iJRASET
International Journal For Research in
Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 6 Issue: VIII Month of publication: August 2018

DOI:

Content Aggregation based Data Extraction and Crawling from URL

Ch Prameeladevi[1], Saradadevi[2], S. K. Muthusundar[3], Mr T.Kumanan[4]
[1, 2]Research scholar, Meenakshi Academy of Higher Education and Research, K.K.Nagar, Chennai, Taminadu, INDIAD.
[3]Professor, Dept. of CSE, Sri Muthukumaran Institution of Tech., Chennai-69
[4]Professor/ CSE, Meenakshi Academy of Higher Education and Research, K.K. Nagar, Chennai, Taminadu, INDIA

Abstract: World Wide Web has transformed an unlimited source of collection. Search engines have made this message available to every Internet exploiter. There is still message available that is not easily accessible through existing operation engines remains ask to make new search engines that would present message better than before. In dictation to present mode that gives quantity, it must equally collect, analysis and transformed. This maestro thesis focuses on collection aggregation location. Mortal information extraction system is presented information that allows highly structured collection anatomy, semi-structured web pages. It complies with bulk of requirements displace for modern data extraction system: it is platform independent, it has powerful semi-automatic wrap generation system and has easy to usage someone interface for annotating structured collection. Specially designed scheme individual allows to extraction to equal performed on whole www site grade without human interaction. We display that presented tool is suitable for action highly accurate data large number of websites and can equal used as a collection origin for product aggregation system to make new measure. Mortal web page does not only comprise main textual and image content, it has also additional collection added as header, footer, area region artefact. These blocks contain direction links and sometimes advertisements. Also web page is decorated with markup which only aim is to modify visual happening. Vast of web content is generated automatically from relational databases. These include Content Management Systems like Word press, web forums and online shops. This means that collection in its original form is structured. Semantic web was designed to change web content machine readable and allowing Message shared beyond originated website. Semantic web allows to only relation inside single web page but also linkage different websites together in a meaningful way and by that creating net of knowledge. Embedding semantic message to web pages can be done indifferent ways. Most popular is micro data

Keywords: Mining, Web Scraping, Web Data Extraction, E-Commerce

I. INTRODUCTION

In nowadays modern world people are used to content on-line. Search engines get become axiomatic tools for every human, they are the entering to the Internet. Most everyday internet searches are done using search engines such as Google, and others. These universal search engines were designed to indicant entire Web as textual information. Over the years universal search engines get transform better on delivering accurate search results to users, but the results are not presented in easy mode and are not organized in comparable pattern. Therefore usually the method from old life – opening search results in new browser /window or bookmarking search results - is needed for further assessment of search results.

This has direct to initiation of vertical search engines that focus only on specific content (user goods, books, light tickets, real estate ads, scientific articles, etc.). Most well experience are SkyScanner1 that allows you search level tickets and Google Scholar2. But also goods price accumulation sites like PriceGrabber3 or Google Shopping4are getting more popular. user product search engines (aggregators) are designed to present search results in e-shop like mode, providing goods name with image and usually lowest price. Each search result link to the item page with full specification reviews that listing the Post of sales. Post of sale can be either an on-line store or physical store. For each Post of sale the price and inventory information is displayed and "acquire" button that directs user to e-shop. Bulk of these aggregators focus on price comparing ordering shops based on price. This allows users to cheapest locations where to acquire the item very easily.

In order to construct such system, detail product content must equal gathered and systematized.

