Touch with Industry

Priyanka Thakur [1], Karuna Sharma [2], Priyanka Chaudhari [3], Sayali Borade [4]

[1,2,3,4] *Student, Computer Department, MET's BKC IOE Nashik, Maharashtra, India*

---***---

Abstract - *Web content mining is the extraction and integration of relevant data, information and knowledge from different forms of web page contents. Extracting data from web pages is challenging task as web data is mainly in semi-structured or unstructured format, while web content mining deals primarily with structured data. We have proposed a system with title "Touch with Industry" that mainly focuses on gathering information from various trusted sites on the basis of company names and fields selected by the users. The provided fields not limited to Company CEO, Address, Contact No, Year of Establishment, Employee Strength, Images of Company, and People on LinkedIn.*

The proposed system uses web crawling, which makes it easier for the application to return the most relevant results to users. The proposed system is advantageous to people like consultants, HR representatives, Employed and Unemployed peoples who need company overview in short period of time, which reduces the searching time and provide relevant information from different web pages in one go.

Key Words: *web mining, web crawling, pattern matchin, filtering.*

1.INTRODUCTION

The World Wide Web has large online databases of the companies. This database needs to be accessed by millions of people. According to recent survey, it has been found that some websites provide fake information to people. Also, it has been observed that people needs to search a lot for relevant information which consumes large amount of time

1.1 Project Idea

The idea behind implementing project is to provide solution to problem of people which they face while they are searching for company information. This system is mainly proposed to promote efficient communication between job seekers and consultants which will lead to ideal and effective system.

1.2 Motivation

Basically the motivation came with the survey among people who faced problem while searching company details. Users need to crawl through a number of pages to get desired information about a company. To overcome this problem the proposed system will be develop a desktop application through which users can easily search about a company in less amount of time and get relevant details.

2. LITERATURE SURVEY

The exhaustive literature survey consisting of the conceptual base for this project is briefly outlined here.

1. Web Data Extraction: Extracting structured data from deep Web pages is a challenging problem due to the underlying intricate structures of such pages. This motivates us to seek a different way for deep Web data extraction to overcome the limitations of previous works by utilizing some interesting common visual features on the deep Web pages. In this paper, a novel vision-based approach that is Web- pageprogramming- language-independent is proposed. This approach primarily utilizes the visual features on the deep Web pages to implement deep Web data extraction, including data record extraction and data item extraction.[1]

2. Data Restruction: It synthesizes ideas from query languages for the Web, for semi structured data and for website restructuring and makes several contributions, most notably, the idea of querying documents by manipulating their abstract syntax trees and the support of the concept of web as a data type. [2]

3. Hoovers: It is an official website to search the company's profile. It searches the world's largest database of company and industry information. It provides less detail company information to non-subscribers. It maintains a database of about 85 million companies and 100 million peoples.[3]

4. Web Information Extraction Systems: The Internet presents a huge amount of useful information which is usually formatted for its users, which makes it difficult to extract relevant data from various sources. This paper surveys the major Web data extraction approaches and compares them in three dimensions: the task domain, the automation degree, and the techniques used. [4]

2.1.1 Existing Systems

The following mentioned applications are available for the various feature on Company Information Extraction but having the limitations with some regards to satisfy the user.

1. Company Profile Search
_ Organised Profile
_ Detailed Search
_ Profit Accomplished

2. Lead Finder Jack
_ Crawls onto Google
_ Searches companies having paid advertisements
_ Low accuracy

3. PROBLEM DEFINATION AND SCOPE

Suppose a user wants to search specific information about the company. For this purpose one makes use of search engine which indeed displays a lot of websites which may be irrelevant to the user and this is a tedious and time consuming task. To overcome the problem specified above we develop a desktop based application that provides the solution to the issues related with existing company information finding system. Using trusted sites to retrieve relevant information related to companies and providing accurate data to the user.

3.1 Goals and objectives

i) Goals: To eliminate the burden of searching company information. To provide user friendly and efficient service.
ii)Objective: The main objective of proposed system is user comfort. The users can easily access the company details and get the desired information at a glance.

3.2 Statement of scope

The scope of proposed system is justifiable because in wide range people search for industry details. So wide range of people can make use of proposed system.

