Volume 6, Issue 8, August 2016 ISSN: 2277 128X
International Journal of Advanced Research in Computer Science and Software Engineering
Research Paper
Available online at: www.ijarcsse.com

A Survey on Automated Web Data Extraction Techniques for Product Specification from E-commerce Web Sites

Harish Rao, M*
M.Tech (CSE), CITECH,
VTU University Bangalore, India

Sashikumar, D R
Dept. Of CSE, CITECH,
VTU University, Bangalore, India

Abstract— In this survey paper we have briefly described the challenges of web data extraction and discussed briefly on published research topics on supervised, semi-supervised and unsupervised web data extraction techniques. We have also discussed on the metrics used for evaluation and improvement of classifiers used for learning algorithm improvement. Our survey is more specific to product specification extraction from e-commerce web sites and therefore we have discussed about the attempts made by different researchers using supervised, semi-supervised and unsupervised algorithms. In the end, we have discussed a recent research work contributing towards automatic product specification extraction from both structured and unstructured product pages. Finally, we discussed about the future direction of research work for automatic product specification extraction and product catalogue updation.

Keywords— Wrapper Induction, Automatic Product specification extraction, web data extraction, E-commerce websites, Learning algorithms, X-Path, Regular Expressions, unsupervised web data extraction

I. INTRODUCTION

The demand for efficient and accurate web data extraction techniques for product specification is spurred by the rapid growth of E-commerce business. Customers, Retailers and service companies make extensive use of web for gathering detailed product information. For e-commerce shops and retailers, accurate information on product specification can help them in demand forecasting, assortment optimization, product recommendations and assortment comparison across the retailers and manufacturers. For consumers, they will be interested in only specific product attributes to make their purchasing decision after comparing the different product offers from various e-commerce sites, shopping portals and product review sites. Producers and retailers need specific attributes of the product on which customers make their purchasing decision to improve their product design and product offers to increase their sales. Ever increasing growth of internet usage and rapid developments in technology and software applications has turned the World Wide Web into a huge knowledge database. Tapping this treasure of knowledge has been the subject of great interest for computer science and information technology researchers. All the research efforts [1-5] in this context have been directed to pursue two fold objectives. The first objective is to develop the web search and extraction applications to reduce the human intervention to the minimum so as to extract the right information and conserve time which is a very precious economic resource. The second objective is to ensure that the reduction in human efforts does not compromise the requirement of high accuracy and precision in extracted data and information while minimizing the probability of garbage collection in the form of irrelevant data and information. Product specification extraction is a subset of many different customized requirements for web data mining. This survey paper confines itself to the web mining techniques for product specification extraction and aims to enlighten how those twin objectives of automation and accuracy are being dealt by different research teams working on software applications for product specification extraction from the world wide web.

II. WEB DATA EXTRACTION

Web pages are complex entities consisting of main textual and image content. A web page today also contains header, footer and side area blocks. There are also navigation links and advertisements embedded in the web page. The most important challenge is the heterogeneity and diversity of web page in the design and the content. The data embedded in the web page can be unstructured, semi-structured or structured. Web pages may contain non-semantic and semantic elements. Non-semantic elements used in old browsers like and do not tell anything about content while semantic elements (HTML5) used in new browsers like <form>, <table> and <img> clearly define the content. As per current available information [1] about 18% of the browsers crawled by web crawlers have semantic elements and thus structured data while others are still using non-semantic elements having fuzzy, unstructured or semi structured data. In some cases, semantic web is used partially and only partial description of data available on them. Besides this, there are syntactical errors to contend with. The early data extraction applications used programs named as "wrappers" [7-11] which are defined as "procedures which implement a family of algorithms to find, the information a user needs, extract this from unstructured sources and transform them into structured data". The main wrappers' assumption is that one can find repetitive patterns in one or more Web pages that conform to a common template.

