Recommendation Generation on Extracted Web Data Using Cosine Similarity

Asst. Prof. Priyanka Patil
Department of Computer Engineering
Atharva College of Engineering
Mumbai, India

Asso. Prof. Anuradha G.
Department of Computer Engineering
St. Francis Institute of Technology
Mumbai, India

***Abstract*— A online databases can be queried and/or updated through World Wide Web (WWW). Online databases are also called as web databases. Online databases generate result pages in the form of response to queries posted by users. Such Result pages are called as Resulted Query pages (QRPs). Nowadays, many analytical applications necessitate the automatic extraction of data from these query result pages. The extracted result from query result pages is very important for many web applications. This is because the QRP is a consolidation of related data which is in fields of database. While trying to extract QRPs initially the webpage data which is in the HyperText Markup Language (HTML) format is converted to Extensible Markup Language (XML) format using web harvest tool. The data is in XML format is semi-structured or non-structured data. The alignment of this non-structured data into structured format is done using cosine similarity approach. Data records which are aligned are used for text mining purposes.**

Keywords*— Cosine Similarity, Data Extraction, Recommendation.

I. INTRODUCTION

Internet and the Web have revolutionized access to information. Today, one finds primarily on the Web, HTML (the standard for the Web) but also documents in pdf, doc, plain text as well as images, music and videos. The public Web is composed of billions of pages on millions of servers. It is a fantastic means of sharing information. It is very simple to use for humans. On the negative side, it is very inappropriate for accesses by software applications. This motivated the introduction of a semi structured data model, namely XML, which is well suited both for humans and machines. data alignment. Data extraction is implemented using web extraction tool and for Data alignment string cosine similarity algorithm is implemented. Upon the aligned or structured data recommendation is given to the user.

Data Extraction:-

Data extraction is the process of retrieving data from unstructured or poorly structured data sources for further data processing or data storage. The majority of data extraction comes from unstructured data sources and different data formats. This unstructured data can be in any form, such as tables, indexes, and analytics.[5] After receiving a user's query, a web database which is semi structured database returns the relevant data values, in structured format. Many web applications need the data from multiple web databases. For these applications to further utilize the data embedded in HTML pages, automatic data extraction is necessary. Once data values are extracted and organized in a structured manner, such as tables, they can be compared and aggregated. Hence, accurate data extraction is vital for these applications. [1].

Web Data Extraction:-

Web Data extraction is the process of retrieving unstructured data from web pages and importing it into a structured data system like a database. Process of extracting data from Web pages is also referred as Web Scraping or Web Data Mining. [5]

Common Problem with Web Data Extraction:-

- Incapable of processing with zero query results they require at least two records in a query result page.
- Vulnerable to optional and disjunctive attributes. It is true for those attributes which are not connected with each other or inconsistence in nature, such kind of attributes can cause data alignment problem.
- Incapable of processing nesting data structures many methods can only process a flat data structure and fail for a nested data structure.

II. PROBLEM DEFINITION

Although the web data can be extracted through the web using wrappers, the resultant data is not fully relevant or really expected by the user. Therefore, it is necessary to have an information-searching or web data extraction system which is capable of extracting the relevant data expected by the user [30].

This research proposes a data extraction methodology that extracts web data using web extraction tool. After extraction, data is aligned using pair wise algorithm depending on every attribute's cosine similarity with another attribute. The proposed system is based on online data extraction and alignment. Till now data extraction is done using different type's wrappers and also for data alignment various algorithms are being used. So, following research mainly focuses on finding an easy and faster way to perform data extraction using extraction tool and data alignment is implemented using a single algorithm which gives better record alignment with this a recommendation is provided to the user using cosine similarity.

III. REVIEW OF LITERATURE

Web database extraction is gaining popularity among the database and information extraction research areas in recent years due to the volume and quality of deep web data. As the returned data from a query are embedded into HTML pages, the current research areas are focused on the extraction of this data. Earlier work focused on wrapper induction methods, which require human assistance to build a wrapper.[1]

a) Current Wrappers:-

Current automatic wrappers except Ontology-assisted Data Extraction (ODE) wrapper [26]) do not use the semantic properties of data records in their design. Earlier automatic wrappers extract data records by checking the patterns inherent in data records using the DOM tree structure. For example, Mining Data Region (MDR) [11] checks the repetitive HTML tags in order to locate data records, while EXALG uses equivalence classes (template which represents all the data records) to determine the template of data records before data extraction is carried out.

b) Ontology-based wrapper:-

ViPER [19] on the other hand, enhances the algorithm of MDR by using primitive tandem repeat, which detects the repetitive sequence of HTML tags using a matrix. VSDR [21] extracts search engine results pages by detecting the centrally located data region in a web page. DeepMiner [27] uses domain ontology to mark up web services. DeepMiner uses the data collected from the query web pages to generate the domain ontology. Recently, ODE wrapper [26] used ontology technique to extract, align, and annotate data from search engine results pages. However, ODE requires training data to generate the domain ontology. ODE is also only able to extract a specific type of data record (single-section data records). Thus, it is not able to extract irregular data records, such as multiple-sections data records and loosely structured data records.

c) Data Alignment:-

Simon and Lausen [19] proposed Multiple Sequence Alignment (MSA) algorithm to align data based on Maximal Unique Matches (MUM). MSA could efficiently align data records in a polynomial time complexity, but to find the MUM requires extensive checking on the DOM tree structure. MUM is also not suitable to represent data, which is iterative. MSA also assumes that if a MUM is created, it may contain more than one text elements. Text nodes in a HTMLDOM tree are atomic entities, and therefore, they should be considered as separate entities when used for data alignment.

IV. PROPOSED WORK

The proposed work extracts QRR from QRPs in an automated fashion. The data extracted is in an unstructured format. The unstructured data is aligned using pair wise mechanism. The system is divided into following modules.

a) Data Extraction module:

The extraction module mainly focuses on the web data extraction. Where, first user submits a query to the response of that query a single web page gets selected. It is also called as Query Result Page (QRP).Extraction is done using a former web extraction tool. i.e. Web Harvest Tool which gives output in the form of XML document. The data in the XML document is in unstructured or semi-structured format.

b) Data alignment module:

The extracted data records are called as Query Result Record(QRR).QRRs are then aligned with the help of pair wise alignment algorithm based on cosine similarity where, accuracy of the result depends upon higher cosine similarity value.

c) Recommendation module:

Once the QRR are aligned in a structured format they can be use for further analysis. QRR can be used to give recommendation to the user. The recommendation is personalized for a particular user. It is based on the user's likes for a product. Depending on the user's like that products cosine value is compared with the other remaining product's value and on bases of their similarity measures final list is generated and given to user as a recommendation.

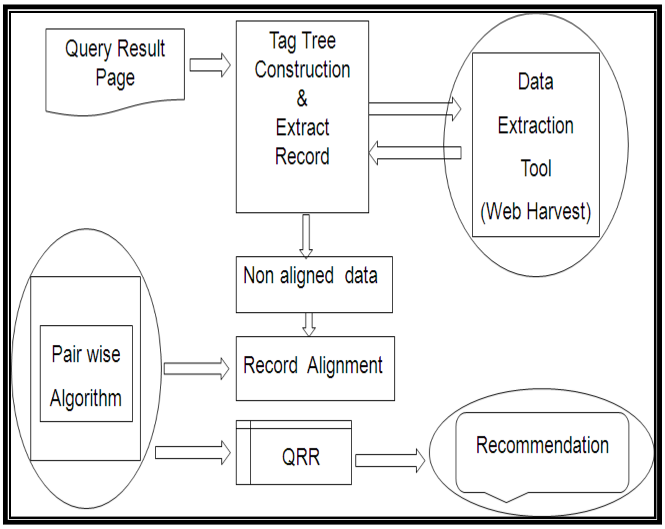


Fig4.1.Proposed System

Fig 4.1 shows proposed system for this research work. It includes detail flow of the system. The proposed system extracts data from web and aligns it into a tabular format. After alignment of data it can be used for analysis purpose.

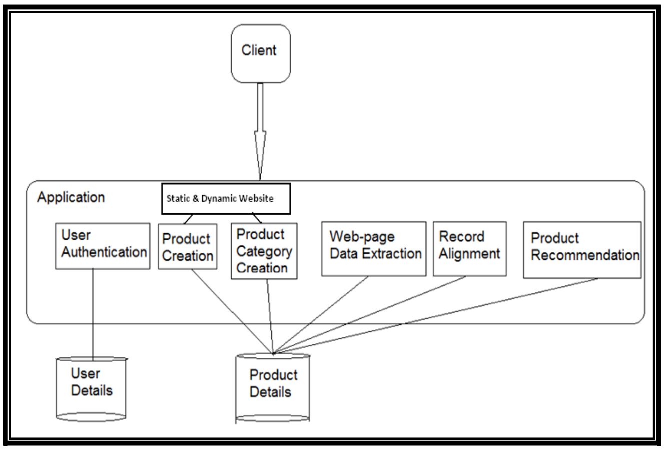


Fig 4.2 System Architecture

System architecture of the proposed work is as shown in Figure. 4.2. The system is implemented for both static and dynamic websites.

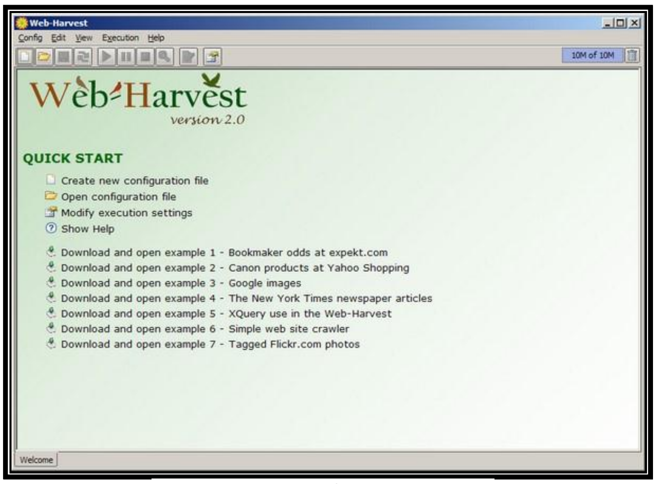
Data Extraction

In the proposed system data extraction is implemented using web harvest tool which makes data extraction process efficient and fast. The steps involved in Query result record (QRR) are as follows:

Web harvest tool:-

Web data on existing websites is mainly formatted in unstructured HTML, even though flexible markup languages. e.g. XML, XHTML, are attracting a lot attention recently. Moreover, HTML is mainly used for presentation of data. While XML provides more suitable data representation by separating data structure from its layout. Imagine a set of XML documents are already in well structured format that is why it can be regarded as a database and can be directly processed by a database application. But, we also know querying relevant data from unstructured HTML content spent a huge amount of time and cost. That is why we need a tool, in order to implement web data extraction.

In this study, E-commerce websites are used for extraction following are some details about their tab structure. [41].



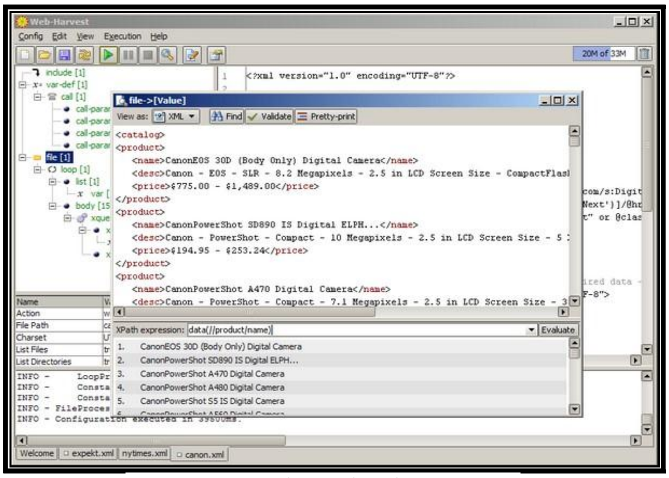
QUICK START

- Create new configuration file
- Open configuration file
- Modify execution settings
- Show Help

- Download and open example 1 - Bookmaker odds at expekt.com
- Download and open example 2 - Canon products at Yahoo Shopping
- Download and open example 3 - Google images
- Download and open example 4 - The New York Times newspaper articles
- Download and open example 5 - XQuery use in the Web-Harvest
- Download and open example 6 - Simple web site crawler
- Download and open example 7 - Tagged Flickr.com photos

Fig 4.3 Home page of Web Harvest tool

Figure 4.3 shows the startup page of web harvest tool. This tool is used for web data extraction. In general, webpage contain data in HTML format this web data is converted into XML format using web harvest tool. To perform such an operation it uses XQuery and XSLT language.



```
<?xml version="1.0" encoding="UTF-8"?>
```

```
<catalog>
<product>
   <name>CanonEOS 30D (Body Only) Digital Camera</name>
   <desc>Canon - EOS - SLR - 8.2 Megapixels - 2.5 in LCD Screen Size - CompactFlasl
   <price>$775.00 - $1,489.00</price>
</product>
<product>
   <name>CanonPowerShot SD890 IS Digital ELPH...</name>
   <desc>Canon - PowerShot - Compact - 10 Megapixels - 2.5 in LCD Screen Size - 5 :
   <price>$194.95 - $253.24</price>
</product>
<product>
   <name>CanonPowerShot A470 Digital Camera</name>
   <desc>Canon - PowerShot - Compact - 7.1 Megapixels - 2.5 in LCD Screen Size - 3
```

XPath expression: data(//product/name)

1. CanonEOS 30D (Body Only) Digital Camera
2. CanonPowerShot SD890 IS Digital ELPH...
3. CanonPowerShot A470 Digital Camera
4. CanonPowerShot A480 Digital Camera
5. CanonPowerShot S5 IS Digital Camera

```
INFO -      LoopPr
INFO -      Consta
INFO -      Consta
INFO - FileProces
INFO - Configuration executed in 3980ms.
```

Fig 4.4 Example of XML file after conversion

Figure4.4 shows XML file format which has been provided by web harvest tool after conversion of web data from a webpage.

The structure of an HTML Documents for E commerce Website Design:

These E commerce Website Design tags are usually laid out in a certain order. HTML is quite flexible in general, but there are two tags that almost all documents need to have: the head and the body. As part of E commerce Website Design the head should contain information about the document, as well as any scripts or style sheets the E commerce Website Design is using, while the body should contain the main text of the document. So, as an example, simple E commerce Website Design might look like this:

For example,

```
<html>
<head>
<title>my page</title>
<style>body { background-color: blue; }</style>
</head>
<body>
<p>some text</p>
</body>
</html>
```

The first thing to notice as part of the E commerce Website Design is the way it starts and ends: with the HTML tag. This is essential when you complete E commerce Website Design. Now, notice what's included in the head and what's in the body: while the head tells you the title of the page and that its background color is blue, it's the body that has the web page's text.

Data Alignment

Data alignment is implemented using pair wise algorithm. Pair wise algorithm used cosine similarity measures. They are totally depended on the similarity measure between two vectors. Here, two different data records are compared for their similarity. For those similarity and dissimilarity measures are as follows:

Similarity Measures:

Many data mining and analytics tasks involve the comparison of objects and determining in terms of their similarities (or dissimilarities)

Many of today's real-world applications rely on the computation similarities or distances among objects

- Recommender systems
- Document categorization
- Information retrieval

Similarity and Dissimilarity:

Similarity:-

- Numerical measure of how alike two data objects are

- Value is higher when objects are more alike
- Often falls in the range [0,1]

Dissimilarity:-

- Numerical measure of how different two data objects are
- Lower when objects are more alike
- Minimum dissimilarity is often 0

Vector Based Similarity Measures:

When the data is very sparse and 0's in the vectors are not significant

In such cases, typically vector-based similarity measures are used

Most common measure: Cosine similarity

X=(x1,x2,….,xn) (1)

Y=(y1,y2,….yn) (2)

The cosine similarity is:

$$\mathrm{Sin}(X,Y) = X.Y / \|X\| \times \|Y\| \quad (3)$$

$$= \frac{\overset{i}{\Sigma}(x_i \times y_i)}{\sqrt{\Sigma x_i^2} \times \sqrt{\Sigma y_i^2}} \quad (4)$$

For example,

Consider there are two data records Row 1 and Row 2. Following steps will give you cosine similarity of both the vectors.

Row 1= <3, 1, 4, 3, 1, 2, 0, 1>

Row 2= <0, 1, 0, 3, 0, 0, 2, 0>

Dot-Product (Row 1, Row 2)

= <3,1,4,3,1,2,0,1> * <0,1,0,3,0,0,2,0>

= 0 + 1 + 0 + 9 + 0 + 0 + 0 + 0 = 10

Norm (Row 1) = SQRT (9+1+16+9+1+4+0+1) = 6.4

Norm (Row 2) = SQRT (0+1+0+9+0+0+4+0) = 3.74

Cosine (Row 1, Row 2) = 10 / (6.4 * 3.74) = 0.42

Cosine similarity= 0.42

In first step dot product of Row 1 and Row 2 is calculated as shown in equation 2. In next step normalize both the vector as shown in equation 3. As shown in equation 4 cosine similarity formula applied on normalized vectors.

Recommendation Module:-

In recent years we see the continuing growth of the Internet. Not only is the number of Internet users and websites increasing, but also the amount of information on the individual websites. Many websites are concerned with presenting their often very semantically versatile information in a concise and efficient way. This is especially true for large E-Commerce websites with large amount of product information.

The motivation for the use of web recommendations comes from both Internet users and website owners. Internet users want to see interesting information; the website owners want their information to reach users quickly and to the full extent. Owners of commercial websites also employ web recommendations in order to sell additional products or services to the users and thus increase the sales turnover of their websites. Many algorithms have been developed in order to generate such potentially interesting web recommendations automatically. These approaches are based on different intuitions about what might be interesting for the given user in a given situation.

Cosine Similarity formula for recommendation:

$$\text{Cosine value} = \frac{\text{No. of Common Terms} * 100}{\sqrt{(\text{Terms in string 1(keywords)})* (\text{Terms in string 2 (keywords)})}}$$

= __ % calculated

= value will be in between (0 to 1)

V. CONCLUSION

The proposed approach has three modules. The first and the most important step is preprocessing. For preprocessing i.e. extraction of web data an open source tool is used. Using web harvest tool data is extraction can be implemented. The proposed system is for automatic extraction of structured query results from a deep web page. It is able to eliminate all the auxiliary information in the web page. This system extracts the result using both the tag and value information which the existing system does not do. This gives more accurate results and the alignment is made easier. Better Alignment gives better Query performance. This system is very useful in many web applications that need data from multiple websites in less time. For Example, Marketing Sites, E-commerce Sites. Proactive analysis helps to give Recommendation.

ACKNOWLEDGMENT

I would like to express my deep gratitude to Associate Professor Anuradha G., my research supervisors, for their patient guidance, enthusiastic encouragement and useful critiques of this research work. I would also like to extend my thanks to the authors. Finally, I wish to thank my parents for their support and encouragement throughout my study.

REFERENCES

[1] Weifeng Su, Jiying Wang, Frederick H. Lochovsky, Member, IEEE Computer Society, and Yi Liu "Combining Tag and Value Similarity for Data Extraction and Alignment" *IEEE Transactions on knowledge and data engineering*, vol. 24, no. 7, july 2012..

[2] Weifeng Su, Jiying Wang, Frederick H. Lochovsky, Member, IEEE Computer Society, and Yi Liu "Combining Tag and Value Similarity for Data Extraction and Alignment" *IEEE Transactions on knowledge and data engineering*, vol. 24, no. 7, july 2012.

[3] Mohammad Shafkat Amin Hasan Jamil, "FastWrap: An Efficient Wrapper for Tabular Data Extraction from the Web," *IEEE IRI* 2009, July 10-12, 2009, Las Vegas, Nevada, USA

[4] M.K. Bergman, "The Deep Web: Surfacing Hidden Value," White Paper, BrightPlanet Corporation, http://www.brightplanet.com/resources/details/deepweb.html, 2001.

[5] K.C.-C. Chang, B. He, C. Li, M. Patel, and Z. Zhang, "Structured Databases on the Web: Observations and Implications," *SIGMOD Record*, vol. 33, no. 3, pp. 61-70, 2004

[6] P.V.Praveen Sundar Research Scholar"Towards Automatic Data Extraction Using Tag and Value Similarity Based on Structural - Semantic Entropy "*International Journal of*

[7] Lukáš Bainka, Ivan Jelínek, "Data Extraction by Visual Matching", International Conference on Computer Systems and Technologies - CompSysTech'09.

[8] D. Chamberlin and al. (Eds.), "XQuery: A query language for XML", http://www.w3.org, 2001.

[9] Web Data Extraction. [Online] Available:http://www.automationanywhere.com/solutions/webDataExt, 10 August 2013.

[10] A.Budanitsky and G.Hirst, "Semantic distance in wordnet: An experimental, application-oriented evaluation of five measures," in Proc. NAACL, 2001, pp. 29–34.

[11] B. Liu, R. Grossman,, and Y. Zhai, "Mining data records in Web pages, "presented at the ACM SIGKDD *Conf*., Washington, DC, 2003.

[12] C. Fellbaum, WordNet: An Electronic Lexical Database. Cambridge, MA: MIT Press, 1998.

[13] C. Leacock and C. Martin, Combining Local Context and WordNet Similarity for Word Sense Identification. Cambridge,MA: MIT Press, 1998.

[14] D. W. Embley, D. M. Campbell, Y. S. Jiang, S. W. Liddle, and D. W. Lonsdale, "Conceptual-model-based data extraction from multiple-record Web pages"*Elsevier Science Publishers B. V. Amsterdam, The Netherlands, The Netherlands.*Volume 31 Issue 3, Nov. 1999 Pages 227 - 251

[15] G. Hirst and D. St-Onge, Lexical Chains as Representations of Context for the Detection and Correction of Malapropisms. Cambridge, MA: MIT Press, 1998.

[16] H. Snoussi, L.Magnin, and J.-Y.Nie, "Heterogeneous web data extraction using ontology," in Proc. Agent-Oriented Inf. Syst., 2001, pp. 99–110.ges," Data Know. Eng., vol. 31, pp. 227–251, 1999.

[17] H. Zhao,W.Meng, Z.Wu,V.Raghavan, and C.Yu, "Fully automatic wrapper generation for search engines," in Proc. ACMWWW, 2005, pp. 66–75.

[18] J. J. Jiang and D. W. Conrath, "Semantic similarity based on corpus statistics and lexical taxonomy," in Proc. Int.Conf. Res.Comput. Linguist., 1997, pp. 19–33.

[19] K. Simon and G. Lausen, "ViPER: Augmenting automatic information extraction with visual perceptions," presented at the ACM CIKM Conf., Bremen, Germany, 2005.

[20] O. Lassila and D. Mc Guinness, "The role of frame-based representation on the semantic web," Know. Syst. Lab., Stanford Univ., Stanford, CA, Tech. Rep. KSL-01-02, 2001.

[21] L. Li, Y. Liu, A. Obregon, and M. A. Weatherston, "Visual segmentation based data record extraction from web documents," in *Proc. IEEE IRI*, 2007, pp. 502–507.

[22] P.W. Lord, R.D. Stevens, A. Brass, and G. CA, "Investigating semantic similarity measures across the gene ontology: the relationship between sequence and annotation," Bioinformatics, vol. 19, pp. 1275–1283, 2003.

[23] R. Rada, H. Mili, E. Bicknell, and M. Blettner, "Development and application of a metric on semantic nets," *IEEE Trans. Syst., Man, Cybern.*, vol. 19, no. 1, pp. 17–30, Jan./Feb. 1989.

[24] M. Rodriguez and M. Egenhofer, "Determining semantic similarity among entity classes from different ontologies," *IEEE Trans. Knowl. Data Eng.*,vol. 15, no. 2, pp. 442–456, Mar./Apr. 2003.

[25] W. Liu, X. Meng, and W. Meng, "ViDE: A vision-based approach for deep web data extraction," *IEEE Trans. Knowl. Data Eng.*, vol. 22, no. 3, pp. 447–460, Mar. 2009

[26] W. Su, J. Wang, and F. H. Lochovsky, "ODE: Ontology-assisted data extraction," ACM Trans. Database Syst., vol. 34, no. 2, pp. 1–35, 2009.

[27] W.Wu,A.Doan, C.Yu, andW. Meng, "Bootstrapping domain ontology for semantic web services from source web sites," in Proc. VLDB Workshop, 2005, pp. 11–22.

[28] Z. Wu and M. Palmer, "Verb semantics and lexical selection," in Proc. Annu. Meet. Assoc. Comput. Linguist, 1994, pp. 133–138.

[29] Y. Zhai and B. Liu, "Web data extraction based on partial tree alignment," in Proc. ACM WWW, 2005, pp. 76–85.

[30] Rudy AG. Gultom ,RiriFitri, Sari BagioBudiardjo,"Implementing Web Data Extraction and Making Mashup with Xtractorz" *IEEE 2nd International Advance Computing Conference*, 2010.

[31] CYC Wrapper. [Online]. Available: http://www.cyc.com, 15 August 2013.

[32] DOLCE Wrapper. [Online]. Available: http://www.loa-cnr.it/DOLCE.html, 15August 2013.

[33] SUMO Wrapper. [Online]. Available: http://www.ontologyportal.org/, 15 August 2013.

[34] Pattern Matching and Structured Matching. [online]. Available: http://en.wikipedia.org/wiki/, 25 August 2013.

[35] Web databases. [Online]. Available: http://www.keithjbrown.co.uk/vworks/php/php_p1.php, 25 August 2013.

[36] Surface web and Deep web. [Online]. Available: http://www.metronomegazette.com/2013/11/surfing-deep-web.html, 25 August 2013.

[37] Wrapper induction method. [Online]. Available:

[38] http://www.isi.edu/integration/Mercury. 1st Septmber 2013.

[39] Robinson, G. B., U.S. Patent No. 5,884,282. Washington, DC: U.S. Patent and Trademark Office, 1999.

[40] Sarwar, B., Karypis, G., Konstan, J., &Riedl, J., "Item-based collaborative filtering recommendation algorithms," In Proceedings of the 10th international conference on World Wide Web, 2001, pp. 285-295. ACM

[41] "Web Harvest Tool", http://sourceforge.net/projects/web-harvest/ 25th September2013.