Real Time Monitoring of Agricultural Schemes using Web Mining

Vrushali S. Tadge
SVIT chincholi,Nashik
222/7,Sector-2,charkop,Kandivali.
Tal-Borivali, Dist-Mumbai

Suvarna M. Chaudhari
SVIT chincholi,Nashik
At vadnagari, post fupnagari,
Tal-Jalgaon,Dist-Jalgaon.

Priyanka A. Ugale
SVIT chincholi,Nashik
At post chincholi,
Tal-Sinner,Dist-Nashik.

Priya S. Wagh
SVIT chincholi,Nashik
Flat No.8,Nitisha Apt.,Indira Nagar,
Tal-Nashik,dist-Nashik.

ABSTRACT
Web mining is concerned with extraction of data from deep web rather than surface web.Web contains enormous amount of data and hence searching for particular information becomes hectic .Earlier VIPER and MSE record level page extraction Systems were being used.In this paper we are going to use page level extraction system FiVaTech for mining various Government schemes and weather forecast and display it on our Designed HTML webpage which will include SMS services to the farmers. In experiments tree matching, Tree alignment and pattern mining techniques are applied to perform required search.

Keywords
Deep web, Web Mining, Page level extraction, Tree matching, tree alignment, Pattern Mining.

1. INTRODUCTION
Farmer contributes most of the income of our country. It means development of the country depends on production from agriculture. Considering this Government always tries to enhance our farmers by providing various schemes and policies. All this information is available on web. But it does not reach to the farmers.

WWW i.e. World Wide Web it is one of the biggest source of information .It stores huge amount of data everyday in which most of the data is stored in unstructured format, so extracting required data becomes quite difficult. So our system will extract data using tools such as web crawler, DOM parser and FivaTech algorithm. And system will operate and display data using technologies like .Net (ASP.net and C#.net).

As weather reports and other factors such as soil, crop information are also important from farmers point of view, it is being disseminated to the farmers through our project in the form of SMS service with the benefit of language. Their regional language will be used for SMS instead of common English language.

2. EXISTING SYSTEM
Agriculture is basic and also main occupation of Indian people. Therefore Government always launches various schemes and policies for farmers. But unfortunately this information does not reach to farmers on time. Production of crops mostly depends on weather conditions. Existing technology does not provide all these features combinly. It does not consider regional language also, which is important for communication in rural area where most people unable to understand English.

3. PROBLEM DEFINATION
Firstly, we have to keep watch on various government websites and regularly collect information from it. Also we have to design a website which will suitably display all information catched by web crawler. Along with this we have to maintain a database which will store contact information of farmers. All information including weather report should be correct and maintained so that farmers would not a get a wrong one. For language barrier we have to consider various SMS APIs which will be suitable for a system.

4. PROPOSED METHODOLOGY
The proposed system works on different modules based on different section first is tree merging and second would be schema detection of the web page.DOM trees i.e. documented object module are generated so that unstructured data available on HTML WebPages is converted into structured data forms. These Dom trees are merged by tree merging into fixed/variant pattern trees which are further used to detect the template or schema of the website .

Web crawlers are an important component to search Data. Making a web crawler work efficiently is a challenging task. There are security ,performance, speed ,efficiency and reliability issues including some social issues. Crawling is the most flexible application as

involves interacting with number of web servers . having control on it is quite difficult .Web crawling speed is the required output and it depends on various factors likewise if instead of crawling on different multiple servers if one does download in parallel the crawling or searching time will reduce .web crawlers work as it Downloads the Web page from the server and craw through all the downloaded page and simultaneously retrieve all the links and similar process goes on repeating for all the retrieved process. The Web crawler crawls through a website on the web. Starting with the given URL the Crawler follows all links found in that particular HTML page .This will find some more links, this links will also be crawled or followed , and so on. A website is nothing

but tree-structure, the root is the given start URL and all links found in that root-HTML-page are Successors of the root node and so on forming hierarchical structure .

In our project we are using DOM trees ,in which the data having same type can be followed using same path travelling from the root so that it becomes easy to find any of the node in particular DOM trees . DOM trees are used to manipulate the contents of Html pages. When pages are parsed, they are represented as a hierarchical tree structure in memory. This DOM trees represents document's elements, attributes, content, etc .This allows programmer dynamically add,remove,fire query for data in a similar way like database.

