning terms are used in the concept based query method. In addition, this work proposes a concept and term based ranking system for ordering the documents from search engine to obtain a ranked list of seed documents. With the appearance of sophisticated search engines, finding materials for e- learning is not a problem. However, the resources that one discovers might have varying styles and may be targeted at different type of audiences. The resources may not have a complete coverage of topics which the instructor actually requires for content authoring. Moreover, a number of resources which are retrieved are highly redundant [4]. Hence, appropriate ranking of documents using concept and topic learning terms possibly will help in retrieving topic related documents and reducing redundancy from retrieved content. In this work, the ranking system exploits the concept-document similarity of the document collection. These ranked documents could then be used as seed documents for our proposed crawling system.

Similar to the work described earlier, the proposed system also used the concepts of the ontology to query the web to obtain seed documents. The ontology used by us is however specially designed a compute science ontology based on the ACM classification hierarchy. The association of terms to concepts for specific purposes has been used by Info Web a filtering system using user profiles in a digital library scenario. Here the semantic network used to represent the user profile has nodes representing concepts and as more information Is gathered about the user the profile is enhanced by associating additional weighted keywords with these concept nodes. This idea has been used in the work described in this paper where in the ontology, each node in addition to having concepts from ACM classification, has an associated set of topic learning terms typically used when teaching this topic. At present this set of associated topic learning terms is manually obtained from typical texts covering the topic. As a future enhancement we propose to enhance this ontology through machine learning techniques. The search using concepts and topic learning terms from the ontology retrieves a set of seed documents.

The pace of growth of the world-wide body of available information in digital format (text and audiovisual) constitute a permanent challenge for content retrieval technologies [1]. The popularity of exchange and dissemination of content through the web has created a huge amount of educational resources and the challenge of locating suitable learning references specific to a learning topic has become a big challenge [2]. As the web grows it will become increasingly difficult for educators to discover and aggregate collections of relevant and useful educational content. There is, as yet, no centralized method of discovering, aggregating and utilizing educational content [3]. This work proposes that Information Retrieval (IR) techniques and technologies could be specifically designed to traverse the WWW and centrally collect educational resources, categorized by topic area. IR systems are generally concerned with receiving a users information need in textual form and finding relevant documents which satisfy that need from a specific collection of documents [3]. Most existing content retrieval techniques rely on indexing keywords. Unfortunately, keywords or index terms alone cannot adequately capture the document contents, resulting in poor retrieval performance [7]. Typically, the information need is expressed as a combination of keywords and a set of constraints. However, here we use learning terms associated with topic under consideration extracted from the domain ontology. These topics and learning terms are used in the concept based query method. In addition, this work proposes a concept and term based ranking system for ordering the documents from search engine to obtain a ranked list of seed documents. With the appearance of sophisticated search engines, finding materials for e- learning is not a problem. However, the resources that one discovers might have varying styles and may be targeted at different type of audiences. The resources may not have a complete coverage of topics which the instructor actually requires for content authoring. Moreover, a number of resources which are retrieved are highly redundant [4]. Hence, appropriate ranking of documents using concept and topic learning terms possibly will help in retrieving topic related documents and reducing redundancy from retrieved content. In this work, the ranking system exploits the concept-document similarity of the document collection. These ranked documents could then be used as seed documents for our proposed crawling system.