There are different ways to treat a web page. If we see a web page as a stream of characters, we can use HTML page as a text document and apply regular expressions (regex) and other standard extraction methods to extract data as per requirements. Considering the web page as DOM tree, we can extract data using X-path. Using X-path make data extraction simple but we may have to sacrifice some information embedded in HTML elements. This is to ensure high precision in data extraction when a HTML mark-up contains multiple h1 tags on a single page. There is an extended X-path, called OXPATH, which allows the execution of user actions such as click, navigation through data paginated Pages and identification of data for extraction. A more declarative language is Ducky, which also does the post processing of extracted data. Nexir is similar to Ducky but using XML and X-path rules.

III. AUTOMATION IN DATA EXTRACTION TECHNIQUES

Kushmeric [2] made trend setting contribution towards reduction of manual programming of wrappers by developing the wrapper induction techniques which led to many path breaking research contribution towards more and more semi-supervised and unsupervised data extraction techniques[12-20].All web data extraction application algorithms based on wrapper induction techniques have one thing common among them. They all need a set of data extraction rules to train the system (machine learning) so that only required and specified data is extracted from the crawled websites. Training is an iteration process through test data until some desired degree of accuracy is achieved in the data extraction. Figure-1 depicts this process [12]

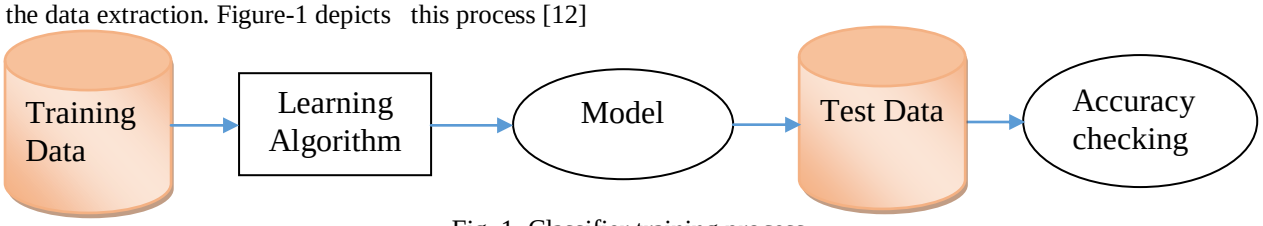

Fig. 1. Classifier training process

If accuracy is not satisfactory, the training and testing process is iterated with modifications to learning algorithm or by data pre-processing until desired accuracy is attained. Data mining techniques are broadly classified [12] as supervised(manual), semi supervised(Semiautomatic) and unsupervised (Automatic) based on degree of automation for generating learning algorithms. Learning algorithms are generated through labelled and unlabelled examples. Classification of data extraction techniques are shown in Table-1

Table I Classification of Data Extraction Techniques

Degree of automation	**Features**		
	Classification techniques	***Learning algorithm***	***Limitations***
Supervised (manual)	Decision Tree induction, Rule Induction and support vector machines(SVMs)	Learning is fully supervised and human dependent. Use heavily Labelled examples for learning	Not scalable and inefficient due to manual involvement
Semi supervised	EM Algorithm, NB classfier, Co-training, self-training, TSVM, graph-based methods	Use partly labelled examples and mostly unlabelled example	Better scalability but still some manual work required for creating labelled examples
Unsupervised	Pattern mining, clustering and ontology	Very low or no use of labelled examples and mostly all unlabelled examples	Scalable but require large number of example pages. In case of no use of labelled examples, accuracy is affected often

Efficiency and accuracy of learning algorithm depends on classifiers or models and there are three metrics used for measuring the predictive capability of classifiers as shown in Table-2. The data set used for learning is called the training data (or the training set). After a model is learned or built from the training data by a learning algorithm, it is evaluated using a set of test data (or unseen data) to assess the model accuracy. The accuracy of a classification model on a test set is defined as:

Accuracy = Number of correct classifications / Total number of test cases

Table II Classfier Performance Metrics

Performance Measure	***Calculation method***	***Significance***
Precision	Ratio of correctly classified positive examples to total number of examples	Exactness of classifiers and low precision indicate large number of