Starting from the root node <HTML> of any web page ,the DOM trees which have the data that is required the multiple string alignment is done for the successors of the root nodes also called as first level child nodes. This reduces the total number of child nodes and hence work is reduced regarding multiple string alignment .another benefit is the nodes with the same name or we can say same tag performing different functionalities are easy to identify by using their sub trees ,this nodes are said as peer nodes.

After peer node recognition we will go for pattern mining .in pattern mining step all the repetitive data that is discovered can be merged or embedded in some another pattern .here ,we will try to find the repetitive patterns i.e. repeat mining step and we will merge them by deleting all the other occurrences and just keeping the first one .

After pattern mining the schema would be detected .the schema detection and template based on page generation model and problem definition .detecting the entire structure of website will decide depending on both of identifying the schema and defining individual template for type constructor of that schema.

Experiment:

India is country of huge population and most of this population is engaged in farming .Agriculture is the main occupation and people are dependent on it to serve for their basic needs of food, clothing and shelter. For farmers various schemes are being sanctioned by the government to benefit them and provide required accessories with minimum investment and, maximum profits or to recover losses that are cause due to rain, droughts or any other climatic disaster. But unfortunately this information is not conveyed to the farmers on time so our proposed system will keep watch on various authenticated government system and help broadcasting this messages to them in their regional languages

4.1 Advantages of Proposed system

Farmers in Rural area can easily get information about various government schemes in their regional language. Along with this a information regarding weather, crops, soil is also provided on our website. A user interface is easy to handle. A database is consist of farmers contact information. If internet connectivity is having good speed then system runs very efficiently. Security provision is made so that no one can change coding other than developer.

5. SYSTEM ARCHITETURE

System architecture consists of six modules:

1) Module 1:Location Detection
2) Module 2:Web Monitoring
3) Module 3:DOM parser
4) Module 4:Pattern mining
5) Module 5:Message Service for farmers
6) Module 6:Graphical user Interface(GUI)

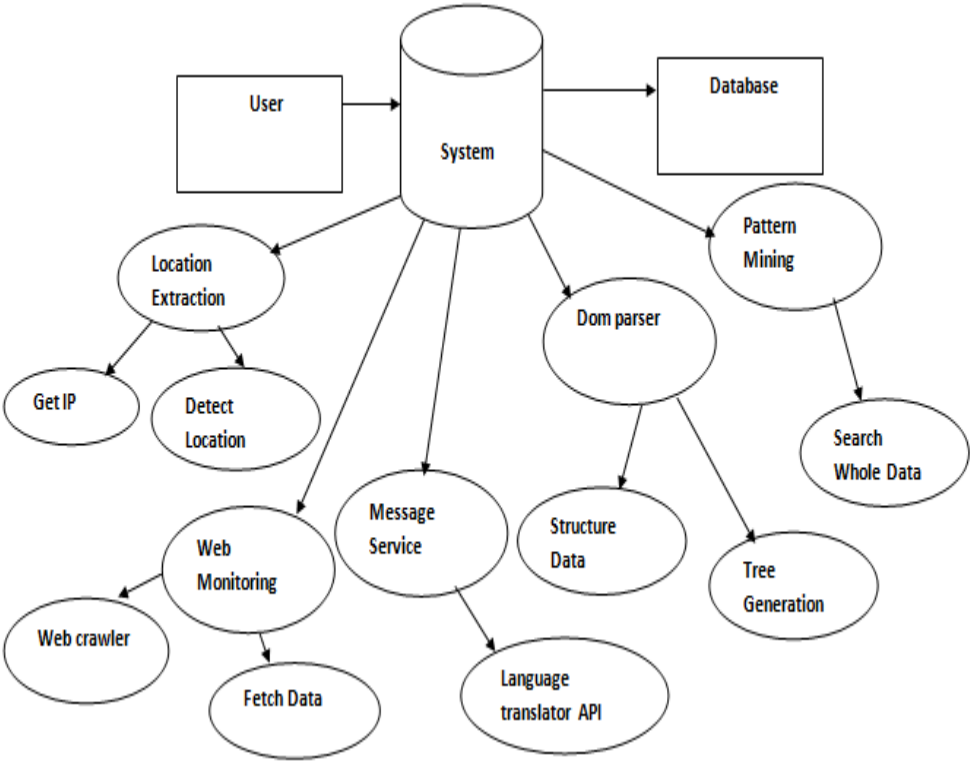

Fig 1. System Architecture.

