A SURVEY ON INFORMATION RETRIEVAL FROM WEB USING WEB SCRAPING TECHNIQUE

Rushabh A. Patel, Mansi Patel
Computer Engineering Department
Silver Oak College of Engineering and Technology, Ahmedabad

Abstract- **In this survey paper, we describe the structural design of web mining and how his work and describe web mining in the third part of the extraction of Web content, Web usage and Web structure mining. Extraction using web content, we put all the data on the WWW and also we manage unstructured information using web content mining web scraping technique. We explain how to retrieve information from the Web and manage in structure, size using an algorithm tree change as the algorithm compares the distance and time of any complexity algorithm.**

Index terms – Web mining, Scraping, tree edit distance, Information Retrieval

I. INTRODUCTION

With the volatile growth of information sources available on the World Wide Web, it has become gradually necessary for users to apply automated tools in innovation the desired information resources, and to track and examine their usage patterns. Such kind of factors provide escalation to the inevitability of creating server-side and client-side intelligent systems which can excellently mine for knowledge. [1]
Web mining can be broadly defined as the discovery and analysis of useful information from the World Wide Web. **Web mining** is the application of data mining techniques to mine knowledge from the Web data.

Web mining is divided into three different parts

a) **Web content mining** (*Text, image, records etc.*)

The extraction of useful information from Web pages

b) **Web structure mining** (*hyperlinks, tags, etc.*)

The development of useful information from the links included in the Web documents.

c) **Web usage mining**

The extraction of useful information from clickstream analysis of Web server logs contain details of web page visits, transactions, etc.

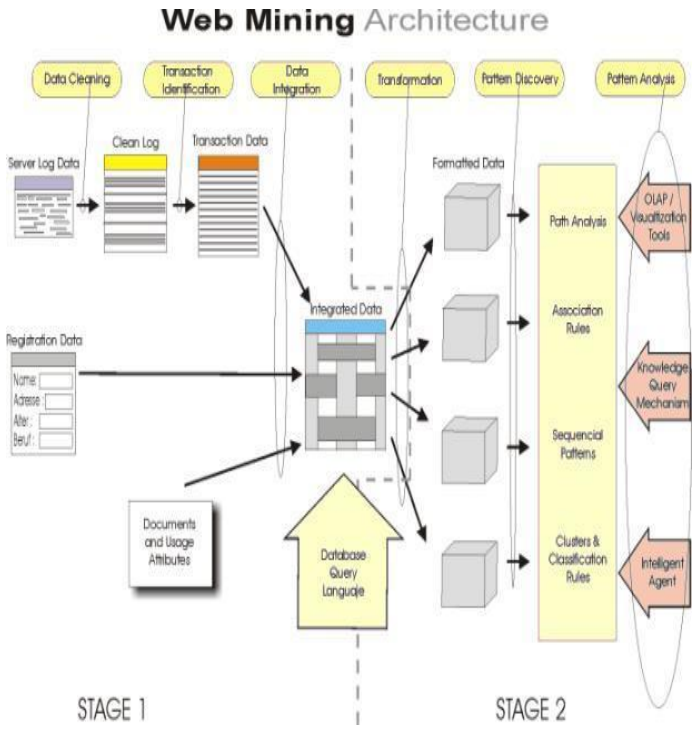

Fig.1 Architecture of Web mining [6]

II. TAXONOMY OF WEB MINING

There are the basic three main categories of Web mining, which is classified in the below section

Web Usage Mining:

Web usage mining is the procedure of concentrating helpful data from server logs e.g. utilization Web use mining is the methodology of figuring out what clients are searching for on the Web. A few clients may be taking a textual at just literary information, though a few others may be intrigued by mixed media information. Web Use Mining is the application of information mining systems to find intriguing use designs from Web information keeping in mind the end goal to comprehend and better serve the needs of Online applications. Use information catches the identity or origin of Web clients alongside their skimming conduct at a Site. Web usage mining itself can be grouped further relying upon the sort of user information considered: [1] [2] [3]

