ISSN 22773061

A FRAME WORK FOR WEB INFORMATION EXTRACTION AND ANALYSIS

Dr Sunitha Abburu, G. Suresh Babu

Professor & Director, Department of Master of Computer Applications, Adhiyamaan College of Engineering, Hosur, Tamil Nadu, India.

drsunithaabburu@yahoo.com

JRF, Department of Master of Computer Applications, Adhiyamaan College of Engineering, Hosur, Tamil Nadu, India.

s_golla@yahoo.com

ABSTRACT

Day by day the volume of information availability in the web is growing significantly. There are several data structures for information available in the web such as structured, semi-structured and unstructured. Majority of information in the web is presented in web pages. The information presented in web pages is semi-structured. But the information required for a context are scattered in different web documents. It is difficult to analyze the large volumes of semi-structured information presented in the web pages and to make decisions based on the analysis. The current research work proposed a frame work for a system that extracts information from various sources and prepares reports based on the knowledge built from the analysis. This simplifies data extraction, data consolidation, data analysis and decision making based on the information presented in the web pages.The proposed frame work integrates web crawling, information extraction and data mining technologies for better information analysis that helps in effective decision making. It enables people and organizations to extract information from various sourses of web and to make an effective analysis on the extracted data for effective decision making. The proposed frame work is applicable for any application domain. Manufacturing,sales,tourisum,e-learning are various application to menction few.The frame work is implemetnted and tested for the effectiveness of the proposed system and the results are promising.

Indexing terms

Web Crawling, Information Extraction, Data Mining, Data Analysis.

Academic Discipline And Sub-Disciplines

Computer Science

SUBJECT CLASSIFICATION

Information Management

TYPE (METHOD/APPROACH)

A frame id proposed and implementd using a application and showed that the results are promising.

1. INTRODUCTION

The World Wide Web (WWW) is a system of interrelated hypertext documents accessed via the internet [1]. The WWW is commonly known as the web. The availability of information in the web is increasing dramatically day by day. Due to the tremendous growth of information availability, World Wide Web (WWW) has become most popular and comprehensive information resource [2]. This massive information in the web has the ability to fulfill any user information need. The web enables people and organizations to connect and access its massive store of information.

Today both individuals and organizations are mostly depending on the web for information to take many important decisions. For example: students to find valuable courses for their bright career, customers to find quality products for the best price, best services, best offers, researchers to find current research areas, current demanding challenges and organizations to analyze customer feedback, product demands, suppliers services, to take effective decisions for customer satisfaction etc. Currently the vast information available the WWW is not serving the user or the organizations for optimum usage since the information available is in different sites and getting information from multiple sources is a difficult task especially when the volume of information is huge. This raises the need for a system frame work for an efficient system that gathers the information from various sources and summarizes the information from which knowledge can be extracted which in turn helps in effective decision making.

The data available in the web is classified as structured, semi-structured and unstructured [3]. Structured information consist a predefined data structure, Semi-structured information does not have a predefined data structure and unstructured information is presented as a text document. In the web majority of the information is available in web pages. The information presented in web pages is semi-structured. It is difficult and time consuming to analyze the information presented in semi-structured due to lack of a predefined structure. The information required for analysis in a context may be scattered in different web pages. It is complex to find desired web pages, extracting information presented in various structures and analyze manually.

To overcome the above limitation the current work proposed an integrated framework. The proposed framework combines the three major technologies: web crawling, information extraction from the web and data mining for analyze the extracted information and to take decisions. The proposed system is implemented and evaluated with mobile service provider information available on the web.

Rest of the paper is organized as follows. Section 2 describes related research work, section 3 presents proposed system architecture, section 4 describes system implementation, section 5 illustrates experimental results and section 6 concludes.

2. RELATED RESEARCH WORK

Effective data analysis and decision making is a very important activity for both individuals and organizations for better prospect. To achieve the task the basic requirement is good and relevant information resource. Web is superior information medium for organizations and people to take many important decisions [4]. Basic steps to full fill the data analysis and decision making with web support are: a) finding web pages b) extracting information from web pages c) data analysis.

2.1 Finding Web Pages

Finding URLs of web pages of desired information is primary activity in web based data analysis. Web crawler is a software program that takes application domain root URL as input and crawl all the pages of the web site. It produces all the URLs of web pages of given application domain [5]. Web crawler also enable user to filter the web pages according to the content of the web pages.

