RESEARCH ON WEB DATA USAGE AND EXTRACTION ASPECTS

Dr. S. Hemalatha[1]

ABSTRACT

As tremendous amount of data are involved in Internet and it is found to be tedious to discover data, ample data mining or machine learning techniques are indulged. All these kinds of techniques are based on intelligent computing approaches, which are based on database, data mining, machine learning, information retrieval, etc,. The growth of web-based data management focuses on the developments of Web applications as services, search engines, as end users, are still facing the problem of overwhelming information and leads to user support. The web users have to deal with the difficulties of finding the relevant information in a maze of data on the web. For example, if a user wants to search the desired information by utilizing a search engine such as Google Chrome, the search engine shows plenty of information both relevant and irrelevant. So, the user finds it tedious to come to a conclusion. Thus, the emerging trends of Web put an end to these kinds of issues by utilizing different types of extraction options found in the web.

Keywords: Web mining, Information Retrieval, Types of Data Extraction and Web Analytics Process

ABSTRACT

According to CRM (Customer Relationship Management) technique, web mining is the process of information obtained via traditional Data mining techniques. Web mining is obtained from data miming which directly retrieves the data from the web in the form of documents, services, content, hyperlinks and server logs. It creates patterns using web data as following the trends that exist in nature. It also helps to know the mentality of the customers by evaluating the web sites to attain the success in marketing. Web mining is divided into content mining, structure and usage mining to form a pattern. Information is evaluated using data mining primitives like clustering classification, association and patterns in sequence as shown in figure 1.

Web data mining techniques are used to discover the data found online and then extract the relevant information from the net. Searching on the web is a process that requires different algorithms [3]. Applying a search query, the relevant pages are searched using the available data, content and hyperlinks. Generally, any type of search engine will support multiple components like,

- A web crawler or spider for collecting web pages.
- A parser that extracts content and preprocesses web pages frequently.
- Indexer places the web pages in data structures.
- Retrieval information system has to be related to a query.

[1]Asst. Professor, Dept of CS, CA & IT, Karpagam Academy of Higher Education, Coimbatore,
Email: drhemashanmugam@gmail.com

- Ranking algorithm ranks the web pages in a sorting manner.

These parts can be divided into web structure mining techniques and web content mining techniques. The web crawler, indexer, and ranking procedures refer to the web structure as hyperlinks. The other parts like parser and retrieval of a search engine are grouped under web content analysis methods.

Figure 1 Structure of Web Mining

2. Categories of Web mining

- Web content mining - It is a kind of process which efficiently mines the needed information from the web page contents and documents, mainly in the form of text, images and audio or video files. Techniques used in these aspects are derived from Natural Language Processing and information retrieval.

  Web content mining is differentiated from distinctive points of view such [6] as information Retrieval and Database views. Based on the studies, tasks are carried out for unstructured information and semi-structured facts from data retrieval view. It clearly reveals that most of the researches use key words, for promoting statistics to determine single words in isolation: for unstructured text, only single word is found in the training mass as features. In case of semi-structured data, HTML structures are being utilized inside the files, and some applied as links between the files for representing it. The database view exposes better record control and queries at the web. However, mining constantly tries to deduce the structure of the website to convert an internet website online to a database.

- Web Structure Mining - Web Structure Mining examines the nodes and connection structures of a website with the help of graph theory. Benefits that are obtained by this are, connection of specific websites with other sites and its document related events to understand the connection between pages of the websites.

Extracting patterns from hyperlinks inside the internet is finished by way of hyperlink that forms a structural aspect connecting the net web page to an exclusive place. Mining of record structures involves the analysis of a tree-like shape of web page structures so that HTML or XML tag usage can be framed.

It highlights the terminology such as,

- Web graph as directed graph thereby representing web.
- Node generates the web page in graph.
- Edges are responsible for hyperlinks of pages.
- In degree forms total number of links pointing to specific node.
- Out degree indicates number of links generated from specific node.

One of the most significant strategies discovered in web structure mining is web page Rank. This frames a set of rules in flip and utilized by Google to rank search outcomes. This name is recommended with the aid of Google-founder Larry web page. The rank of a page decides the quantity of links pointing to the precise node.

- Web Usage Mining - Usage Mining is popular in case of extracting patterns and information from server logs to facilitate the user activity according to features like where they are from, total number of clicks over particular product on the site and all the other activities being done on the site to know the market strategy.

Web usage mining can be further classified based on its kind of data usage like,

- Web Server Data - The user logs are collected through Web server such as IP address, page reference and access time.
- Application Server Data- Commercial application servers have prominent features to enable e-commerce applications to be built on top of them with little effort. Its key benefit is the ability to track different types of business strategy and log them in application server logs.
- Application Level Data- New kinds of events can be defined in an application, and logging can be turned on for users to generate histories of specially enabled events. It will be analyzed to meet the requirements of users by combining one or more techniques applied from the categories found above [1].

3. Web Analytics Process

Many basic web analytics processes can be grouped under four intrinsic steps, [5] which are as follows,

- Data Collection is commonly done in starting stage to know the elementary data needed as a whole.
- Data processing takes place automatically after confirming it as useful knowledge.
- Developing key performance indicator factors is based on the number of counts combining with business strategies. Usually KPI is responsible for the conversion aspects depending upon the organization involved.
- Formulating online strategy concerns with the goals made objectives and perspective standards of the organization/ business. This concentrates on creating money in the form of saving, making towards market strategy. Another foremost quality that must be seen in an analysis websites optimization as shown in figure 2.
- Experiments and testing are controlled by two variants, for online features, such as web enlargement [2].