## 2.SYSTEM ARCHITECTURE:

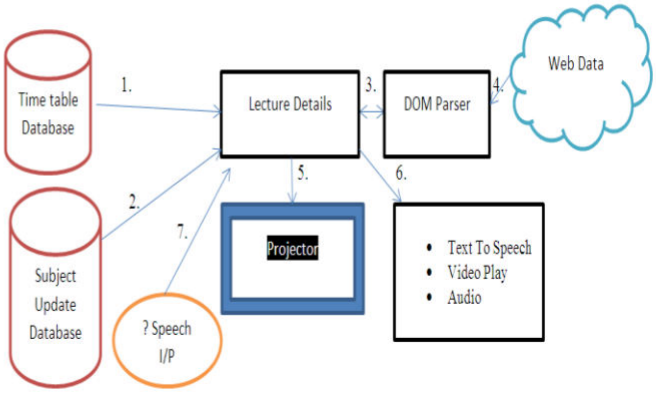


**Fig.1.System Architecture**

1. **Web Parsing:** The Document Object Model Parser interface provides a ability to parse XML or HTML source code from a string into the DOM Document. DOM parser is intended for working with XML as an object graph (a tree like structure) in memory so called Document Object Model (DOM). Firstly, the parser traverses the input XML file and creates DOM objects corresponding to the nodes in XML file. These DOM objects are linked with each other in a tree like structure. Once the parser is done with parsing process, we get a tree-like DOM object structure back from it. Now we can traverse the structure of DOM back and forth as we want to get/update/delete data from it.
2. **Text To Speech:** Text-to-Speech (TTS) encoder decoder architectures. These auto encoders learn the features from speech only and text only a datasets by switching the encoders and decoders used in ASR and TTS models.
3. **Pattern Mining:** This pattern step is designed to handle set-typed data, where multiple values occur, thus a naive approach is to discover repetitive patterns in the input. However, there can be many repetitive patterns discovered and the pattern can be embedded in the form of another pattern, which makes the deduction of the template difficult. The good news is that we can neglect the effect of missing attributes (optional data) since they are handled in the previous step. Thus, we should focus on how repetitive patterns are merged to deduce the data structure. In this section, we detect every consecutive repetitive pattern (tandem repeat) and merge them (by deleting all occurrences except for the first one) from small length to large length.
4. **DOM Parser:** According to our page generation model, data instances of the same type have the same path from the root in the DOM trees of the input pages. Thus, our algorithm does not need to merge similar subtrees from different levels and the task to merge multiple trees can be broken down from a tree level to a string level. Starting from root nodes ¡html¿ of all input DOM trees, which belong to some

type constructor we want to discover, our algorithm applies a new multiple string alignment algorithm to their first-level child nodes. There are at least two advantages in this design. First, as the number of child nodes under a parent node is much smaller than the number of nodes in the whole DOM tree or the number of HTML tags in a Webpage, thus, the effort for multiple string alignment here is less than that of two complete page alignments in RoadRunner. Second, nodes with the same tag name (but with different functions) can be better differentiated by the subtrees they represent, which is an important feature. Instead, our algorithm will recognize such nodes as peer nodes and denote the same symbol for those child nodes to facilitate the following string alignment. After the string alignment step, we conduct pattern mining on the aligned string S to discover all possible repeats (set type data) from length 1 to length jSj=2. After removing extra occurrences of the discovered pattern, we can then decide whether data are an option or not based on their occurrence vector. The four steps, peer node recognition, string alignment, pattern mining, and optional node detection, involve typical ideas that are used in current research on Web data extraction. However, they are redesigned or applied in a different sequence and scenario to solve key issues in page-level data extraction.

## 3.RESULT:

### 1. Frontpage of Bot Virtual Guide:

This is the front page of our system. Whenever we will start the Bot Virtual Guide system this page appears.


**Fig.2.Frontpage of Bot Virtual Guide**

### 2. Login Page:

This is the login page. It is created for authentication purpose, means to provide access of the system to only authorised users only. The authorised user needs to use the user id and password provided to login with the system.


**Fig.3.Login Page**

### 3. Menu Framework:

This page shows the menu provided for the teachers to add the subject, timetable, subject update info, links extract etc. Here the add subject tab is given for the teacher to add the subjects for the respective year and the add timetable tab is for adding the timetable for the respective years. Whereas the subject update info is for the teachers to update the last taught topic in the class.

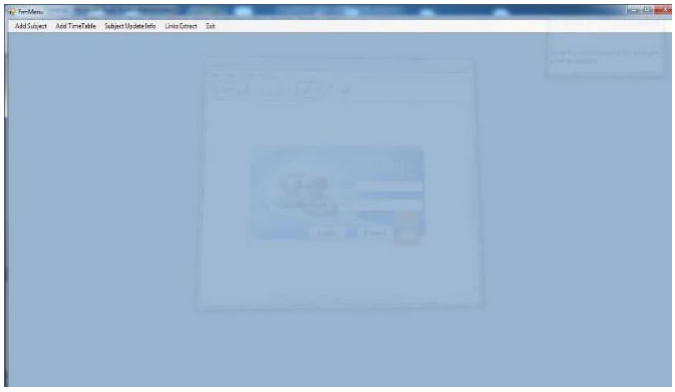

**Fig.4.Menu Framework**

### 4. Add Subject:

In Add Subject the teacher has to add the subject id, year, branch, subject name, faculty name. It will contain the data about the subjects which are to be taught to the respective years. This information is to be filled by the faculty members.


**Fig.5.Add Subject**

### 5. Add Timetable:

In Add Timetable the faculty member needs to enter the timetable of the respective years. The timetable contains the weekly lecture schedule of all years. The timetable contains the days of the week, lecture of the various subjects to be conducted during the whole day.


**Fig.6.Add Timetable**

### 6. Subject Information :

Subject info contains the information of the particular subject. It keeps track of the topics taught by the teacher in every lecture taken. The subject info should be updated everyday by the faculty members so that whenever the faculty member is not available to take the lecture the system is aware of the last topic taught by the faculty.