5.1 Location detection

In this module, we are fetching IP address of particular PC from where we are accessing and getting weather reports from Yahoo service. We are making use of codes allocated for cities in Yahoo for weather report detection. According to IP address and cities entered by user weather will be generated and displayed on our web page.

5.2 Web Monitoring

To keep our farmer up to date with Governments schemes we are keeping watch on various government websites such as agricoop.in etc.Now days almost all data available on web is dynamic.In this module use of web crawler is made . Using web crawler only a required data is fetched.A *Webcrawler*is nothing but program,web crawler automatically travels the web by downloads documentsand follows links from page to page. They are mainly used by web search engines togather data for indexing. The data is refreshed by continuously firing a query after particular interval of time. It need not to be done manually ,included code will generate queries automatically and will return results.

5.3 DOM Parser

The output of web crawler obtained in previous module is given as input to the DOM parser. This module converts unstructured data into structured format through DOM parser. A tree is generated so that the extraction process will be easier. We cannot directly get exact information from unstructured deep web. So DOM parser and tree generation algorithm is used.

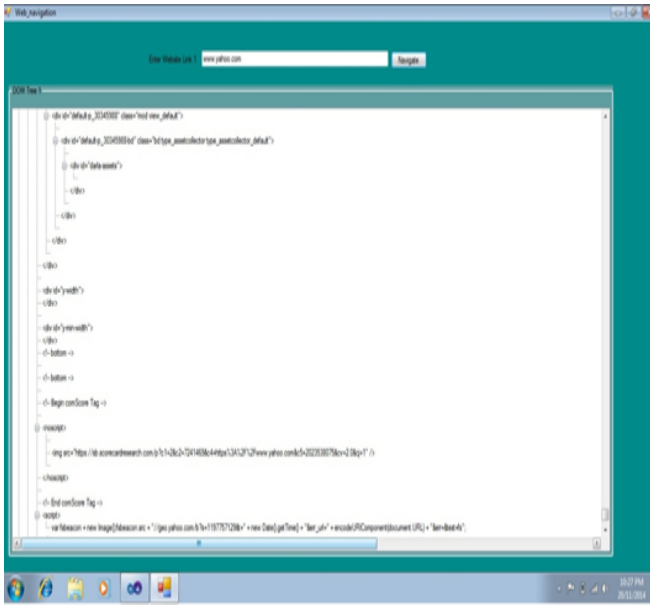

Fig 2. Screen shot of module for DOM parser

5.4 Pattern Mining

Here a fivaTech algorithm has played a major role. A structured data in tree format obtained through DOM parser and tree generation undergoes pattern mining algorithm to extract exact required data. We get only government schemes and policies from various web sites omitting unnecessary data. FivaTech algorithm has much higher efficiency than any other page level extraction algorithm.

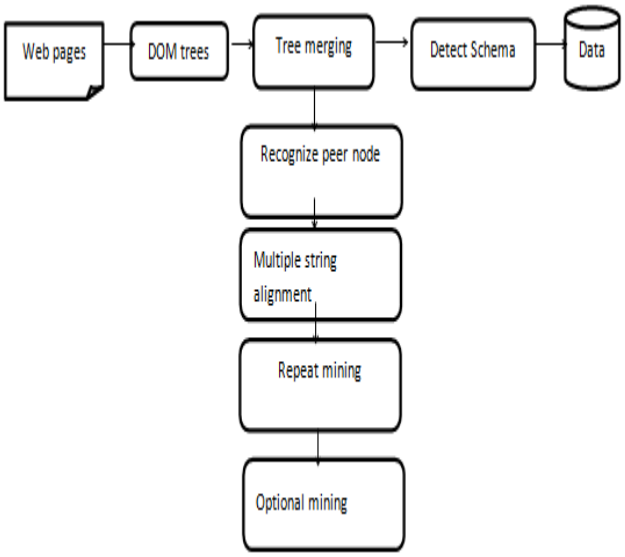

Fig 3. Execution of FivaTech algorithm

5.5 Message Service for farmers

We are providing a message service to farmers located in different cities or villages. A farmer in particular area may not understand a English or any other regional language. So we are giving message alerts to farmers in their understandable languages. In this message alert new policies and schemes with weather report are included. For that we are using various API of languages translator and SMS service provider.