Testing helps in identifying changes on web pages that have been increased or maximized statistically from the tested result. An impact of each stage performs the stage preceding or following it, whereas data that enrich collection crash over online strategy. Otherwise the online strategy deployment influences the data that are being collected.

4. Web analytics technologies

Web analytics technologies applied for both off-site and on-site.

- Off-site web analytics deals with web extension and analysis apart from its ownership or maintenance. It measures the potential website and shares visibility and comments that are carried out in Internet today.
- On-site web analytics measures the behavior of any users who are visiting the website. This includes both drivers and conversions, as for instance, all the pages they visit in the form of links before the purchase of any item online in marketable context. The data are compared with key performance indicators' KPI factor for performance analysis, so as to improve the marketing campaign and user's response. Users mainly concentrate on Google Analytics, and Adobe Analytics are the most widely used in on-site category, since many emerging trends are there to accomplish better information like heat maps and session replay and so on.

Conventionally, web analytics considers the measurement of on-site visitor in general. But it does not seem to be not true, because dealers are incorporating tools that withstand both on-site and off-site characteristics. Different dealers enclose on-site web analytics software along with services. There are two kinds of methodology implemented to collect data namely, server log files analysis and page tagging. Both the methods produce web traffic reports during data collection and processing.

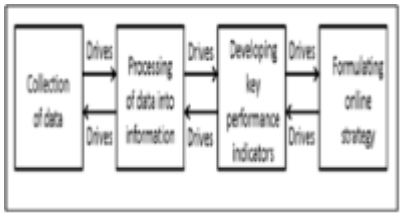

Figure 2 Steps to implement Web Analytical Process

5. Types of Data Extraction

5.1 Web scraping and web harvesting

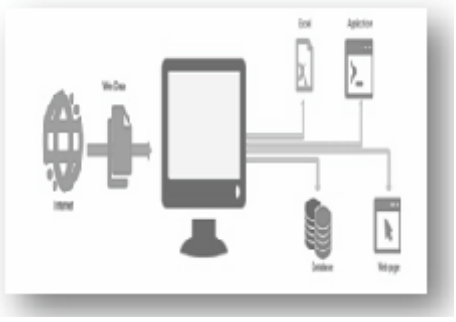

Figure 3 Pictorial Representation of Web Scraping

Web scraping is the technique of automating manual copy and paste action. In general, harvesting is the superset of all data collection methods, but web scraping is a specific method of harvesting. Web scraping, web harvesting or web data extraction depends upon data scraping to enable extraction of data on websites [7]. Web scraping software even access the World Wide Web directly using Hypertext Transfer Protocol (HTTP) or across a web browser. If web scraping is done manually by a user, then automated processes can be implemented using a bot or web crawler in the form of copying, where data are gathered and copied from the web, over central local database or spreadsheet to promote retrieval or analysis when it is needed.

Web scraping of a web page performs fetching or extracting from it [4]. Normally when a user views any page, fetching or downloading is done automatically. Meantime, web crawling is a main component within web scraping, to fetch pages for later processing. Once the pages are fetched, then extraction takes place automatically. This content of pages will be parsed, searched, reformatted, and the data copied into a spreadsheet, and so on. Web scrapers may gain some knowledge over these processes. In case of contact

scraping any person's name would be found and copied along with his number and company with the respective URL.

Web scraping in turn supports contact scraping, as a component of applications used for web indexing, web mining and data mining, online price fluctuation monitoring and comparison, product review scraping, real estate status, weather monitoring, website change, research, tracking online presence and reputation, mash-up and data integration over web.

Thus, web pages are built using text-based mark-up languages like HTML and XHTML, but frequently store useful data in the form of text. Mainly, web pages are designed to facilitate human end-users and not for ease of automated use resulting in the creation of tools to scrape the web content. A web scraper is a form of Application Programming Interface (API) to retrieve data from a web site. Sites like Amazon, AWS and Google enhance the web scraping tools, services and public data.

6. Conclusion

Intellectual incorporation and correlation of voluminous information from sources like logs and file index logs can tap information which cannot be known to others. Though current data mining techniques helps to give useful information from the repository, it needs many useful tools to support the aspects by including statistical criteria, visual effects and human support to accomplish the above said features in future.

References

1. Available at: https://www.techopedia.com/definition/15634/web-mining
2. Jansen, B. J. Understanding user-web interactions via web analytics, Synthesis Lectures on Information Concepts, Retrieval, and Services, 1(1), 1-102, 2009.
3. Kosala, R. and H. Blockeel, Web Mining Research: A Survey. SIGKDD Explorations, 2(1): p. 1-15, 2000.
4. Vargiu & Urru, Exploiting web scraping in a collaborative filtering- based approach to web advertising, Artificial Intelligence Research, 2 (1). Doi:10.5430/air.v2n1p44, 2013.
5. WAA Standards Committee, Web analytics definitions. Washington DC: Web Analytics Association 2008.
6. Wang, Yan, Web Mining and Knowledge Discovery of Usage Patterns, 2000.
7. Web scraping Boeing, G.; Waddell, P, New Insights into Rental Housing Markets across the United States: Web Scraping and Analyzing Craigslist Rental Listings, Journal of Planning Education and research, 2016.