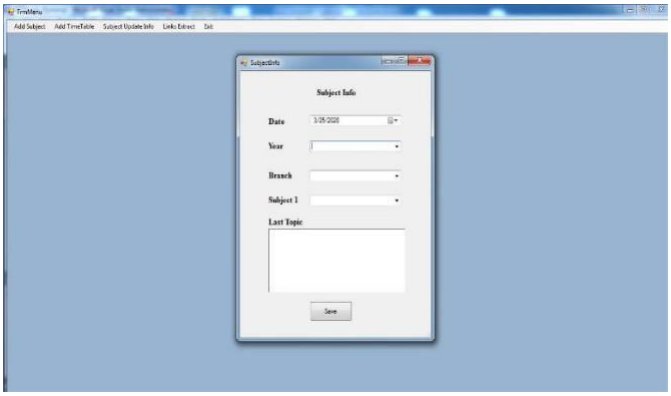


**Fig.7.Subject Information**

### 7. Web Data Extractor:

This image shows the page where once we have entered the topic for search the system gives the relevant information regarding that topic entered in the search query area.Here get last chapter button is also provided which helps to get the last topic taught by the faculty and hence from that the system selects the next topic from the timetable which is to be taught and the system shows relevant information regarding that topic on the screen.

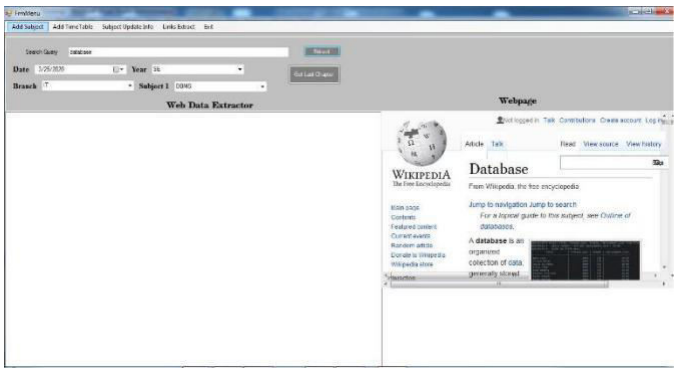


**Fig.8.Web Data Extractor**

### 8. Student and Staff Database:

This is the backend of the system i.e database which stores all the information of the system such as the login credentials of the authorised user so that it gives access to only authorised users. It also stores the information of the subjects of the respective years which is entered by the faculty and also contains the information of the timetable details of all the years.

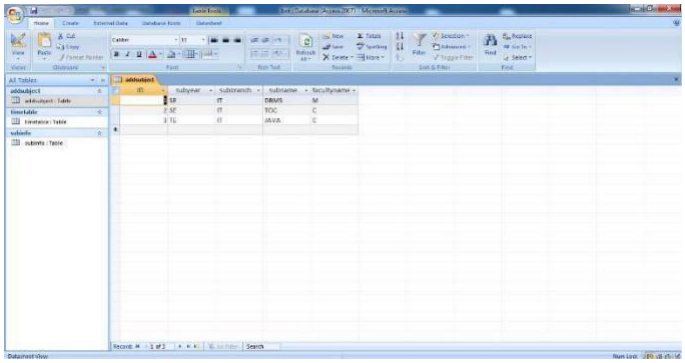

**Fig.9.Student and Staff Database**

## 4. CONCLUSIONS

We have presented BOT Virtual Guide which will obtaining seed documents from search engine and presented a concept-focused crawling system for the discovery of educational content from the web. Also the web being a rich repository of learning content, we attempt to collect high volume of learning material from web using web miner. In this system we have presented concept based on ranking system for obtaining seed documents. Also this system enables speech recognition means making the computer understand what the student speak to solve their problems.

### FUTURE SCOPE:

In Future We will provide an Android application for the same working.

## ACKNOWLEDGEMENT

We are thankful to Dr. Y. R. Kharde,Principal,SVIT for providing useful resources for the completion of this project work & We are also thankful to our project guide Mr. U.R.Patole for the guidance and constructive suggestions that were helpful to us in the preparation of this project.We are also thankful to the all staff members of Computer Engineering Department,SVIT ,Nashik.

## REFERENCES

1. Nitisha Tungar, Nutan Avhad, Pranoti Gayakhe, Rutuja Musmade, Mr.U.R.Patole "Bot Virtual Guide" , International Research Journal of Engineering and Technology (IRJET),2019.
2. Nitisha Tungar, Nutan Avhad, Pranoti Gayakhe, Rutuja Musmade, Mr.U.R.Patole "Bot Virtual Guide" , International Journal of Scientific Research in Engineering and Management (IJSREM),2020.
3. Chakrabarti, S., Punera, K., Subramanyam, M. Accelerated Focused Crawling through Online Relevance Feedback. In proceedings of the Eleventh International World Wide Web Conference, WWW2002, Honolulu, Hawaii, USA. May 7-11, 2002.
4. Lawless, S."Leveraging Content from O p e n Corpus Sources for Technology Enhanced Learning", Ph.D Thesis, Submitted to the University of Dublin, Trinity College, 2009.

## BIOGRAPHIES

**Name:** Nitisha R. Tungar
**Educational Details:**
BE Computer (Pursuing)

**Name:** Nutan V. Avhad
**Educational Details:**
BE Computer (Pursuing)

**Name:** Pranoti P. Gayakhe
**Educational Details:**
BE Computer (Pursuing)

**Name:** Rutuja V. Musmade
**Educational Details:**
BE Computer (Pursuing)

**Name:** Uttam R. Patole
**Educational Details:**
M.Tech(CSE)