		False positives
Recall	Ratio of correctly classified positive examples to total number of actual positive examples	Indicates completeness of classifier and a low recall indicates large number of False negatives
F1 score	F score is the harmonic mean of Precision and Recall	F1 indicates balance between precision and recall

IV. WEB EXTRACTION OF PRODUCT SPECIFICATIONS FROM THE E-COMMERCE SITES

All product specifications extraction methods make use of the attribute-value pairs [21-29] which are called either Name Value pair (NVP) or product property-value pairs (PVP). The basic steps required are crawling the e-commerce or producers websites, extracting the product page followed by extracting the product attribute and value data and storing them in a suitable structured format. Product specification extraction poses lot of challenges due to following issues on commercial websites

- Product information on most e-commerce websites is mostly volatile and gets frequently updated and in some cases generated dynamically on the fly when user browses for the product.
- HTML pages are mostly optimized for human browsing rather than machine processing and product specification web pages vary across the e-commerce web sites
- Product attribute value pairs are differently represented with some in structured tabular form and some in unstructured text formats.
- Usage of different product properties or different names for the same attributes or values as well as representing them in different order from site to site.
- Missing out on some attribute-value pairs even on the same web site
- Presence of many other blocks along with product information like tool bars, navigation bars with browsing links of the website, product visuals which poses a great challenge for automatic extraction of product specification on the product page

Researchers applied different algorithms and methods to overcome above limitations for successful extraction of product specification. Both semi-supervised and unsupervised approaches have been tried by different application developers. Walthers described a web extraction technique [24] which takes product name and manufacturer's name to retrieve product page and product detail page using popular web search engines such as Google, Bing and Yahoo. A scoring system is used to rank the product pages and key-value pair based algorithm extracts product details using X-path and DOM tree. Both supervised as well as unsupervised algorithms used with and without domain knowledge. Implementation is done through Fedseeko for 100 products, 40 different producers and 10 diverse application domains. They found higher success rate with domain knowledge to get only 10% spurious data as compared to 35% spurious data with out domain knowledge

Bo Wu described a template independent semi supervised method [25] with simultaneous extraction of attribute-value pairs They have used a co-training algorithm which combines labelled and unlabelled data together to reduce the requirement of labelled data in training the classifiers. They have tried this model on three product categories of Laptops, Digicams and Books. They have compared the metrics using only name or value or Name-value together and found good precision, recall and F-scores while using simultaneous extraction of NVP. George and Ebrahim [26] made an improvement over Walters [24] and Bo wu [25] techniques using common short (3 word length) property-value pairs (PVPs). Product specification extraction by their methods have four steps- pre-processing, seed learning, pattern discovery, and pattern based extraction. They have used classifier algorithms such as BPM, SVM and NB for predicting the P or V labels from text chunks based on maximum token limit. They have implemented the methods on 100 distinct electronic goods retailer sites from two domains of digicams and smart phones. They found all three metrics above 99 % while using known seeds .Petrovisky used a feature extraction method [27] which uses learned Regular expressions from microdata on html pages from highly structured websites such as Amazon. They have used automatic induction of regular expressions from positive and negative examples using genetic programming. In a twostep experiment they applied algorithm to 500 products on structured Amazon catalogues. Learned examples from Amazon experiment are then applied to 5000 other web product pages having both structured and unstructured product information. They used a Fitness function measure to evaluate the metrics for their models. With numeric and semi numeric data they got F-measure of 89%. Ghani employed a semi-supervised co-EM algorithm [28] to collect a vast amount of unlabelled data and use a dependency parser to link the extracted attributes with corresponding values.