2.2 Extracting Information from Web pages

In the web, web pages present information only for visualization not for data exchange [6]. To achieve data exchange in applications, web information extraction technology is introduced. Information extraction can be performed in three ways: manually, semi-automatically and automatically [7]. Manual information extraction from large number of web pages is extensively labor intensive. Many researchers and software agenesis have proposed tools to perform information extraction automatically. Programs that perform information extraction are information extractors or wrappers. Easy Web Extract [8] is one of software for extracting data from millions of web pages.

2.3 Data Analysis

Data mining is defined as finding hidden information in large volumes of databases. Data mining is used for exploratory data analysis [9]. Data mining has attracted a great deal of attention in the information science. Due to data mining powerful techniques, it is used in various commercial applications such as retail sales, e-commerce, bioinformatics etc [10]. Several algorithms for data mining techniques and data mining systems have proposed by different researchers [11]. Today data mining plays major role in data analysis, effective decision making and knowledge discovery.

ISSN 22773061

3. PROPOSED SYSTEM

The proposed system is divided into three modules for better management. The modules are a) web crawling, b) Information Extraction and c) Mining. Fig 1 shows the proposed system architecture. Following section describes each module in detail.

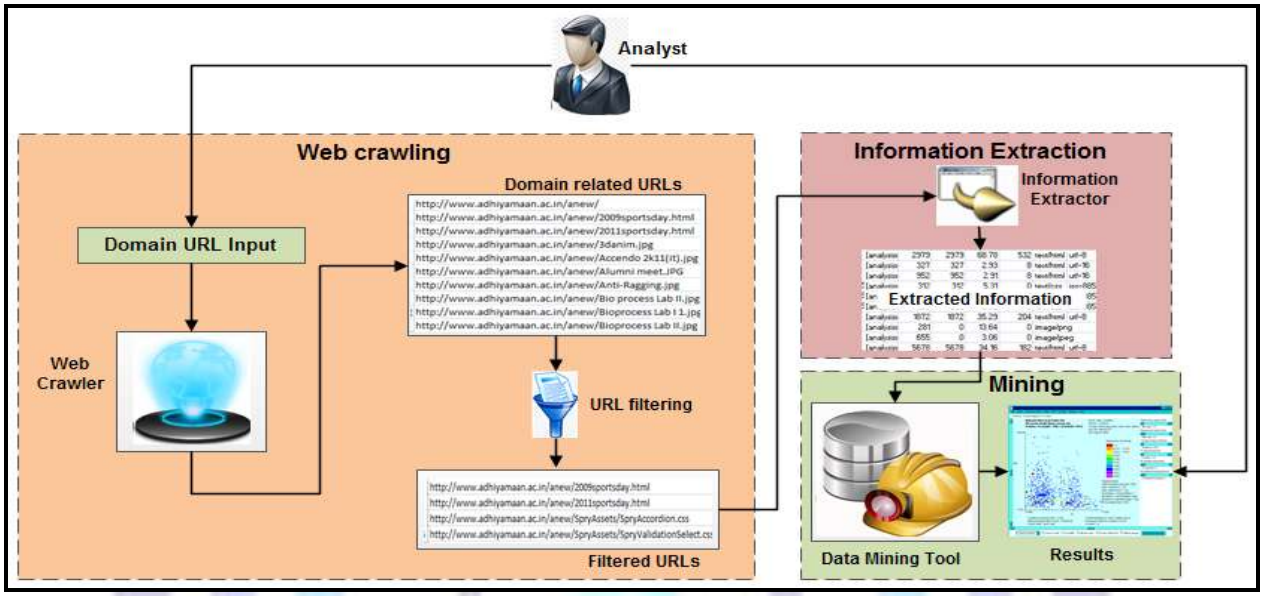


Fig 1: Proposed system architecture

3.1 Web Crawling

This is primary module of the proposed system. The web crawling module takes application domain URL as input and produces URLs of web pages of the application domain. It uses a web crawler for crawling the application domain. After crawling the application domain, it produces all the domain related URLs as output. In the output there may exist URLs of images, audio and video resources along with the URLs of web documents. URL filter is used to obtain only the URLs of web documents.