One implementation to get this content is to ask manufactures or wholesale companies. Contacting and asking each organization for data is time consuming and updating these datasets is hard to proceed, if it is not done by the companies themselves. But usually no one wants additional workload and cost from their part. The second choice is to purchase this data from commercial service provider. The question with this approach is a) it's expensive for beginning companies dataset is usually limited to specific aggregation (ex. user electronics) only covers few languages. Because of these limitations the easiest manner to acquire data for start-up organization or individual is to exercise web scraping technologies to interact publicly available data from WWW Coping information from other sites might not equal seen as absolutely legal but document does not assist factual aggregation and goods description can equal considered as factual collection. In item analysis of legal aspects can equal found, this system presents modern papers for medium budget and structured data extraction from all semi structured web pages. First parts gives existing work in the area of Web data mining and is divided into two separate topics are web data extraction and web crawling. Next part describes the implementation of built system followed with assessment of the system together with communicating on last part is future work ideas are presented.

II. OVERVIEW OF WEB DATA EXTRACTION METHODS

Web search engines have been present more than two generations and get evolved from simple full text search engines into complex systems that analyses web page content together with links that level into them. But still even most modern web search engine has three main components Web crawling and data use and extracting link and text data, Data storage and online processing Data earnings and aggregation and modify search index artifact, Query processing search index and ranking results based of search status. As mentioned in introduction, this work will direction on web content extraction, therefore only web crawling and data acquisition is covered in this area. Other portion (data aggregation, product classification task, reduplicate matching) required to construct fully functional goods search engine are not discussed as these are implemented as separate organization

In next subsections we look content extraction and web crawling in more detail.

A. Semantic Web

Semantic web was designed to change web content machine readable and allowing Message shared beyond originated website. Semantic web allows to only relation inside single web page but also linkage different websites together in a meaningful way and by that creating net of knowledge. Embedding semantic message to web pages can be done indifferent ways. Most popular is micro data which **is** shown in Figure1 . Although Schema.org is becoming de facto standard to define schemas for semantic web it is not necessary to activity their schema. But consolidating schemas makes their use easier and allows spreading. For commercial products GoodRealations schema was created by Martin Hepp and is now partially merged into Schema.org. When defining your own schema it is advisable to usage as it will be persistent over time.