Dexter an open source application [29] makes use of focussed crawler technique .It makes use of vote-filter-iterate model. From a small set of seed web pages from large popular websites and product specification in a category ,Dexter iteratively obtains a large set of product specifications Each iteration has voting and filtering to prune irrelevant web sites and web pages to reduce the noise in the pipeline to result in high quality product specification. It was tested on 1.46 M product specifications, 3005 websites and 9 product categories they got an F-core of 87 % for unknown websites and 94-97% for known websites. A hybrid approach including both domain dependent and domain independent classification methods used to get a precision of 92% and recall of 95%. However no semantic integration is done for the collected product specification across the web sites. Dexter [29] provides only inputs for data integration studies

V. AUTOMATIC SPECIFICATION EXTRACTION FOR CONSOLIDATED PRODUCT CATALOGUES

All the above techniques implemented the product specification extraction but none of them worked on generating automatic consolidated catalogues from the extracted specification data. Stuthi [30] developed a novel specification extractor which can extract product specifications from both structured as well as unstructured web pages. For structured HTML it uses DOM tree for extraction and for unstructured web pages it uses regular expressions for pattern matching. The following steps describe the process flow in table-3

Table III Process Steps For Automatic Product Specifcation Extraction

Steps	***Process name***	***Process input***	***Process output***
1	Crawling/Scraping	Traversed e-commerce websites	Scraped product pages stored in database
2	Branding	Product pages	Identified Brand names
3	Categorization	Product pages	Product placement into respective categories
4	Disambiguation	Product pages	Same products with different names are regrouped to single product identification
5	Spec.Extraction and mapping	Product pages	Extracted specification of products
6	Updation into catalogue	Mapped specification	Updated catalogue with addition of new products

The components of specification extractor is shown in figure-2 as in [30]

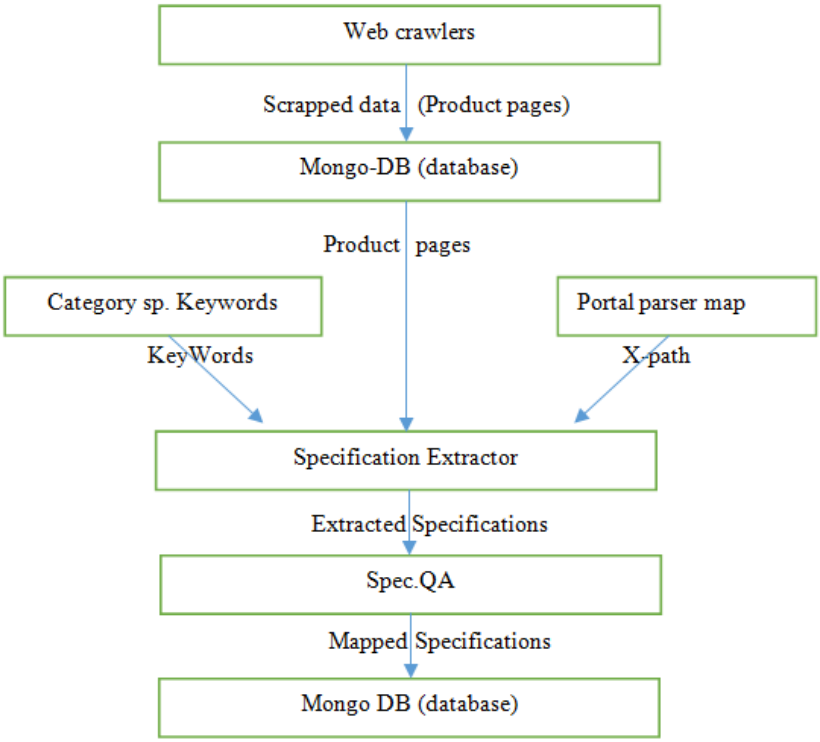

Fig. 2 Specification Extractor components

Category specific key words list is provided as an initial input to the extraction algorithm. Example keywords for a washing machine could be capacity and for a TV set it could be display size. The algorithm is self-adaptive in the sense that the algorithm adds new keywords to its existing list in case the product pages have keywords which are not found in the keywords file supplied to it. In cases of product pages having structured data of product specifications X-path based parsers locate the specific elements in the element tree for extracting the product specification data from the tree nodes. In the case of product pages where product specifications are in unstructured format, a portal specific parser map is created for each portal which provided the information as to where to look for the specifications of a particular portal's products. The map is organized as a dictionary structure having a few parsers for each portal. For specification extraction from tree structures, a heuristically computed match score metric is applied to identify the node with maximum score for specification extraction. For unstructured data python built in functions are used pattern matching using regular expressions.