3.3 Outcome

The outcome of the proposed system is,

i)Accuracy of information

ii)Selection of features according to users choice

iii)Efficient searching

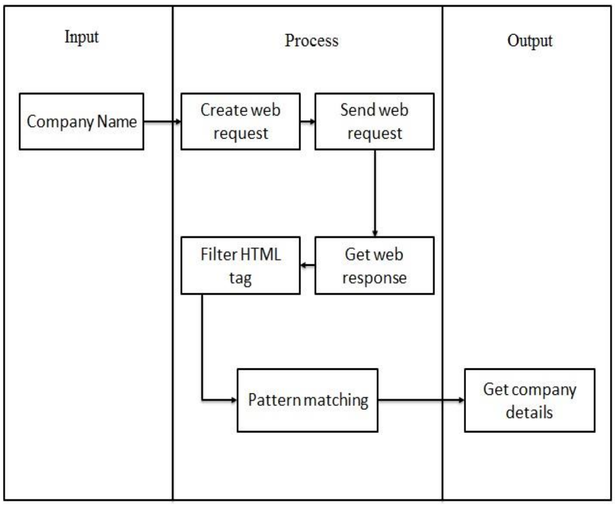
4. DETAILED DESIGN

4.1 Introduction:

System architecture will benefit in such a way that every user of the system gets advantage. When a user searches about a company he needs to register into the system. After logging into the system user has to enter the company name and select the filters as per his requirement. The system will return the required information by applying data extraction techniques.

4.2 Architecture Design:

The input to the system is the company name and the feature of the company. The proposed system will then create a web request and send it to the search engine. The search engine then replies to the system, the reply contains all the information of the company. This system supports GET and SET protocol method for the web request. SET method creates a web request and GET method sends the request to the system. After, receiving the information the proposed system filter the HTML tags i.e. filters the irrelevant information based on the feature entered as input. Further pattern matching is carried out in which the input pattern (features of input query) is matched with the responded received by the system after filtering. After the pattern matching, the information whose pattern matches with the input query are displayed to the user i.e. user receives the relevant information as output. The proposed system will make use of only trusted sites, so only the correct information will be received by the tool and display as output.



4.3 Algorithm

Extracting structured data from web pages :

On entering the company name the process of the system will get initiated. company_name is the variable taken for company name. The system sends the web request to the websites with company_name. The system receives the HTML response from the websites. Then we apply regular expressions and extract text node and data node. Pattern matching is performed and illegal characters are removed. After extracting relevant data related company links will be displayed to the user. The user has to enter the correct company name . After that system again sends web request with CIN no and get HTML response. The variable used for CIN no is cin_no. Then html tags are removed by using regular expressions and replacing it by newline \n. Split lines by '\n'. For each line do (find heading nodes(attribute names) at position i attribute[i].value= lines[i+1]). Then the system save attribute value and show data.

5. CONCLUSIONS AND FUTURE WORK

We have presented Touch With Industry system, which provides a framework for approaching many web data management task from a unified perspective. This system operates in three levels: Input, process and output. The proposed system provides the user with the relevant information by filtering the HTML tags. The system also follows pattern matching process after filtering HTML tags which allows the system to match the response with the input parameters given by user. The Touch With Industry system can be extended to other domain of application. Additionally, as future work we can notify the logged in users about the jobs vacancy available in various companies. We can also provide the facility to the user for uploading their resume for getting job in their area of interest.

REFERENCES

[1] ViDE: A Vision-Based Approach for Deep Web Data Extraction by Wei Liu, Xiaofeng Meng.

[2] WebOQL: Restructuring Documents, Databases and Webs by Gustavo O. Arocena, Alberto O. Mendelzon

[3] S. Huffman, "Learning Information Extraction Patterns from Examples," Connectionist, Statistical, and Symbolic Approaches to Learning for Natural Language Processing, Springer-Verlag, 1996.

[4] A Survey of Web Information Extraction Systems by Chia-Hui Chang, Mohammed Kayed and Khaled F. Shaalan

[5] Using Visual Clues Concept for Extracting Main Data from Deep Web Pages by Satish Pusdekar , Shaikh Chhaware

[6] A Survey of Web Information Extraction Systems by Chia-Hui Chang , Mohammed Kayed

[7]D. Freitag, "Information Extraction from HTML: Application of a General Learning Approach," Proc. 15th Conf. Artificial Intelligence (AAAI '98), 1998.

[8] J. Wang and F.H. Lochovsky, "Data Extraction and Label Assignment for Web Databases," Proc. 12th Int'l Conf. World Wide Web (WWW), pp. 187-196, 2003.

[9] A. Arasu and H. Garcia-Molina, "Extracting Structured Data from Web Pages," Proc. ACM SIGMOD Int'l Conf. Management of Data, pp. 337-348, 2003.

BIOGRAPHIES

Priyanka S. Thakur appearing for BE degree from the Department of Computer Engineering, MET's Bhujbal Knowledge City IOE, Nashik.

Karuna R. Sharma appearing for BE degree from the Department of Computer Engineering, MET's Bhujbal Knowledge City IOE, Nashik.

Priyanka B. Chaudhari appearing for BE degree from the Department of Computer Engineering, MET's Bhujbal Knowledge City IOE, Nashik.

Sayali P. Borade appearing for BE degree from the Department of Computer Engineering, MET's Bhujbal Knowledge City IOE, Nashik.