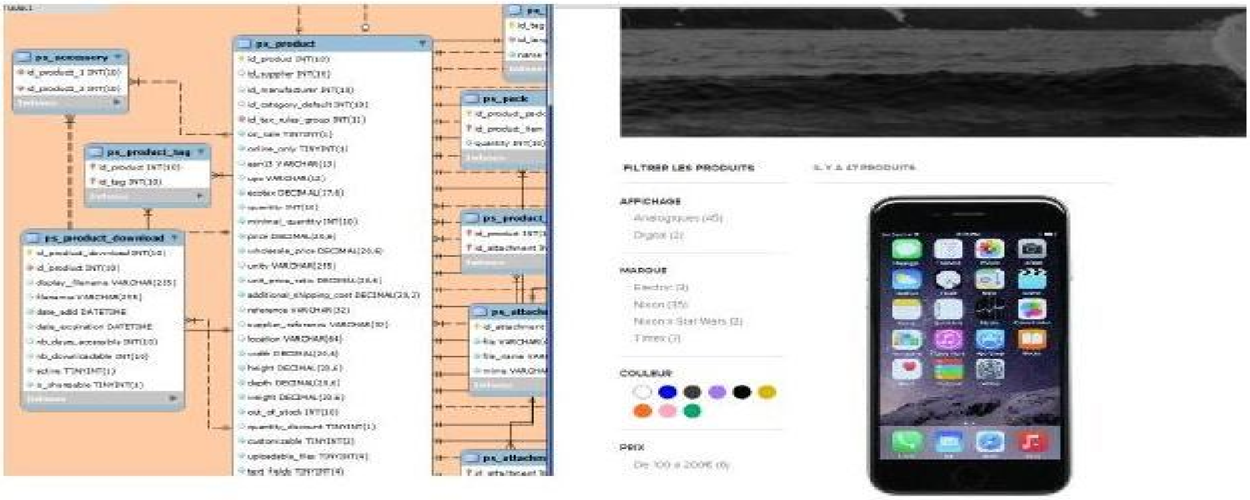


Figure 1 : Poorvika Database Schema and mobile online shopping store

Together with linked data, semantic element were introduced in HTML5 standard. These are element like <article>, <figure>, <header>, <nav> and others. These were designed to supply meaning to website structure and regenerate HTML scheme like <div class="header">. As semantic web is getting more widespread (17% of domains crawled by commons Crawl used semantic annotation) and with growing of differents API it is becoming questionable, to we ask data extraction from fuzzy semi structured webpages and instead focusing on structured data extraction and its meaningful investigation . But unfortunately,other ways to include semantic and JSON-LD (JSON Linked collection) embedding's. at current time (2016) we still can't be without extracting data from non semantic web pages. Even when semantic direction is used, it is often used partially and describes only portion of data made available. Also syntactical mistakes are common.

B. Web data extraction

Web pages comprise messy data. Mortal web page does not only comprise main textual and image content, it has also additional collection added as header, footer, area region artefact. These blocks contain direction links and sometimes advertisements. Also web page is decorated with markup which only aim is to modify visual happening. Vast of web content is generated automatically from relational databases. These include Content Management Systems like Wordpress, web forums, online shops. This means that collection in its original form is structured. This structured message is embedded into HTML example and decorated with visual content. Figure shows underlying database schema (left) of popular open source e-commerce resolution and timefy.com online shop. In order to acquire original data, the example must be removed or data must dening your own schema it is advisable to use
Together with linked data, semantic element were introduced in HTML5 standard. These are element like <article>, <figure>, <header>, <nav> and others. These were designed to provide meaning to website construction and regenerate HTML syntax like <div class="header">. As semantic web is getting more widespread (18% of domains crawled by Common Crawl used semantic annotation) and with growth of dierent API-s it is becoming questionable, to we demand data extraction from fuzzy semi structured webpages and instead concentration on structured data extraction and its meaningful analysis . But unfortunately,
at current time (2016) we still can't reside without extracting data from non semantic web pages. Even when semantic markup is used, it is often used partially and describes only portion of data made available. Also syntactical mistakes are common. data must wrapper is a Technic that implements a family , that finds the message that individual needs, extracts this from an unstructured origin and change them into structured data. Over the last generation large quantity of research and tools have been created that support with web mining project.product studies have been performed to analyse existing solutions, these include and more recent overview about web mining including web data extraction[11]. Together with web content mining some of these surveys give summary also about other parts of web mining, such as web structure mining and web utilization mining.
Data extraction methods can be divided into three segments depending on the level of automation[]: Manual approach: Human observers web page and its source code and writes
down rules or software code extract data. Also tools that change the activity simpler for programmers, such as pattern specication languages and human interfaces are placed to this segment. Wrapper stimulation: In this approach supervised learning where extraction rules
are learned from a manually labeled collection records. Automatic action: This uses unsupervised learning to find repetitive patterns on single or multiple pages. As this is fully automatic, then it can remain applied in web standard.
Although it is generally accepted, that manual approach is not scalable to large amount of sites, it can be viewed as labeling or notation task for Wrapper elicitation or to generate experiment dataset for automatic extraction establishment as it is stone gives most accurate results. And with visual assist tools this process can be signicantly accelerate up when compared to manually creating extraction rules. This approach also plant well, if we have highly structured data, meaning that we can create wrapper by only observing single web page per site.

1) *Expressing Extraction Rules:* Web pages can be treated as just as movement of characters, structured data where fields are separated using tokens (HTML tags) or as Document Object Model (DOM) that represents HTML page in tree structure. When using HTML page as text representation we can exercise regular expressions (regex) or other standard text extraction methods to take data we involve. For example extracting nonfiction headline we can exercise following rules (Figure 3). When we appear web page as DOM tree, then we can exercise XPath to express the extraction rules (see Figure). Using XPath makes rules simple but we also loose knowledge to obtain partial content inside HTML component.