Results obtained by this specification extractor were tested [30] manually. For results of extractions from structured product pages it is checked to verify the selection of correct node, as spec-node, the retrieval of proper specification in the form of key-value pairs and whether the dictionary is blank in absence of structured data instead of garbage values. For unstructured product pages where product specification are in descriptive form, the results were consolidated and dumped into a tab-separated format where they could be easily validated.

A screen shot of the test run for the unstructured data for the category Baby Products as in [16] is shown in figure.3

1	Maclaren Techno XT - Scarlet 53	7 kgs	3 Years	16.5 cm
2	Maclaren Techno XT - Sky Blue 63	15 kg	8 - 18 months	16.5 cm
3	MeeMee Baby Pram	35 kg	1 - 3 Years	530 x 900 x 330 mm
4	Hauck Disney Baby Rio - Cars Trac		3 Years	490 mm
5	Safety 1st Easy Go		1 Year	
6	Maclaren Volo Baby Pram	15 kg	3 - 6 Years	11.4 cm
7	Maclaren Twin Techno	12.2 kgs	6 - 10 Years	15.2 cm
8	Hauck Disney Baby Shopper 6 - Po		1 year and above	490 mm
9	Hauck Esprit - Speed Sun			18 cm
10	Mothercare Stroller M Bu	15 kgs	0 Month +	108 W x 25 D x 21 H
11	Hauck Run Buggy		3 Month	460 mm
12	Safety 1st Compa'City Stroller	15 kg	0 Month +	490 x 1030 x 770 mm
13	Hauck Roadster Stroller		3 Month	24 cm
14	Hauck Pram		3 Month	101 cm

Fig. 3 Extracted specifications for Baby products from unstructured product page

Figures 4 and 5 give the sample cases of HTML Input page and corresponding extracted data for structured and unstructured data from product pages as in [30]

