BOT Virtual Guide

Nitisha Tungar[1], Nutan Avhad[2], Pranoti Gayakhe[3], Rutuja Musmade[4], Mr. U.R. Patole[5]

[1,2,3,4]BE Computer, [5]Assistant Professor, [1,2,3,4,5] Computer Department, 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 taught. An enormous amount of learning material is needed for a e-learning content management system to be effective. This has led to the difficulty of locating suitable learning materials for the 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 world wide web for an e-Learning 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 hence obtained and 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 user's 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 a 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. RELATED WORK

2.1 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.

2.2 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.

2.3 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.

World situations. It requires a small amount of training data to estimate the parameters.

2.4 Speech Recognition:

Speech recognition is a interdisciplinary subfield of computational linguistics that develops methodologies and technologies that enables the recognition and translation of spoken language into text by computers. It can also known as automatic speech recognition (ASR), computer speech recognition or speech to text (STT). It incorporates knowledge and research in a linguistics, computer science, and electrical engineering fields. Some speech recognition systems require "training" (also called as"enrollment") where an individual speaker reads text or isolated vocabulary into a system. The system analyzes the person's specific voice and uses it to fine-tune the recognition of that person's speech, resulting in increased accuracy. Systems that do not use training are called "speaker independent" systems. Systems that uses training are called as "speaker dependent".

3. LITERATURE SURVEY

The pace of growth of the world-wide body of available information in digital format (text and audiovisual) constitute a permanent challenges for content retrieval technologies [1]. The popularity of exchange and dissemination of content through a web has created a huge amount of educational resources and a 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 user's 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 a 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 by [11], 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 computer science ontology based on the ACM classification hierarchy. The association of terms to concepts for specific purposes has been used by InfoWeb [13] 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 a 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.

4. PROPOSED SYSTEM

To provide virtual guide using web crawler BOT. 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.

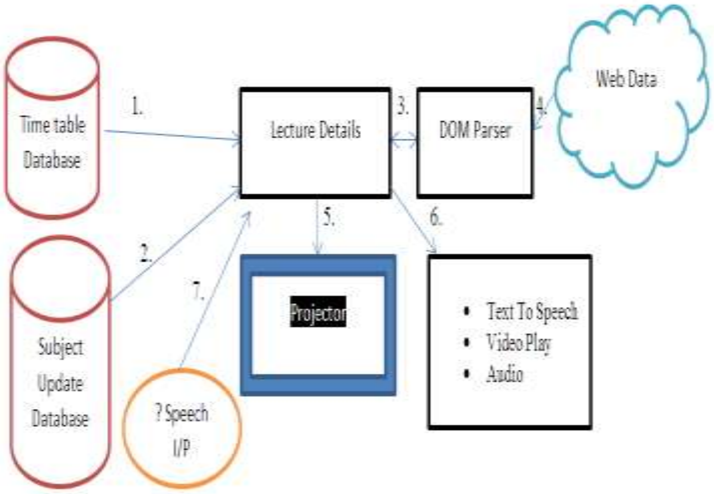

Fig 1: Block Diagram

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.

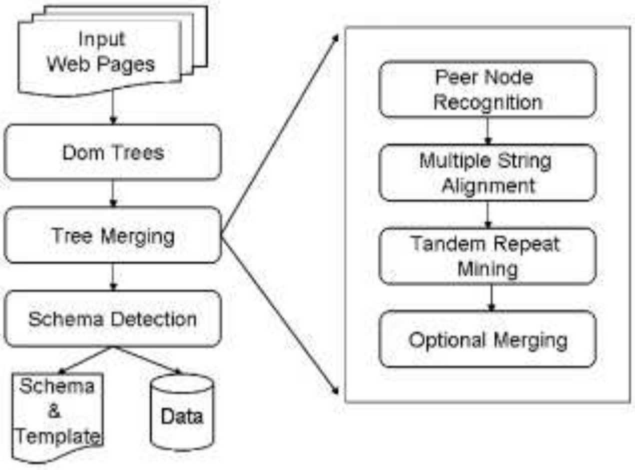

Fig 2. DOM Parser Working

3. 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.

REFERENCES

1. David Vallet, Pablo Castells, Miriam Fernández, Phivos Mylonas, and Yannis Avrithis, “Personalized Content Retrieval in Context Using Ontological Knowledge”, Circuits and Systems for Video Technology, IEEE Transactions on In Circuits and Systems for Video Technology, IEEE Transactions on, Vol. 17, No. 3. (05 March 2007), pp. 336-346.

2. Khairil Imran Bin Ghauth, Nor Aniza Abdullah, “Building an E-Learning Recommender System using Vector Space Model and Good Learners Average Rating”, Advanced Learning Technologies, 2009. ICALT 2009 Ninth IEEE International Conference (15-17 July 2009), pp 194 - 196

3. Lawless, S."Leveraging Content from Open Corpus Sources for Technology Enhanced Learning", Ph.D Thesis, Submitted to the University of Dublin, Trinity College, 2009.

4. Brusilovsky, P. & Henze, N. “Open Corpus Adaptive Educational Hypermedia”. In The Adaptive Web: Methods and Strategies of Web Personalisation, Lecture Notes in Computer Science, vol. 4321, Berlin: Springer Verlag, pp. 671-696. 2007.

5. 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.

6. Sparck-Jones, K (1972).”A statistical interpretation of term specificity and its application in retrieval”, Journal of Documentation 2004, Volume 60 Number 5 pp. 493-502

7. DIK L. LEE, “Document Ranking and the Vector-Space Model”, Software, IEEE (Mar/Apr 1997) Volume: 14 Issue: 2 pp 67 – 75.

8. Mehrnoush Shamsfard, Azadeh Nematzadeh, and Sarah Motiee, “ORank: An Ontology Based System for Ranking Documents”, International Journal of Computer Science, Vol. 1, No. 3. (2006), pp. 225-231.

9. Udit Sajjanhar, “Focused Web Crawling for E-Learning Content”, M.Tech Thesis to be submitted Indian Institute of Technology Kharagpur , April 2008

10. Jun Li Kazutaka Furuse Kazunori Yamaguchi, “Focused Crawling by Exploiting Anchor Text Using Decision Tree”, Proceeding WWW '05 Special interest tracks and posters of the 14th international conference on World Wide Web ACM, pp.1190-1191

11. Hiep Phuc Luong Susan Gauch Qiang Wang, “Ontology-based Focused crawling”, International Conference on Information, Process, and Knowledge Management, Cancun, Mexico, Feb. 1-7, 2009, pp123-128.

12. Marc Ehrig Alexander Maedche, “Ontology- Focused Crawling of Web Documen`ts”, Proceeding SAC '03 Proceedings of the 2003 ACM symposium on Applied computing, ACM

13. Gentili, G., Micarelli, A., Sciarrone, F.: Infoweb: “An Adaptive Information Filtering System forthe Cultural Heritage Domain”. Applied Artificial Intelligence 17(8-9) (2003) 715-744

14. Chakrabarti, S., van den Berg, M., Dom, B. “Focused Crawling: A New Approach to Topic- Specific Web Resource Discovery”. In The International Journal of Computer and Telecommunications Networking, Vol.31(11-16), Elsevier North-Holland, Inc. New York, USA. pp. 1623-1640, May 1999.

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)