5.6 Graphical User Interface

The most important part of any system is its ease of use. So we are providing a good, easy, simple and descent web user interface. Along with that a registration portal is also provided so farmers can register themselves with proper information to get benefits of the system. All the extracted information is displayed on our web page in suitable manner using ASP.net technology.

In addition to all these modules we are having service of providing information regarding crop, soil and fertilizers.

6. CONCLUSION

The proposed system is checked under various condition such as network availability, work load and database quantity. The project runs very efficiently if network provides good speed. The message service also shows correct result as per the specification. The GUI designed is easy to handle and understand. All the information is correctly displayed on web page. Hence we have executed our project successfully. We hope that this system will come out to be very helpful for all farmers on large scale also.

7. ACKNOWLEDGMENTS

We would like to thank our head of department and guide, Prof. S.M.Rokade for providing us with valuable information and courage needed for our project. We really appreciate his valuable time given for the betterment of this system. We would like to thank all the professors from our Computer Engineering Department who have helped us to improve our knowledge and for all their support. Finally we would like to

thank all our friends and our families for all their support and belief in us and for being with us during rough times.

8. REFERENCES

[1] Mohammed Kayed and ChiaHui Chang,"FiVaTech: Page-Level Web Data Extraction from Template Pages" IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, VOL. 22, NO. 2, FEBRUARY 2010

[2] A. Arasu and H. Garcia-Molina, "Extracting Structured Data from Web Pages," *Proc. ACM SIGMOD*, pp. 337-348, 2003.

[3] C.-H. Chang and S.-C.Lui, "IEPAD: Information Extraction Based on Pattern Discovery," *Proc. Int'l Conf. World Wide Web (WWW-10)*, pp. 223-231, 2001.

[4] C.-H. Chang, M. Kayed, M.R. Girgis, and K.A. Shaalan, "Survey of Web Information Extraction Systems," *IEEE Trans. Knowledge and Data Eng*., vol. 18, no. 10, pp. 1411-1428, Oct. 2006.

[5] V. Crescenzi, G. Mecca, and P. Merialdo, "Knowledge and DataEngineerings," *Proc. Int'l Conf. Very Large Databases (VLDB)*, pp. 109-118, 2001.

[6] C.-N. Hsu and M. Dung, "Generating Finite-State Transducers for Semi-Structured Data Extraction from the Web," *J. Information Systems*, vol. 23, no. 8, pp. 521-538, 1998.

[7] N. Kushmerick, D. Weld, and R. Doorenbos, "Wrapper Induction for Information Extraction," *Proc. 15th Int'l Joint Conf. Artificial Intelligence (IJCAI)*, pp. 729-735, 1997.

[8] A.H.F. Laender, B.A. Ribeiro-Neto, A.S. Silva, and J.S. Teixeira, "A Brief Survey of Web Data Extraction Tools," *SIGMOD Record*, vol. 31, no. 2, pp. 84-93, 2002.

[9] Muslea, S. Minton, and C. Knoblock, "A Hierarchical Approach to Wrapper Induction," *Proc. Third Int'l Conf. Autonomous Agents (AA '99)*, 1999.

[10] K. Simon and G. Lausen, "ViPER: Augmenting Automatic Information Extraction with Visual Perceptions," *Proc. Int'l Conf. Information and Knowledge Management (CIKM)*, 2005.

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

[12] Y. Yamada, N. Craswell, T. Nakatoh, and S. Hirokawa, "Testbed for Information Extraction from Deep Web," *Proc. Int'l Conf. WorldWide Web(WWW-13)*, pp. 346-347, 2004.

[13] W. Yang, "Identifying Syntactic Differences between Two Programs," *Software—Practice and Experience*, vol. 21, no. 7, pp. 739- 755, 1991.

[14] Zhixian Zhang, Kenny Q. Zhu, Haixun Wang, Hongsong Li "Automatic Extraction of Top-k Lists from the Web".