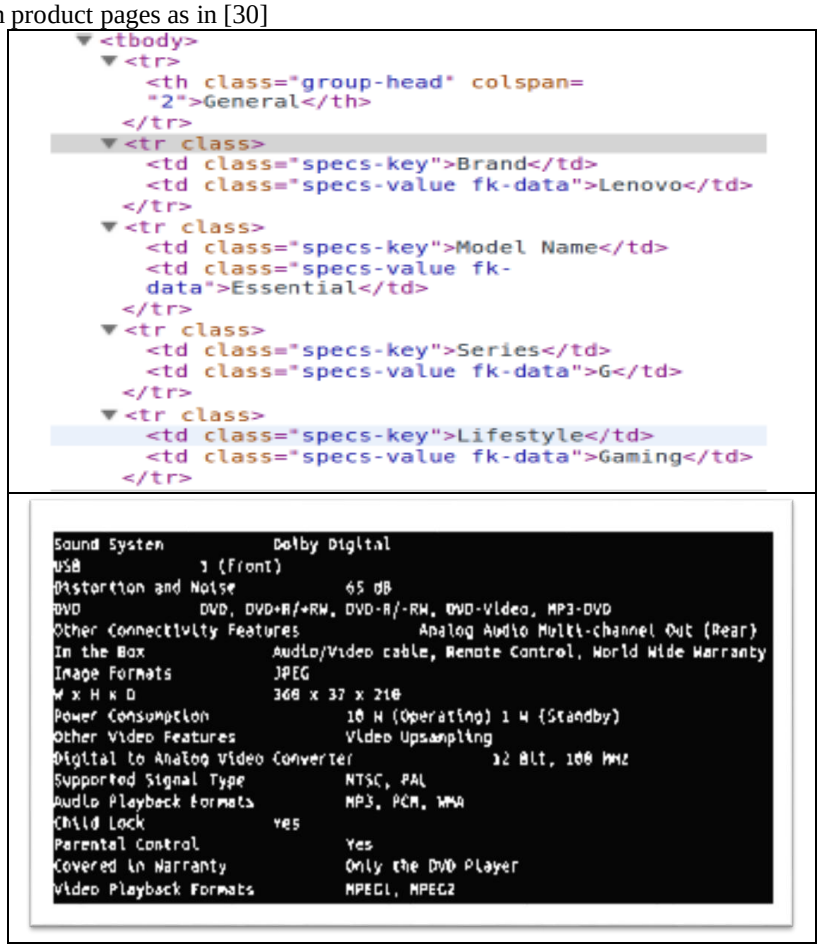

Fig. 4: Sample input HTML page with structured data and its extraction results

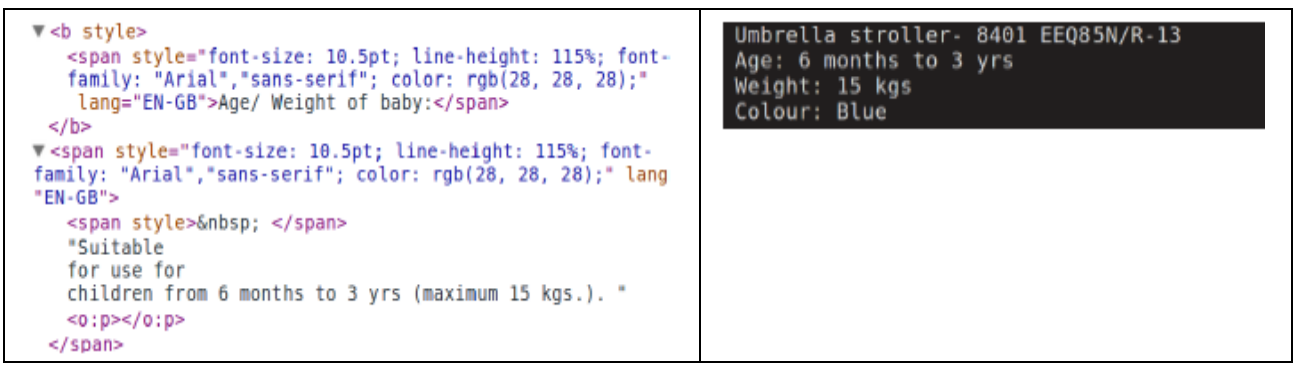

Fig 5. Sample input HTML page with unstructured data and its extraction results

Thus it is tested and proved [30] that the method successfully extracts product specifications from both structured and unstructured product data pages without collecting any garbage values during the process

According to contributors of this automatic product specification extraction application , the future enhancement possibilities for specification extractor could be to extend it to (1) extract product attributes such as colour, (2) use NLP instead of HTML parser for specification extraction, (3) use frequency of keyword occurrence to more efficient extraction and (4) use of semantic understanding instead of pattern matching.

VI. CONCLUSIONS

Existing published information shows that web data extraction techniques are continuously explored to achieve the twin objectives of accuracy and speed. Many novel approaches are successfully tried to extract accurate data from heterogeneous (structured, unstructured) web pages using template dependent and template independent methods using X-path, DOM tree and regular expressions. Product specification extraction from e-commerce sites has been a popular research topic due to its high importance with growth of e-commerce websites. A great degree of success has been shown in both semi-supervised as well as unsupervised specification extraction attempts by using different classification algorithms. The most significant contribution in this area is the automatic product specification extraction from structured and unstructured websites using python and mongo DB which can produce final output in the form of product catalogue. There is more work needed to improve the accuracy and efficiency of automatic product specification extraction in terms of adding non-numerical attribute-value pairs and use of NLP instead of HTML parser