3.2 Information Extraction

Information extraction module takes the crawled URLs of the web documents of the application domain as input. This module uses an information extractor to extract information from the crawled web pages. It enables user to store the extracted information in native system or in database.

3.3 Mining

This module is the major complement for the current work. Mining module enables user or analyst to perform several data mining techniques on the extracted information. Popular data mining techniques are classification, clustering forecasting and comparison analysis. This module uses a data mining tool to perform mining operations on the extracted information.

4. SYSTEM IMPLEMENTATION

The proposed system is implemented using java net beans IDE [12]. Excel CSV format is used to store information extracted from web pages. The system uses win web crawler [13], easy web extractor [8] and weka data mining tool [14] as plug-ins. Win web crawler is a popular web crawler. It supports URL filter and domain filter etc. Easy web extractor can automatically extract semi-structured information from web pages and has support for export extracted data in structured formats like CSV, Access and ODBC data source etc. weka is a data mining software developed in java. Weka consist comprehensive set of data preprocessing tools, learning algorithms and evaluation methods [15]. Fig. 2. shows main interface of the proposed system implementation.

The proposed system uses three tools, web crawler, extractor and mining tool. Web crawler enables the user to easily obtain all the URLs of the application domain and filtering of the URLs. The crawled URLs are taken by information extractor. Information extractor allows data analyst to select data attributes and information regions to be extracted according to the users preference. The information extractor performs extraction operation automatically over the selected region and extracts values for the data attributes. User has to store the extracted information in a convenient format for the further process (data analysis). The information extractor module allows the user to export and save the extracted information in various formats such as excel (CSV), MS Access database, SQL Script file and ODBC data base etc.

The major objective of the proposed system is to analyze the extracted information. Data mining is a powerful and popular technology for effective and efficient data analysis. It provides several techniques such as classification, clustering,

association etc. It often uses statistical techniques to perform various calculations such as mean, variance, correlation etc. testing methods to derive conclusions by checking the consistency between the values of various data attributes.

Mining module of the proposed system provides user friendly interface to perform analysis over the extracted data. This module takes input in various formats like CSV, ARFF, binary and SQL database using JDBC URL. Once the data is imported into the system user can:

- Perform classification and clustering of data records based on the values of a particular data attribute.
- Find association between various attributes of the information.
- Check data consistency between two attribute values.
- View the relationship between data attributes in clear visuals.

Fig 2: System interface

5. RESULTS

The proposed system is implementd and evaluated with mobile service provider information available in the web. Fig 3a shows the crawled URLs. Fig. 3b shows the extracted information. The extracted information is sorted in tabular format. In fig 3c the extracted information sorted on offer cost. Fig 3d present a bar diagrams that represents number of models available in each mobile brand. Fig 3e shows comparative analysis between various attributes of the mobile product such as brand name, mobile type, offers, mobile cost, offer cost etc. Fig 3f shows scattered diagram of brand name and their offer costs. From this experiment analyst can find the following information (see fig. 3):

- Numbers of models are available in each brand in the current market.
- Which model and brand is at the lowest and highest price.
- Models and brands available for below and above the mean cost.

6. CONCLUSION AND FUTURE WORK

The paper describes an integrated frame work for data extraction from web pages and data analysis for effective decision making. The frame work is an efficient system that gathers the information from various sources and summarizes the information from which knowledge can be extracted which in turn helps in decision making. The proposed system enable user to easily find all the web pages of given domain URL, extract the data from the all web pages and to make analysis on the extracted data using a data mining tool. The proposed system is quite useful for individuals: students, researchers, managers etc as well as the organizations to collect web links and data from various sources of the web. And mine the data to produce meaningful reports with the appropriate knowledge about the domain. These reports or knowledge could be used in crucial decision making. The proposed approach is a general frame work which is applicable for any application domains like e-learning, research and development, products: demand forecast, manufacture, suppliers, services, sales, tourism, recruitment, agriculture etc. The system is implemented and evaluated with mobile service provider information available in the web.

Automatic filtering of crawled web pages according to user need and effective method for extracting unstructured information from the web and analysis are the future enhancements of the proposed work.