When creating extraction rules we must assure that they are general enough to extract data from all pages and specific to only extract data that this concept is designed for. For example when HTML markup for single page contains multiple element1 tags then we must change the rule less general to locate the eld with higher precision. Several programming languages or module extensions has made to create manual approach faster for technologist.
The intensity of web crawling lies in the information that it covers all the necessary bases when it comes to direct reproduction. Data is harvested, structured, categorized and organized in such a manner that businesses can easily exercise the data provided for their marketing leads. As discussed earlier, cold and detached lists no longer furnish you with enough actionable leads. You ask to appear at various factors and contemplate them during your direct generation efforts are

Interaction details of the expectation
Purchasing quality and purchasing history of the potential
Past purchasing trends, disposition to purchase and past of buying preferences of the potential
Social markers that are indicative of behavioral patterns
Commercial and business markers that is indicative of behavioral patterns, transactional details
Other factors including years, sex, sociology, social circles, communication and interests

All these factors demand to be taken into account and considered in detail if you get to guarantee whether a lead is viable and actionable, or not. With web scraping you can acquire enough data about every single potential, associate all the data collected with the assist of onboarding, and determine with condemnation whether a particular potential will be viable for your business.

REFERENCE

[1] Robert Baumgartner, Thomas Eiter, Georg Gottlob, Marcus Herzog, and Christoph Koch.Information Extraction for the Semantic Web. In Norbert Eisinger and Jan Maªuszy«ski, editors, Reasoning Web, number 3564 in Lecture Notes in Computer Science, pages 275289. Springer Berlin Heidelberg, 2005. DOI: 10.1007/11526988_8.

[2] ValterCrescenzi, Paolo Merialdo, and Paolo Missier.Clustering Web Pages Based on Their Structure. Data Knowl. Eng., 54(3):279299, September 2005.

[3] Chia-Hui Chang, Mohammed Kayed, MohebRamzyGirgis, and Khaled F. Shaalan. ASurvey of Web Information Extraction Systems.IEEE Trans. on Knowl.and Data Eng., 18(10):1411 1428, October 2006

[4] RuiCai, Jiang-Ming Yang,Wei Lai, YidaWang, and Lei Zhang. iRobot: An Intelligent Crawler forWeb Forums. In Proceedings of the 17th International Conference on World Wide Web,WWW'08, pages 447456, New York, NY, USA, 2008. ACM.

[5] Jefirey Dean and Monika R. Henzinger. Finding Related Pages in theWorld Wide Web. In Proceedings of the Eighth International Conference on World Wide Web, WWW '99, pages 14671479, New York, NY, USA, 1999. Elsevier North-Holland, Inc.

[6] Emilio Ferrara and Robert Baumgartner. Intelligent Self-repairable Web Wrappers. In Roberto Pirrone and FilippoSorbello, editors, AI*IA 2011: Articial Intelligence Around Man and Beyond, number 6934 in Lecture Notes in Computer Science, pages 274285. Springer Berlin Heidelberg, September 2011. DOI: 10.1007/978-3-642-23954-0_26.

[7] Emilio Ferrara, Pasquale De Meo, GiacomoFiumara, and Robert Baumgartner.Web data extraction, applications and techniques: A survey. Knowledge-Based Systems, 70:301 323, November 2014.

[8] TimFurche, Georg Gottlob, Giovanni Grasso, Giorgio Orsi, Christian Schallhart, and Cheng Wang.AMBER: Automatic Supervision for 30 Multi-Attribute Extraction. arXiv:1210.5984 [cs], October 2012. arXiv: 1210.5984.

[9] Tomas Grigalis and AntanasCenys.Using XPaths of inbound links tocluster template-generated web pages. Computer Science and InformationSystems, 11(1):111131, 2014.

[10] Tomas GRIGALIS. STRUCTURED DATA EXTRACTION FROM TEMPLATE-GENERATED WEB PAGES.DOCTORAL DISSERTATION, VILNIUS GEDIMINAS TECHNICAL UNIVERSITY, Vilnius, 2014.