ACKNOWLEDGMENT

The Author acknowledges the support given by the Cambridge Institute of Technology, Bangalore, Karnataka, India for providing the necessary library resources for this survey

REFERENCES

[1] Andres Viikmaa, "Web Data Extraction For Content Aggregation From E-Commerce Websites", Master's thesis, University of Tartu,2016,p.7

[2] N.Dalvi,Ravi Kumar, and Md.Soliman," Automatic Wrappers for Large Scale Web Extraction", Proceedings of the VLDB Endowment, Volume 4,No 4, August 29th - September 3rd 2011, Seattle, Washington,pp.219-230

[3] V.Crescenzi,G.Mecca, and P.Merialdo," RoadRunner: Towards Automatic Data Extraction from Large Web Sites", VLDB '01 Proceedings of the 27th International Conference on Very Large Data Bases,Italy,pp109-118.

[4] R.Manjula and A.Chilambuchelvan," An Effective Approach to Extract Information from web pages", International Research Journal of Engineering and Technology (IRJET),Volume 3,Issue 4, April,2016,pp.1770-1776.

[5] V.Gupta and G.S.Lehal," A Survey of Text Mining Techniques and Applications", JOURNAL OF EMERGING TECHNOLOGIES IN WEB INTELLIGENCE, VOL. 1, NO. 1, AUGUST 2009, pp.60-76.

[6] N.Kushmerick, "Wrapper Induction for Information Extraction" Ph.D. thesis,Univeristy of Washington,1997

[7] A.H. F. Laender ,B.A.Ribeiro-Neto,A.S,de Silva, Newsletter,ACM SIGMOD Record, Volume 31, Issue 2,June,2002,pp84-93.

[8] S.Flesca,G.Manco,E.Masciari,E.Rende,and A.Tagarelli, "Web Wrapper Induction: A Brief Survey", I Communication, Volume 17,Issue 2, April,2014,pp.57-61.

[9] G.sigletos,G.Paliouras,C.D.Spyropoulos and M.Hatzopoulos," Mining Web sites using wrapper induction, named entities and post-processing", First European Web Mining Forum, EWMF 2003, Cavtat-Dubrovnik, Croatia, September 22, 2003, Invited and Selected Revised Papers, ISBN: 978-3-540-23258-2 (Print) 978-3-540-30123-3 (Online)

[10] A.Firat,D.Peleshchuk and Prakash Rao, "IWrap: Instant Web Wrapper Generator", Working Paper CISL# 2000-10,June,2000,pp.1-9

[11] M.Nekvasil," The Use of Ontologies in Wrapper Induction" J. Pokorný,Dateso 2007, pp. 132–135

[12] Y.C.Chun Chu, C.C. Hsu, C.J. Lee andY.T. Tsai , "Automatic data extraction of websites using data path matching and alignment", Digital Information Processing and Communications (ICDIPC), 2015,ISBN: 978-1-4673-6831-5,pp 60-64

[13] S.W.Liddle,S.H.Yau and ,D.W.Embley, "On the Automatic Extraction of Data from the Hidden Web", Conceptual Modeling for New Information Systems Technologies, Lecture Notes in Computer Science,ISBN: 978-3-540-46140-1,volume 2465,2002,pp-212-226

[14] P.YesuRaju and P.KiranSree," A Language Independent Web Data Extraction Using Vision Based Page Segmentation Algorithm",IJRET,Volume 2,Issue 4,ISSN:2319-1163,April,2013,pp 635-639

[15] E.Ferraraa, ,P. De Meob, , G. Fiumarac, , and R. Baumgartnerd, "Web Data Extraction, Applications and Techniques: A Survey", Knowledge Based Systems, Elsevier, Volume 70,Nov.2014,pp 301-323

[16] A.Bhagat and V.Raut," A Survey on Unsupervised Extraction of Product Information from Semi-Structured Sources", International Journal of Computer Science and Information Technologies, Vol. 6 (3) , 2015, pp.2893-2897

[17] Devika,K, Subu,S.," An Overview of Web Data Extraction Techniques", International Journal of Scientific Engineering and Technology (ISSN : 2277-1581),Volume 2 Issue 4, April 2013, pp : 278-287

[18] K.Thirunarayan,ABerkovich and D.Sokol," Semi-automatic Content Extraction from Specifications",Natural Language Processing and Information Systems,Lecture Notes in Computer Science,Volume 2553,pp40-51.