ISSN 22773061

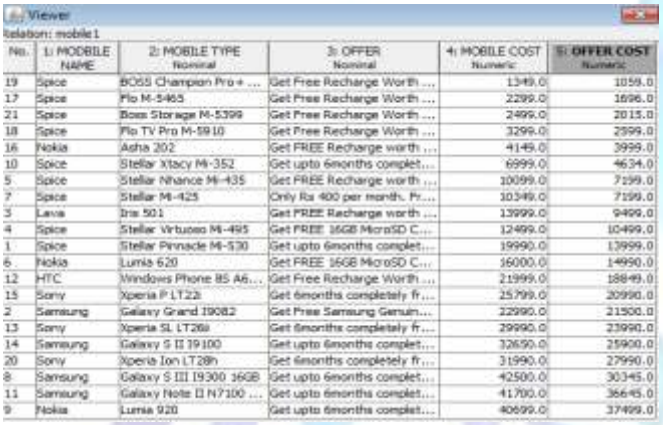
URL LIST OF WEB CRAWLER:

http://www.saholic.com/mobile-phones/spice-stellar-pinnacle-mi-530-1005699
http://www.saholic.com/mobile-phones/karbonn-retina-a27-1006405
http://www.saholic.com/mobile-phones/karbonn-a9+-1005109
http://www.saholic.com/mobile-phones/samsung-galaxy-grand-i9082-1005576
http://www.saholic.com/mobile-phones/lava-iris-501-1005730
http://www.saholic.com/mobile-phones/lg-optimus-g-e975-1006076
http://www.saholic.com/mobile-phones/spice-stellar-virtuoso-mi-495-1005573
http://www.saholic.com/mobile-phones/karbonn-a15-1005188
http://www.saholic.com/mobile-phones/spice-stellar-nhance-mi-435-1005698
http://www.saholic.com/mobile-phones/nokia-lumia-620-1005724
http://www.saholic.com/mobile-phones/karbonn-a21-1004960
http://www.saholic.com/mobile-phones/karbonn-android-a1+-1004941
http://www.saholic.com/mobile-phones/spice-stellar-mi-425-1004195
http://www.saholic.com/mobile-phones/sony-xperia-tipo-dual-st21i-1004842
http://www.saholic.com/mobile-phones/karbonn-a9-1004501
http://www.saholic.com/mobile-phones/karbonn-a2-1005734
http://www.saholic.com/mobile-phones/karbonn-smart-a111-1005893
http://www.saholic.com/mobile-phones/sony-xperia-z-1005704
http://www.saholic.com/mobile-phones/lava-iris-502-1006062
http://www.saholic.com/mobile-phones/xolo-a1000-1005991

Fig 3a: Crawled URLs

No.	1: MOBILE NAME	2: MOBILE TYPE Nominal	3: OFFER Nominal	4: MOBILE COST Numeric	5: OFFER COST Numeric
1	Spice	Stellar Pinnacle Mi-530	Get upto 6months complet...	19990.0	13999.0
2	Samsung	Galaxy Grand I9082	Get Free Samsung Genuin...	22990.0	21500.0
3	Lava	Iris 501	Get FREE Recharge worth ...	13999.0	9499.0
4	Spice	Stellar Virtuoso Mi-495	Get FREE 16GB MicroSD C...	12499.0	10499.0
5	Spice	Stellar Nhance Mi-435	Get FREE Recharge worth ...	10099.0	7199.0
6	Nokia	Lumia 620	Get FREE 16GB MicroSD C...	16000.0	14990.0
7	Spice	Stellar Mi-425	Only Rs 400 per month. Fr...	10349.0	7199.0
8	Samsung	Galaxy S III I9300 16GB	Get upto 6months complet...	42500.0	30345.0
9	Nokia	Lumia 920	Get upto 6months complet...	40699.0	37499.0
10	Spice	Stellar Xtacy Mi-352	Get upto 6months complet...	6999.0	4634.0
11	Samsung	Galaxy Note II N7100 ...	Get upto 6months complet...	41700.0	36645.0
12	HTC	Windows Phone 8S A6...	Get Free Recharge Worth ...	21999.0	18849.0
13	Sony	Xperia SL LT26ii	Get 6months completely fr...	29990.0	23990.0
14	Samsung	Galaxy S II I9100	Get upto 6months complet...	32650.0	25900.0
15	Sony	Xperia P LT22i	Get 6months completely fr...	25799.0	20990.0
16	Nokia	Asha 202	Get FREE Recharge worth ...	4149.0	3999.0
17	Spice	Flo Mi-5465	Get Free Recharge Worth ...	2299.0	1696.0
18	Spice	Flo TV Pro Mi-5910	Get Free Recharge Worth ...	3299.0	2599.0
19	Spice	BOSS Champion Pro+ ...	Get Free Recharge Worth ...	1349.0	1059.0
20	Sony	Xperia Ion LT28h	Get 6months completely fr...	31990.0	27990.0
21	Spice	Boss Storage Mi-5399	Get Free Recharge Worth ...	2499.0	2015.0