[11] K. Kanaoka, Y. Fujii, and M. Toyama. Ducky:A data extraction system for various structured web documents. In ACM International Conference Proceeding Series, pages 342347, 2014.

[12] Kei Kanaoka and Motomichi Toyama.Effective Web Data Extraction with Ducky. In Proceedings of the 19th International Database Engineering &#38; Applications Symposium, IDEAS '15, pages 212213, New York, NY, USA, 2014. ACM.

[13] Bing Liu. Web Data Mining: Exploring Hyperlinks, Contents, and Usage Data. Springer-Verlag New York, Inc., Secaucus, NJ, USA, 2011.

[14] Zaki, J. X. Yu, B. Ravindran, and V. Pudi,editors, Advances in Knowledge Discovery and Data Mining, Pt Ii, Proceedings, volume 6119, pages 222 229. Springer-Verlag Berlin, Berlin, 2010. WOS:000281629400022.

[15] Sean O'Reilly. Nominative Fair Use and Internet Aggregators: Copy- right and Trademark Challenges Posed by Bots, Web Crawlers and Screen- Scraping Technologies [article]. Number 3. 2006. TY: GEN; ID: Accession Number: hein.journals.lyclr19.20; Item Citaton: Loyola Consumer Law Review, Vol. 19, Issue 3 (2007), pp. 273-288,.

[16] K. Pol, N. Patil, S. Patankar, and C. Das.A Survey on Web Content Mining and Extraction of Structured and Semistructured Data.In First International Conference on Emerging Trends in Engineering and Technol- ogy, 2008. ICETET '08, pages 543546, July 2008.

[17] Andrew Sellers, Tim Furche, Georg Gottlob, Giovanni Grasso, and Christian Schallhart. Taking the OXPath Down the Deep Web. In Proceedings of the 14th International Conference on Extending Database Technology, EDBT/ICDT '11, pages 542545, New York, NY, USA, 2011. ACM.

[18] Andrew Jon Sellers, Tim Furche, Georg Gottlob, Giovanni Grasso, and Christian Schallhart. OXPath: Little Language, Little Memory, Great Value. In Proceedings of the 20th International Conference Companion on World Wide Web, WWW '11, pages 261264, New York, NY, USA, 2011. ACM.