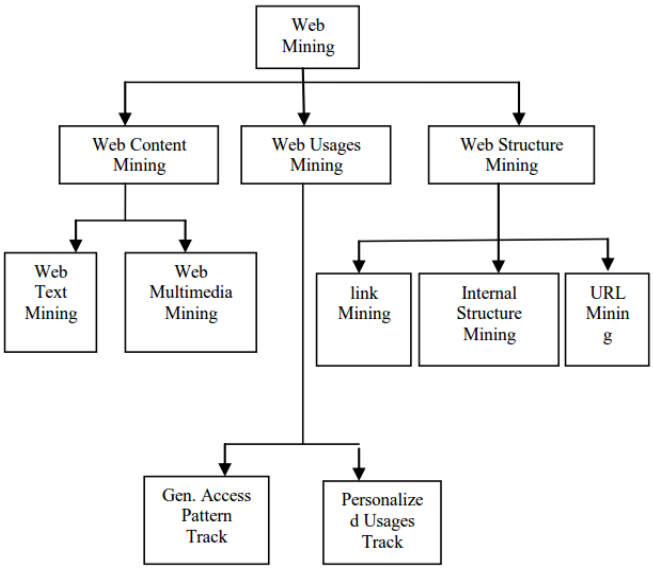

Fig.2 Taxonomy of Web Mining

Web Server Data: The client logs are collected by the Web server. Generally, facts and figures include Internet Protocol address, sheet citation and get access to them. [2]

Application Server Data: monetary proposal servers have significant characteristics to grant e-commerce submissions to be built on top of them with tiny effort. A key feature is the proficiency to pathway diverse kinds of enterprise events and logs them in application server logs. [1]

***Application Level Data*:** New sorts of events can be portrayed in an application, and logging can be wound on for them in this way creating histories of these particularly depicted events. It should be noted then again, that different end entrances require a union of one or a more noteworthy measure of the frameworks facilitated in the classes above [1] [2] [3]

Web Structure Mining:

Web structure mining is the procedure of utilizing graph theory, hypothesis to break down the node and the association structure of a web site. As per the sort of web structural information, web structure mining can be separated an into two sorts: [1]

1. ***Extracting examples from hyperlinks on the web***: a hyperlink is a structural part that unites the Web page to a different location.

2. ***Mining the record structure***: examination of the tree-like structure of page structures to portray HTML or XML tag use. [1]

Web Content Mining:

A. Web content mining is the mining, extraction and coordination of helpful information, data and learning from Page content.

B. The heterogeneity and the absence of structure that saturates a significant part of the constantly growing data sources on the Internet such as Lycos, Alta Vista, WebCrawler, ALIWEB , Meta Crawler, and others provide some relaxation to users, but they do not generally provide structural information nor classify, filter, or interpret documents.

C. In recent years, these causes have prompted researchers to develop more intelligent tools for information retrieval, such as intelligent web agents, as well as to extend database and data mining techniques to make accessible a higher level of association for semi-structured data available on the web.

D. The agent-based approach to web mining involves the development of refined AI systems that can act unconventionally or semi-conventionally on behalf of a particular user, to discover and organize web-based information.

E. Web content mining differentiate two different points of view: [1]

Information Retrieval View and Database View.
Summarising the research works prepared for two kind of data, those are unstructured data and semi-structured data from the information retrieval view.

III. WEB SCRAPING AND STRUCTURE

Web scraping is a process of collecting information from the web using computer software programmatically. For example, you wish to get all the links in the Google search result of a term. If you know the URL for the search result page, you could get the links programmatically using web scrapping technique.

Web scraping should be done only if you are authorized to the targeted data source. Data scraping on a site without their (targeted data source) permission is illegal. The aim of the web scraping is to reduce the time and money spent on manual scraping.

Terms used in web scrap

1. *Scrap/Targeted Data Source*
   This could be a web page, web service, etc.
2. *Data Extractor*

This component extracts the required data from the response retrieved from the scrap data source