Fig 3b: Extracted Information



No.	1: MOBILE NAME	2: MOBILE TYPE Nominal	3: OFFER Nominal	4: MOBILE COST Numeric	5: OFFER COST Numeric
19	Spice	BOSS Champion Pro+ ...	Get Free Recharge Worth ...	1349.0	1059.0
17	Spice	Flo Mi-5465	Get Free Recharge Worth ...	2299.0	1696.0
21	Spice	Boss Storage Mi-5399	Get Free Recharge Worth ...	2499.0	2015.0
18	Spice	Flo TV Pro Mi-5910	Get Free Recharge Worth ...	3299.0	2599.0
16	Nokia	Asha 202	Get FREE Recharge worth ...	4149.0	3999.0
10	Spice	Stellar Xtacy Mi-352	Get upto 6months complet...	6999.0	4634.0
5	Spice	Stellar Nhance Mi-435	Get FREE Recharge worth ...	10099.0	7199.0
7	Spice	Stellar Mi-425	Only Rs 400 per month. Fr...	10349.0	7199.0
3	Lava	Iris 501	Get FREE Recharge worth ...	13999.0	9499.0
4	Spice	Stellar Virtuoso Mi-495	Get FREE 16GB MicroSD C...	12499.0	10499.0
1	Spice	Stellar Pinnacle Mi-530	Get upto 6months complet...	19990.0	13999.0
6	Nokia	Lumia 620	Get FREE 16GB MicroSD C...	16000.0	14990.0
12	HTC	Windows Phone 8S A6...	Get Free Recharge Worth ...	21999.0	18849.0
15	Sony	Xperia P LT22i	Get 6months completely fr...	25799.0	20990.0
2	Samsung	Galaxy Grand I9082	Get Free Samsung Genuin...	22990.0	21500.0
13	Sony	Xperia SL LT26ii	Get 6months completely fr...	29990.0	23990.0
14	Samsung	Galaxy S II I9100	Get upto 6months complet...	32650.0	25900.0
20	Sony	Xperia Ion LT28h	Get 6months completely fr...	31990.0	27990.0
8	Samsung	Galaxy S III I9300 16GB	Get upto 6months complet...	42500.0	30345.0
11	Samsung	Galaxy Note II N7100 ...	Get upto 6months complet...	41700.0	36645.0
9	Nokia	Lumia 920	Get upto 6months complet...	40699.0	37499.0