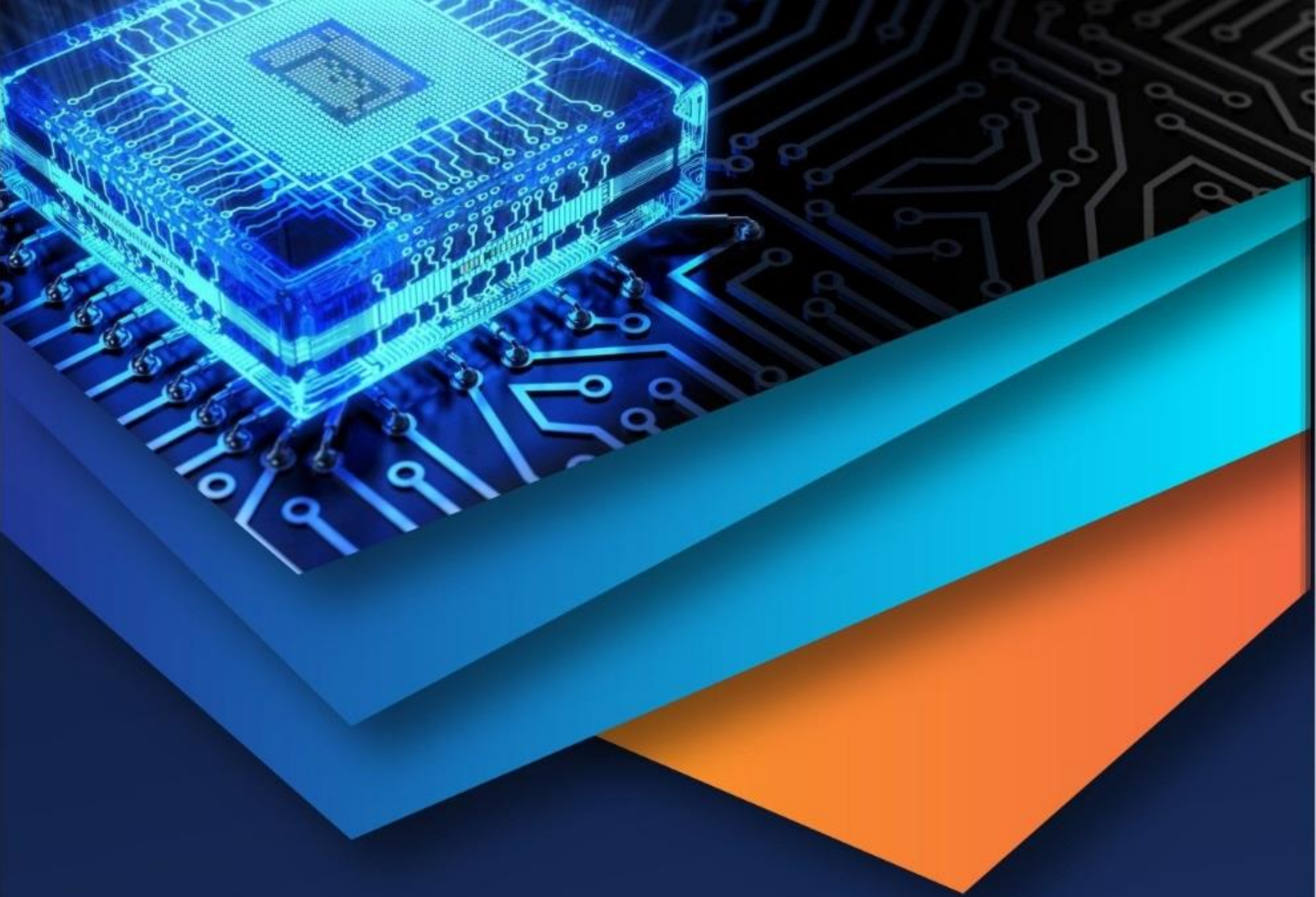
[19] Shengsheng Shi, Wu Wei, Yulong Liu, Haitao Wang, Lei Luo, Chunfeng Yuan, and Yihua Huang. NEXIR: A Novel Web Extraction Rule Language toward a Three-Stage Web Data Extraction Model. In Xuemin Lin, YannisManolopoulos, DiveshSrivastava, and Guangyan Huang, editors, Web Information Systems Engineering WISE 2013, volume 8180 of Lecture Notes in Computer Science, pages 2942. Springer Berlin Heidelberg, 2013. 32

[20] Maarten Truyens and Patrick Van Eecke.Legal aspects of text mining. Computer Law & Security Review: The International Journal of Technology Law and Practice, 2014.

[21] Juan D. Velásquez.Web mining and privacy concerns: Some important legal issues to be consider before applying any data and information extraction technique in web-based environments. Expert Systems with Appli- cations, 40(13):52285239, October 2013.

[22] Tim Weninger, Rodrigo Palacios, ValterCrescenzi, Thomas Gottron, and Paolo Merialdo.Web Content Extraction - a Meta-Analysis of its Past and Thoughts on its Future.arXiv:1508.04066 [cs], August 2015. arXiv: 1508.04066.

[23] T Yang and AGerasoulis.Web Search Engines: Practice and Experience. In Computer Science Handbook. Chapman & Hall/CRC Press, 3rd edition, 2014.



10.22214/IJRASET

45.98

IMPACT FACTOR:
7.129

IMPACT FACTOR:
7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY