www.ijecs.in
International Journal Of Engineering And Computer Science ISSN: 2319-7242
Volume 5 Issue 1 January 2016, Page No. 15616-15622

Sequential Implementation of Web Content Mining Using Data Analysis in Online Sales Domain

Dr.S.P.Victor*, Mr. M. Xavier Rex

Associate Professor CS, St.Xaviers College, Tirunelveli,drspvictor@gmail.com

Research Scholar M.S.University, Tirunelveli.

Abstract-In the web data mining retrieving and storage analysis of web content plays a vital role in online sales domain which provides the systematic way of novel implementation towards real-time data with different level of implications. Our experimental setup initially focuses with retrieval of web content. This paper perform a detailed study of web content retrieval schema towards variant effect of periodic web page content in the field of online sales marketing domain which can be carried out with expected optimal output strategies. We will implement our experimental image restoration techniques with real time implementation of object representation in the motive of commercial household product Domains such as an Online marketing required for an open data analysis system. We will also perform algorithmic procedural strategies for the successful implementation of our proposed research technique in several sampling domains with a maximum level of improvements. In near future we will implement the Optimal multiple product poly comparison techniques for the pricing structure of online sales domain.

Keywords: Web Mining, parse, Web Content Mining, Hyperlink, online sales

I.INTRODUCTION

Web mining is the use of data mining techniques to automatically discover and extract information from Web documents and services [1].

There are three general classes of information that can be discovered by web mining:

- Web activity, from server logs and Web browser activity tracking.
- Web graph, from links between pages, people and other data.
- Web content, for the data found on Web pages and inside of documents [2].

At Scale Unlimited we focus on the last one – extracting value from web pages and other documents found on the web. There's no explicit reference to "search" in the above description. While search is the biggest web miner by far, and generates the most revenue [4], there are many other valuable end uses for web mining results. A partial list includes:

- Business intelligence
- Competitive intelligence
- Pricing analysis
- Events
- Product data
- Popularity
- Reputation

When extracting Web content information using web mining, there are four typical steps [3].

1. Collect – fetch the content from the Web
2. Parse – extract usable data from formatted data (HTML, PDF, etc)
3. Analyze – tokenize, rate, classify, cluster, filter, sort, etc.
4. Produce – turn the results of analysis into something useful (report, search index, etc)

The term **Web Data Mining** is a technique used to crawl through various web resources to collect required information, which enables an individual or a company to promote business, understanding marketing dynamics, new promotions floating on the Internet, etc. There is a growing trend among companies, organizations and individuals alike to gather information through web data mining to utilize that information in their best interest [5].

Data Mining is done through various types of data mining software. These can be simple data mining software or highly specific for detailed and extensive tasks that will be sifting through more information to pick out finer bits of information. For example, if a company is looking for information on doctors including their emails, fax, telephone, location, etc., this information can be mined through one of these data mining software programs [6]. This information collection through data mining has allowed companies to make thousands and thousands of dollars in revenues by being able to better use the internet to gain business intelligence that helps companies make vital business decisions [7].

Before this data mining software came into being, different businesses used to collect information from recorded data sources. But the bulk of this information is too much too daunting and time consuming to gather by going through all the records, therefore the approach of computer based data mining came into being and has gained huge popularity to now become a necessity for the survival of most businesses [8].

This collected information is used to gain more knowledge and based on the findings and analysis of the information make predictions as to what would be the best choice and the right approach to move toward on a particular issue [9]. Web data mining is not only focused to gain business information but is also used by various organizational departments to make the right predictions and decisions for things like business development, work flow, production processes and more by going through the business models derived from the data mining[10].

II.PROPOSED METHODOLOGY

The proposed methodology describes the input to our research model emphasize on the web page resources. Which will be then parsed into several components using the coded programmed and each resultant component is stored in the data mining structure. The text, images and document components are then retrieved for the specified necessity in order to obtain the optimal pricing strategy in online marketing domain.

DOI: 10.18535/ijecs/v5i1.26

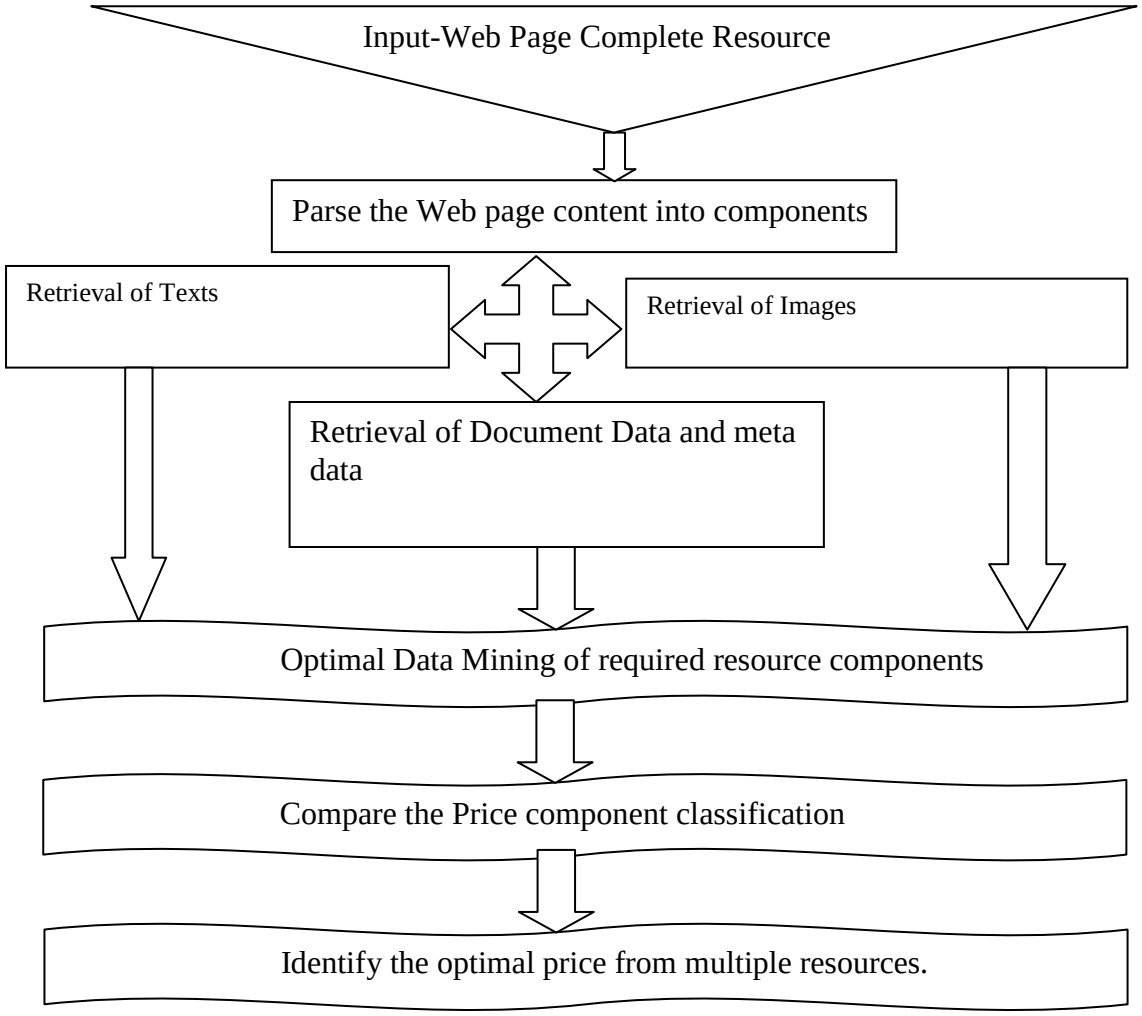

Figure-1: Proposed methodology for web data mining in online sales domain

III.IMPLEMENTATION

The implementation of the web page parsing is done in the basically procedure as follows,
1. Extracting the textual content.
2. Extracting the images in a web page.
3. Extracting the document data and Meta data from the corresponding product pages.

The actual implementation of web content extraction can be utilized by using the following java programming codes.

1. Extracting text content

```
public static String get(String url) throws Exception {
  StringBuilder sb = new StringBuilder();
  for(Scanner sc = new Scanner(new URL(url).openStream()); sc.hasNext(); )
    sb.append(sc.nextLine()).append('\n');
  return sb.toString();
}
public static void main(String[] args) throws Exception {
  System.out.println(get("http://www.yahoo.com"));
}
```

2. Extracting Images

```
public class ExtractAllImages {
   public static void main(String args[]) throws Exception {

      String webUrl = "http://www.hdwallpapers.in/";
      URL url = new URL(webUrl);
      URLConnection connection = url.openConnection();
      InputStream is = connection.getInputStream();
      InputStreamReader isr = new InputStreamReader(is);
      BufferedReader br = new BufferedReader(isr);

      HTMLEditorKit htmlKit = new HTMLEditorKit();
      HTMLDocument htmlDoc = (HTMLDocument) htmlKit.createDefaultDocument();
      HTMLEditorKit.Parser parser = new ParserDelegator();
      HTMLEditorKit.ParserCallback callback = htmlDoc.getReader(0);
      parser.parse(br, callback, true);

      for (HTMLDocument.Iterator iterator = htmlDoc.getIterator(HTML.Tag.IMG); iterator.isValid(); iterator.next()) {
         AttributeSet attributes = iterator.getAttributes();
         String imgSrc = (String) attributes.getAttribute(HTML.Attribute.SRC);

         if (imgSrc != null && (imgSrc.endsWith(".jpg") || (imgSrc.endsWith(".png")) || (imgSrc.endsWith(".jpeg")) ||
(imgSrc.endsWith(".bmp")) || (imgSrc.endsWith(".ico")))) {
            try {
               downloadImage(webUrl, imgSrc);
            } catch (IOException ex) {
               System.out.println(ex.getMessage());
            }
         }
      }
   }
   private static void downloadImage(String url, String imgSrc) throws IOException {
      BufferedImage image = null;
      try {
         if (!(imgSrc.startsWith("http"))) {
            url = url + imgSrc;
         } else {
            url = imgSrc;
         }
         imgSrc = imgSrc.substring(imgSrc.lastIndexOf("/") + 1);
         String imageFormat = null;
         imageFormat = imgSrc.substring(imgSrc.lastIndexOf(".") + 1);
         String imgPath = null;
         imgPath = "C:/Users/Machine2/Desktop/CTE/Java-WebsiteRead/" + imgSrc + "";
         URL imageUrl = new URL(url);
         image = ImageIO.read(imageUrl);
         if (image != null) {
            File file = new File(imgPath);
            ImageIO.write(image, imageFormat, file);
         }
      } catch (Exception ex) {
         ex.printStackTrace();
      }

   }
}
```

3.Extracting Content and Meta data from a document:

```
import java.io.File;
import java.io.FileInputStream;
```

```
import java.io.IOException;

import org.apache.tika.exception.TikaException;
import org.apache.tika.metadata.Metadata;
import org.apache.tika.parser.ParseContext;
import org.apache.tika.parser.html.HtmlParser;
import org.apache.tika.sax.BodyContentHandler;

import org.xml.sax.SAXException;

public class HtmlParse {

   public static void main(final String[] args) throws IOException,SAXException, TikaException {

      //detecting the file type
      BodyContentHandler handler = new BodyContentHandler();
      Metadata metadata = new Metadata();
      FileInputStream inputstream = new FileInputStream(new File("example.html"));
      ParseContext pcontext = new ParseContext();

      //Html parser
      HtmlParser htmlparser = new HtmlParser();
      htmlparser.parse(inputstream, handler, metadata,pcontext);
      System.out.println("Contents of the document:" + handler.toString());
      System.out.println("Metadata of the document:");
      String[] metadataNames = metadata.names();

      for(String name : metadataNames) {
        System.out.println(name + ":   " + metadata.get(name));
      }
   }
}
```

After extracting the codes and document data from the WebPages are stored in the data warehouse and the retrieval of the content using data mining analysis of sorted data approach can be utilized for the future reference.

IV.RESULTS AND DISCUSSION

The actual stored content from a data warehouse after the retrieval of web content from the corresponding trusted web resource is as follows for the product " **Philips HD4938 Induction Cook Top " in Table-1.**

Table-1:Price data from online resources

Philips HD4938 Induction Cook Top	
E-Seller Site	**Price**
Pay Tm	**Rs.4455**
Infibeam	**Rs.3837**
Amazon. in	**Rs.3039**
India Times	**Rs.4211**
Shop clues	**Rs.3325**
Flipkart	**Rs.3795**

Now performing the sorted table data with respect to the price and store it in the warehouse and the final sorted data in table-2 as follows,

Table-2:Sorted Price data from online resources

Philips HD4938 Induction Cook Top	
E-Seller Site	**Price**
Amazon. in	**Rs.3039**
Shop clues	**Rs.3325**
Flipkart	**Rs.3795**

DOI: 10.18535/ijecs/v5i1.26

Infibeam	**Rs.3837**
India Times	**Rs.4211**
Pay Tm	**Rs.4455**

Select the first minimal row data extraction for the optimized price for the required product "**Philips HD4938 Induction Cook Top** " with the price of Rs.3039 from Amazon.in.The comparison price chart is as follows,