Fig 3c: Extarcted information

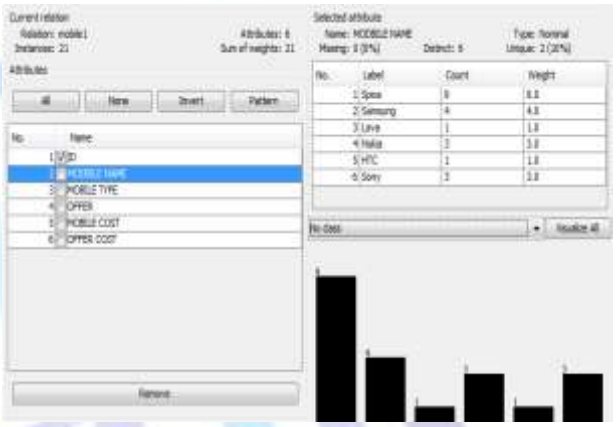


Fig 3d: Models available in each brand

Fig 3e: compative analysis between mobile attributes

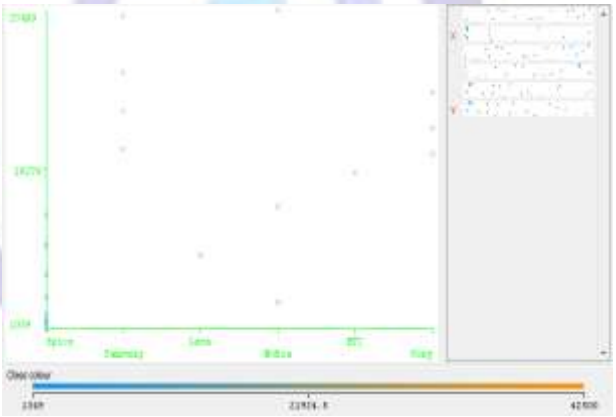


Fig 3f: scattered analysis between brands and their costs

Fig. 3. Experimental results

ACKNOWLEDGMENTS

The work presented in this paper is done as part of a sponsored project founded by government of India, Ministry of Defence, DRDO (ER&IPR), and done in the labs of Adhiyamaan College of Engineering where the author is working as a Professor& Director in the department of Master of Computer Applications. The author would like to express her sincere thanks to DRDO for providing the support.

REFERENCES

[1] Paul Deitel, Harvey Deitel, Abbey Deitel, “Internet and World Wide Web How to Program”, 5th Edition, pearson, 2011.

[2] M. G. Costa Júnior and Z. Gong, “Web Structure Mining: An Introduction,” In Proc. 2005 IEEE International Conference on Information Acquisition, Hong Kong and Macau, pp. 590-595, China, June 27 - July 3, 2005.

[3] K. Pol, N. Patil, S. Patankar and C. Das, “A Survey on Web Content Mining and Extraction of Structured and Semistructured Data,” In Proc. First International Conference on Emerging Trends in Engineering and Technology, IEEE, pp. 543-546, 2008.

[4] N. Joseph-Williams, R. Evans, A. Edwards, R. GNewcombe, P. Wright, R. Grol and G. Elwyn, “Supporting Informed Decision Making Online in 20 Minutes: An Observational Web-log Study of a PSA Test Decision Aid”, Journal of medical internet research (JMIR), vol 12, No. 2, 2010; available at http://www.jmir.org/2010/2/e15/.

[5] X. Ren , X. Kang, et al, “Web Crawlers Design and Implementation Based on Dynamic Tunneling“, New technology of library and information service, no.6, pp.83-87, 2008.

[6] Xiaoyan Ren and Yunxia Fu, “Web Information Extraction Based on IEBIDTech”, In Proc. 2010 Conference on Dependable Computing (CDC’2010), Yichang, China, pp. 239-241, November 20-22, 2010.

[7] C. Chang, M. R. Girgis, “A Survey of Web Information Extraction Systems”, Transactions on Knowledge and Data Engineering, IEEE, VOL. 18, NO. 10, pp. 1411-1428, October 2006.

[8] http://webextract.net/

[9] Brijendra Singh and Hemant Kumar Singh, “Web Data Mining Research: A Survey”, In Proc. IEEE International Conference on Computational Intelligence and Computing Research (ICCIC), pp. 1-10, 2010.

[10] Margaret H. Dunham, “Data Mining Introductory and Advanced Topics”, Pearson Education

[11] Y. Wang, Z. Gu and H. Wang, “A Survey of Data Mining Softwares Used for Real Projects”, In Proc. 2011 International Workshop on Open-Source Software for Scientific Computation (OSSC), IEEE, pp. 94-97, 2011.

[12] https://netbeans.org/

[13] http://www.winwebcrawler.com/

[14] http://www.cs.waikato.ac.nz/ml/weka/

[15] R. R. Bouckaert, E. Frank, M. Hall, R. Kirkby, P. Reutemann et al., “WEKA Manual”, Version 3-7-5, October 28, 2011.

Dr. Sunitha Abburu, received BSc and MCA from Osmania University, A.P, India. M.phil and Ph.D from Sri Venkateswara University, A.P, India. She is having 16 years of teaching experience and 3 years of industrial experience. Currently she is working as a Professor and Director, in the Department of Master of Computer Applications, Adiyamaan College of Engineering, Hosur, Tamilnadu, India.

G. Suresh Babu, received BSc and MCA from Sri Venkateswara University, A.P, India. He is having 3.5 years of teaching experience and 1.5 years of research experience. Currently he is working as a JRF, in the Department of Master of Computer Applications, Adiyamaan College of Engineering, Hosur, Tamilnadu, India.