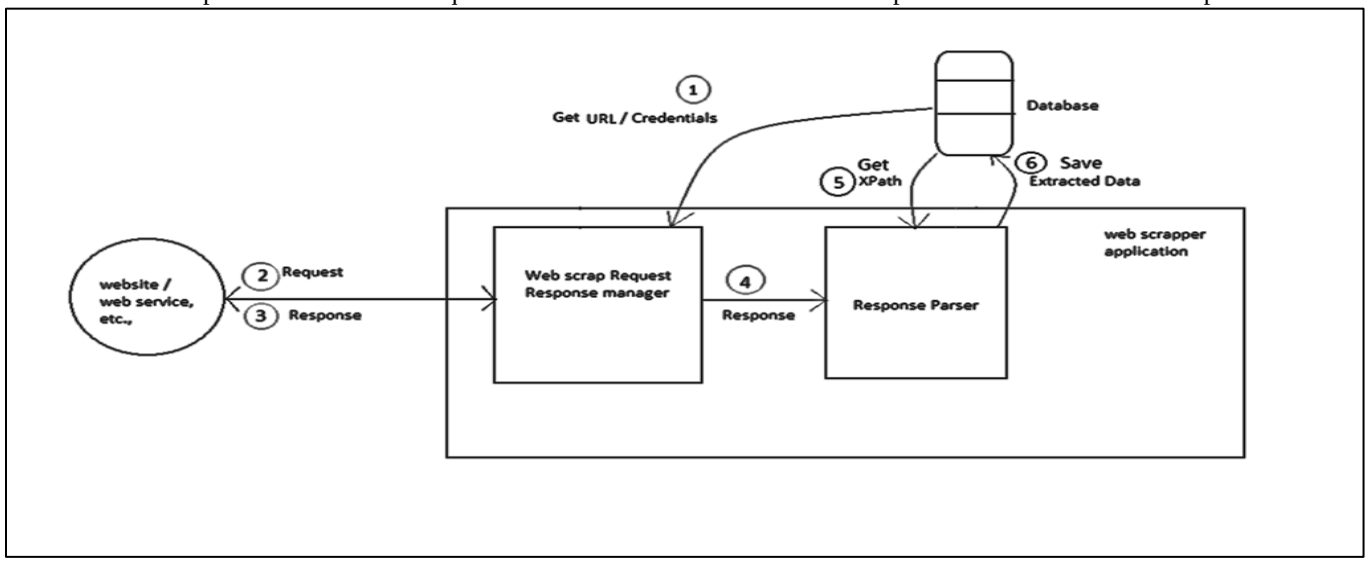

Fig.3 Architecture of Web Scraping [7]

3. Response

The data fetched from the scrap source could be an HTML, XML, simple text that depends on the targeted data source. For e.g. If you are going to scrap a web page you could get the HTML as your response.

4. Extracted Data

This could just be a text, link, any crucial information, etc.

A simple web scrap application that scraps a web page, parses/extracts the required data from the response and then saves the extracted data to a data storage.

This logically consists of two components

I. Web Request/Response Manager

1. Fetches the URL, credentials from the data storage (for e.g. From an XML file, database)
2. Makes the request to the targeted data source and if requires it authenticates your application with the targeted data source (using the privilege, authentication information like username and password, security key, etc.
3. Gets the response
4. Passes it to the Response Parser component

II. Response Parser

5. Gets the XPath that tells the data to be extracted in the response (assume the response is an XML /HTML).
6. Extracts the required data from the response located in the XPath and stores the data into a storage medium like database, flat file, XML, etc.,.

IV. WEB SCRAPING ALGORITHM AND ITS COMPARISON

(Tree Edit Distance algorithm)

The tree edit distance between ordered labelled trees is the **minimal-cost sequence of node edit operations** that transforms one tree into another. We consider following three edit operations on labelled ordered trees:

- **Delete** a node and connect its children to its parent keeping the order.
- **Insert** a node between an existing node and a subsequence of successive children of this node.
- **Change** the label of a node.

Zhang-Shasha [11]

Let's consider three kinds of operations. Changing node n *means* altering the label on n. Deleting a node *n* means constructing the children of *n* become the children of the parent of *n* and then removing *n.* Inserting is the accompaniment of delete. This means that inserting n *as* the child of *n'* will create *n* the parent of a following subsequence of the current children of *n'.* Figs. 1-3 illustrate these editing operations. [11]

A. Change. To make the change one node label to another.

B. Delete. to do the delete a node. (All children of the deleted node b become children of the parent a.*)*

C. Insert. to do the insert a node. (A consecutive sequence of siblings among the children of *a* become the children of *b.)*

Following [WF] and [T], we characterize an edit operation as a pair (*a, b) =; t=* (A, A),

sometimes transcribed as *a ~ b,* where *a* is either A or a label of a node in tree *T,* and *b* is either A or a label of a node in tree *T2•* We call *a ~ b* a change operation if *a =;t=* A and *b =;t=* A; a delete operation if *b* = A; and an insert operation if *a* = A. From the time when many nodes may have the same label as earlier, this notation is theoretically ambiguous. It could be made accurately by recognizing the nodes as well as their labels.