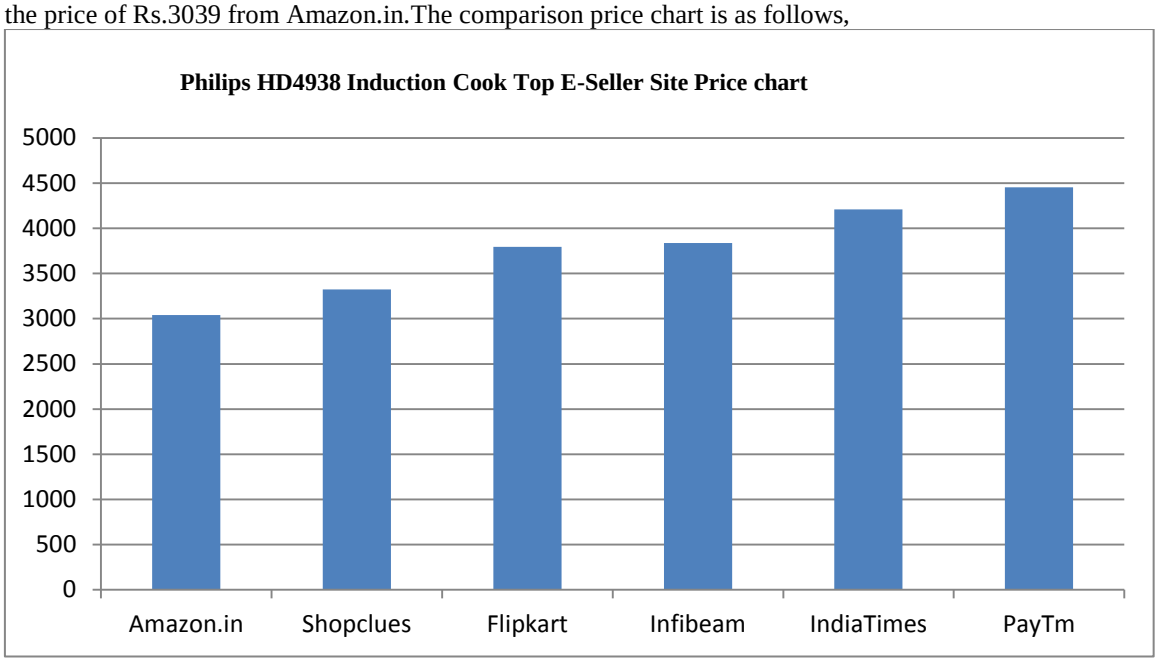

Figure-1:Price comparison analysis over web data mining in online sales domain

V.CONCLUSION

Web data mining is a combination of web mining with the data mining techniques of data filters. Restoring web content is an optimal or an expected form which is a highly technical process to implement in an efficient way. The selection of text, images, document and metadata with the proper utilization of web processing tool is a scientific methodology to implement.

Our proposed methodology make it as an easy process by the novel view of periodic web data level storage and retrieval combinations, further focusing of their mutual proportion along with variational effects we achieved an data analysis process with 99 % efficiency.. For a given product data, we are limited in the periodic price variations. For multiple product comparisons, we may be limited by how many combinations we can obtain. So we are limited in both cases by how many combinations we can average over, and this profoundly affects our estimations.

This is one of the main drawbacks that we found in the combination sector of periodic multiple product price range analysis techniques. In near future this research will focus on an optimal algorithmic identification of Universal Combination of multiple product data analysis process.

References

DOI: 10.18535/ijecs/v5i1.26

1. **Baraglia, R. Silvestri, F. (2007) "Dynamic personalization of web sites without user intervention", In Communication of the ACM 50(2): 63-67**
2. **Cooley, R. Mobasher, B. and Srivastave, J. (1997) "Web Mining: Information and Pattern Discovery on the World Wide Web" In Proceedings of the 9th IEEE International Conference on Tool with Artificial Intelligence**
3. **Cooley, R., Mobasher, B. and Srivastava, J. "Data Preparation for Mining World Wide Web Browsing Patterns", Journal of Knowledge and Information System, Vol.1, Issue. 1, pp. 5–32, 1999**
4. **Costa, RP and Seco, N. "Hyponymy Extraction and Web Search Behavior Analysis Based On Query Reformulation", 11th Ibero-American Conference on Artificial Intelligence, 2008 October.**
5. **Kohavi, R., Mason, L. and Zheng, Z. (2004) "Lessons and Challenges from Mining Retail E-commerce Data" Machine Learning, Vol 57, pp. 83–113**
6. **Lillian Clark, I-Hsien Ting, Chris Kimble, Peter Wright, Daniel Kudenko (2006)"Combining ethnographic and clickstream data to identify user Web browsing strategies" Journal of Information Research, Vol. 11 No. 2, January 2006**
7. **Eirinaki, M., Vazirgiannis, M. (2003) "Web Mining for Web Personalization", ACM Transactions on Internet Technology, Vol.3, No.1, February 2003**
8. **Mobasher, B., Cooley, R. and Srivastava, J. (2000) "Automatic Personalization based on web usage Mining" Communications of the ACM, Vol. 43, No.8, pp. 142–151**
9. **Mobasher, B., Dai, H., Luo, T. and Nakagawa, M. (2001) "Effective Personalization Based on Association Rule Discover from Web Usage Data" In Proceedings of WIDM 2001, Atlanta, GA, USA, pp. 9–15**
10. **Nasraoui O., Petenes C., "Combining Web Usage Mining and Fuzzy Inference for Website Personalization", in Proc. of WebKDD 2003 – KDD Workshop on Web mining as a Premise to Effective and Intelligent Web Applications, Washington DC, August 2003, p. 37**