[19] A.V.A.Mary,S.J.Samuel, and D.J.Rajam," Automated Trinity Based Web Data Extraction for Simultaneous Comparison", Contemporary Engineering Sciences, Vol. 8, 2015, no. 11, pp.491 – 497.

[20] P.V.P.Sundar," Towards Automatic Data Extraction Using Tag and Value Similarity Based on Structural - Semantic Entropy", International Journal of Advanced Research in Computer Science and Software Engineering, Volume 3, Issue 4, April 2013,pp..226-231.

[21] K.Probst,R.Ghani,M.Krema, A fano and Y. Liu,," Semi-Supervised Learning to Extract Attribute-Value Pairs from Product Descriptions on the Web", Accenture Technology Labs, Chicago.IJCAI,2007,pp.2838-2843.

[22] W.Hu,Z.Gong, and J.Guo," Mining Product Features from Online Reviews", IEEE 7th International Conference on e-Business Engineering (ICEBE), 2010,pp.24.29.

[23] W.Choochaiwattana," An Algorithm of Product Information Extraction from Web Pages: a Document Object Model Analysis Approach", 2nd International Conference on Information Communication and Management (ICICM 2012) IPCSIT vol. 55 ,2012, IACSIT Press, Singapore,pp.103-107.

[24] M.Walther,L.Hahne,D.Schuster,and A.Schill," Locating And Extracting Product Specifications From Producer Websites", Proceedings of the 12th International Conference on Enterprise Information Systems, Funchal, Madeira, Portugal, ISBN: 978-989-8425-07-2 , 2010, pp 13-22

[25] B.Wu, X.Cheng, Y.wang, Y.Guo, and L.Song," Simultaneous Product Attribute Name and Value Extraction from Web Pages" Web Intelligence and Intelligent Agent Technologies, WI-IAT '09. IEEE/WIC/ACM International Joint Conferences , ISBN: 978-1-4244-5331-3, Volume:3 ,pp. 295-298

[26] G.Krys,and E.Bhageri," Semi-Supervised Product Specification Extraction From The Web", Proceedings of Science and Technology Innovations, Faculty of Science and Engineering, Athabasca university,Australia,ISBN: 978-1-987973-01-3,2015,pp.105-119

[27] P.Petrovski,V.Bryl and C.Bizer, "Learning Regular Expressions for the Extraction of Product Attributes from E-commerce Microdata", LD4IE'14 Proceedings of the Second International Conference on Linked Data for Information Extraction , Aachen,Germany 2014,Volume 1267 pp. 45-54

[28] R.Ghani,K.Probst,Y.Liu,M.Krema,A.Fano," Text mining for product attribute extraction", ACM SIGKDD Explorations Newsletter, Volume 8,Issue 1,June2006,pp 41-48.

[29] D.Qiu,L.Barbosa,X.L. Dong,Y.Shen and D.Srivastava " DEXTER: LargeScale Discovery and Extraction of Product Specifications on the Web", Proceedings of the VLDB Endowment - Proceedings of the 41st International Conference on Very Large Data Bases, Kohala Coast, Hawaii , Volume 8 Issue 13, September 2015 ,pp. 2194-2205.

[30] S.Hareendran,A.Parashar,F.U.Khan," Automated Specification Extraction for Consolidated Product Catalogue", Electrical, Electronics and Computer Science (SCEECS), 2014 IEEE Students' Conference,ISBN: 978-1-4799-2525-4,March,2014,pp.1-7.