Let Us be a sequence S1, •••, •••, ••• , S_k *of* edit operations. An S-derivation from *A* to *B* is a sequence of trees *Ao, ... ,Ak* such that *A = Ao, B = Ak,* and A-I ~ A, via *s,* for *l~i~k.*

Let y be a cost function that assigns to each edit operation a ~ *b a* nonnegative real number y (a ~ *b).* This cost can be dissimilar for different nodes, so it can be used to provide greater weights to, for example, the higher nodes in a tree than to lower nodes.

We constrain y to be a distance metric. That is,

(I) *y(a~b)~O; y(a~a)=O*

(ii) *y(a~b)=y(b~a);and*

(iii) y(*a ~ c)* ~ y(*a ~ b)* + y(*b ~ c).*

We extend y to the sequence S by letting y(S) = t:IISI y(s;). Formally the distance between *T1* and *T2* is defined as follows:

8(*Tt, T2)* = min {y(S) I S is an edit operation sequence taking *TI* to *T2}.*

The definition of y makes 8 a distance metric also.

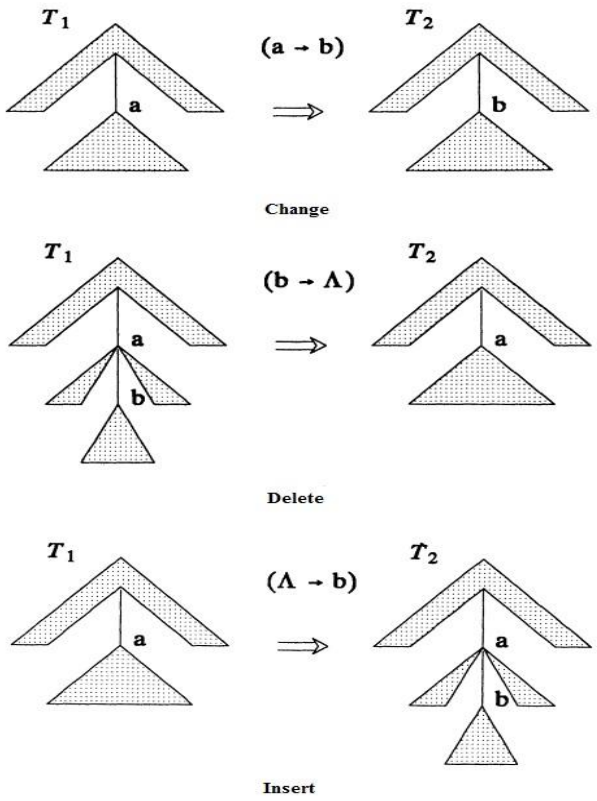

Fig.4 Tree Edit Operations[11]

Formulate a dynamic program and show how to compute a solution bottom-up. They reduce space requirements by identifying which sub problems that are encountered more than once and discard those that are not. The time complexity is reduced because the algorithm exploits that the solution to some sub problems is a byproduct of a solution to another.

It's time complexity is $O(|T_1||T_2|) \cdot$ Min (leaves (T_1), height $(T_1)) \cdot$ Min (leaves (T_2), height (T_2)) time (worst case $O(|T_1|^2|T_2|^2)$) and $O(|T_1||T_2|)$ space.

Philip Klein [15]

Based on the same formulation of a dynamic program as Zhang and Shasha's algorithm, the author proposes an algorithm that requires fewer sub problems to be computed in the worst case. In a top down implementation, the algorithm alternates between the formulation by Zhang and Shasha and its symmetric version based on the sizes of the trees in the sub forest of T_1.

It's time complexity of $O(|T_1|^2|T_2|. \log (|T_2|))$,While maintaining a $O(|T_1||T_2|)$ space bound.

Demaine[8]

Demaine et al., 2009- presents an algorithm that alternates between the two formulations of the dynamic program similarly to Klein's algorithm. However, the conditions for choosing one over the other, i.e. the recursion strategy, are more elaborate. The algorithm runs in $O(|T_1|^2|T_2|(1+\log T_2/T_1))$ time (worst case $O(|T_1|^2|T_2|)$) and $O(|T_1||T_2|)$ space.

K. C Tai[12]

K.C Tai's paper presents the algorithm to solve the tree edit distance problem. But it is complicated and impractical to implement.
Its time complexity is: $O(|T_1||T_2|.\ \text{height}\ (T_1)^2\ .\ \text{height}(T_2)^2)$ Which in the worst case is $O(|T_1|^3\ |T_2|^3)$.

