IMPACT: International Journal of Research in Engineering & Technology (IMPACT: IJRET)
ISSN(E): 2321-8843; ISSN(P): 2347-4599
Vol. 2, Issue 12, Dec 2014, 9-12

"A REVIEW: WEB DATA SCRAPPING"

SWATI SANDIP PATEL

MCA, LDRP-ITR, Gandhinagar, Gujarat, India

ABSTRACT

Information drives today's businesses and the Internet is a powerhouse of information. Most businesses rely on the web to gather data that is crucial to their decision making processes. Companies regularly assimilate and analyze product specifications, pricing information, market trends and regulatory information from various websites and when performed manually, this is often a time consuming, error-prone process. So it is very important to create a simplified algorithm/tool that can easily extract data from web page and publish the extracted data in desired manner.

KEYWORDS: Association, Clustering, Crawling, IP Blackage, IP Rotation, Page Rank, Scrapping, Web Mining

INTRODUCTION

- Web data mining is a kind of data mining. Basically it's 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.
- 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 and has become a necessity for the survival of most businesses

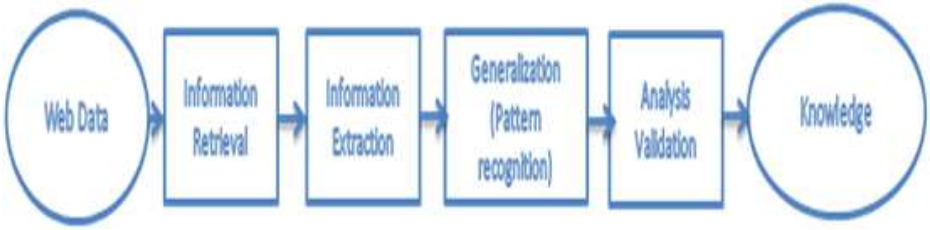

Figure 1

TYPES OF WEB DATA MINING

Web Data is Classified as

- **Web Content** – text, image, records, etc. e.g. extracting business information from web site.
- **Web Structure** – hyperlinks, tags, etc. for e.g. finding out number of links of particular website.
- **Web Usage** – http logs, app server logs, etc. for e.g. finding out number of request per day.

- **Web Content**

Web Content Mining is the process of extracting useful information from the contents of Web. It may consist of text, images, audio, video, or structured records such as lists and tables. Research activities in this field also involve using techniques from other disciplines such as Information Retrieval (IR) and natural language processing (NLP).

- **Web Structure**

Identifying interesting graph patterns or preprocessing the whole web graph to come up with metrics such as Page Rank. Web Structure Mining can be is the process of discovering structure information from the Web. This type of mining can be performed either at the (intra-page) document level or at the (inter-page) hyperlink level. It's a useful source for extracting information such as quality of web page, interesting web structure etc.

- **Web Usage**

User identification, session creation, robot detection and filtering, and extracting usage path patterns

CURRENT TRENDS AND LIMITATIONS

- Currently the mining tools available in the market include Kapow, Automation Anywhere etc., which facilitate the web mining by creating crawler/robot/spider/bots etc. for different websites
- But all tools are quite expensive and have limited support for complex html structure.
- The tools available these days are not so user friendly so it is very difficult for business/novice user to create bots/spider on fly.
- The tools available this days requires programming knowledge which will not be meaningful for the general user
- The tools are lacking IP rotation facility that is very important especially when you are dealing with famous web sites like Google, Facebook etc.
- The tools are not designed to support large data scrapping so scalability is also one of the concerns when scrapping huge data over the web.
- The tools are limited to avoid duplicate scrapping and usually get stuck up the iterative job

PROBLEM FORMULATION AND CHALLENGES

- Frequent Changes in Web site structure.
- Java script based loading of page content. (e.g. Ajax requests)
- Collaborative Crawling
  - Avoid lots of redundant crawling. (e. g recursive references)
  - Avoid IP blockage from third party site (possible solution. need IP rotation program).
  - Efficient fetching (hundreds of pages per second. Possible solution: implementing tread or task based architecture)

- Deep Web Crawling (Deeply structured websites)
- Crawling multimedia
  - Bigger load on web sites since files are bigger.
  - More apparent copy right issues.
  - More resources (band width, storage place etc.) required
  - More complications may arrive to resolve duplicates (possible solution, assign unique #hashid to record inserted)

PROPOSED WORK

Objectives

- The overall objective of this research is to create a optimized algorithm/tool that can be used to extract desired information of the web page without any programming knowledge. Going beyond simple screen scraping or cutting and pasting information from a website into an application or file
- Algorithm / tool should be smart enough to grab information from text, images or even site behind the login. Additionally, it should also support the scrapping of https (secured site) sites plus Ajax enabled sites too.
- Extraction of the data can be based on the repetitive data pattern, semantic search etc.
- Algorithm/ tool should be smart enough to intelligently transfer information from web to desired format (excel, pdf, database etc.)
- Tool should capable enough to publish extracted data via web service so it can be consumed on any platform.
- The more basic and popular data mining techniques include:
  - Classification
  - Clustering
  - Associations
- Proposed system is meant to provide scalability through Parallel Processing concepts
- A Scheduler can also be made to do interval based tasks automatically
- Media Content can also be published
- Possible implementation in real applications
  - Extract competitor's price list from web page regularly to stay ahead of competition
  - Extract people's data from web page and put it in a database.
  - Extract opening and closing price of stock from a web page.

- Extract Mutual Funds information from a website daily.
- Extract data from one online system and transfer it to another online system.
- Scrape unstructured data from the web and transfer it to Excel
- Regularly download updated web images of weather maps

REFERENCES

1. Extracting Content Structure for Web Pages based on Visual Representation; Deng Cai 1*, Shipeng Yu 2*, Ji-Rong Wen* and Wei-Ying Ma* * Microsoft Research Asia {jrwen, wyma}@microsoft.com 1 Tsinghua University, Beijing, P. R. China caideng00@mails.tsinghua.edu.cn 2 Peking University, Beijing, P. R. China ysp@is.pku.edu.cn
2. SOCIAL MEDIA PROFILING: A PANOPTICON OR OMNIOPTICON TOOL?; Lilian Mitrou; Miltiadis Kandias; Vasilis Stavrou; Dimitris Gritzalis
3. Focused crawling: a new approach to topic-specific Web resource discovery; Soumen Chakrabartia, Martin van den Bergb, Byron Domc a Computer Science and Engineering, a Indian Institute of Technology, Bombay, 400076, India b FX Palo Alto Laboratory,3400 Hill view Ave, Bldg4, Palo Alto, CA 94304, USA c IBM Almaden Research Center, 650 Harry Rd, SanJose, CA
4. Link Analysis in Web Information Retrieval; Monika Henzinger Google Incorporated Mountain View, California monika@google.com
5. Web Mining: Accomplishments & Future Directions; Jaideep Srivastava University of Minnesota USA srivasta@cs.umn.eduhttp://www.cs.umn.edu/faculty/srivasta.html
6. http://searchcrm.techtarget.com/definition/Web-mining
7. http://www.web-datamining.net/
8. http://www.academia.edu/484699/Web_Mining_Today_and_Tomorrow
9. http://www.slideshare.net/denshe/icwe13-tutorial-webcrawling
10. Bing Liu; Web Data Mining Exploring Hyperlinks, Contents, and Usage Data Web Data Mining Exploring Hyperlinks, Contents, and Usage Data.