RTED[10]

RTED works on $O(n^2)$ runtime and $O(n^3)$ space complexity.

V.REAL LIFE APPLICATION [16]

Some real-world examples of how our customers extract data:

Extract online pricing data, subtract one cent and add it to your online store.Update internal systems with the latest exchange rates and stock-market quotations.Gather leads from online business directories.Gather search engine rankings.Gather company information from many different directory websites.Monitor order status from ecommerce portals. See what orders you still need to fulfil, when they were ordered, and all applicable details. Gather bookings for any type of resort, or area. Gather price, quantity, item name, description, etc., from a supplier's website. Check competitor's shipping rates on major shopping sites. Monitor web-server availability and status. Perform keyword and PPC research. Extract product images and specification documents. Extract useful information from an encyclopedia and journal websites. Check the Meta information on the pages of a website (description, keywords, page title).

Comparison

No	Year	Author(s)/ Algorithm	Approach/ Overview of Work	Time Complexity
1	1989	Zhang-Shasha	Bottom-up Implementation	$O(\|T_1\|^2 \|T_2\|^2)$
2	1998	Philip Klien	Top-down Implementation	$O(\|T_1\|^2 \|T_2\|.\log(\|T_2\|))$
3	2009	Demaine	Top-Down Implementation	$O(\|T_1\|^2 \|T_2\|)$
4	1979	K. C. Tai	Complicated & Impractical to Implement	$O(\|T_1\|^3 \|T_2\|^3)$
5	2011	RTED	Upper bound	$O(n^2)$

VI. CONCLUSION AND FUTURE WORK

In this paper, we discuss that web mining, classification and their taxonomy and from this, we describe web content mining. In web content mining Information evocation technique like web scraping using their tree, edit distance algorithm we can retrieve data with structure format. We also explained different algorithm and their time complexity.

On Behalf of this, my future work I will check all algorithm complexity and their results after that I will select best from this and in existing algorithm, it applies to all the key routes. Which can be modified, mapping function can be applied on the required node of the tree for the specific operation with consideration of cost.

REFERENCE

[1] R. Cooley, B. Mobasher, and J. Srivastava, "Web Mining: Information and Pattern Discovery on the World Wide Web",IEEE 1997

[2]. Ashika gupta,Rakhi Arora, Ranjana Sikhavar,Neha Saxena, "Web Usage Mining Using Improved Frequent Pattern Tree Algorithms",IEEE 2014

[3].Brijendra Singh,Hemant Kumar Singh," WEB DATA MINING RESEARCH: A SURVEY",IEEE2010

[4].Ajith Abraham,"Miner: A Web Usage Mining Framework Using Hierarchical Intelligent Systems"

[5]. Dongkwon Joo and Songchun Moon," Scalable Web Mining Architecture for Backward Induction in Data Warehouse Environment", IEEE 2001

[6] Architecture of web Mining
http://meilingbieportfolio.wordpress.com/

[7]Architecture of web Mining
http://dotnet4features.wordpress.com/2011/04/30/web-scraping-%E2%80%93-a-developer%E2%80%99s-perspective/

[8] E.D. Demaine, S. Mozes, B. Rossman and O. Weimann," An Optimal Decomposition Algorithm for Tree Edit Distance", ACM Trans, 2009.

[9] Philip Bille,"A Survey on Tree Edit Distance and Related Problems",IEEE 2001

[10] Mateusz Pawlik," RTED: A Robust Algorithm for the Tree Edit Distance",arxiv 2011

[11] Kai Zhong Zhang And Dennis Shasha," Simple Fast Algorithms For Editing Distance Between The Trees And Related Problems"

[12]Kuo - Chung Tai,"The Tree-To-Tree Correction Problem".ACM 1979

[13] Tree edit distance
http://www.inf.unibz.it/dis/projects/tree-edit-distance/tree-edit-distance.php

[14] Shihyen Chen · Kaizhong Zhang," An improved algorithm for tree edit distance with applications for RNA secondary structure comparison",Springer 2012
[15] Philip N. Klein," Computing the Edit-Distance Between Unrooted Ordered Trees",Springer 1998

[16]Real World Application of Web Scraping
http://imacros.net/overview/data-extraction