ISSN:0975-9646

Deep Web Mining and its Application in Business Intelligence

RAKESH KUMAR BALODA[1], PRAVEEN KANTH[2]

[1]*Research Scholar, BRCM CET Bahal, Haryana, India*

[2]*Assistant Professor, BRCM CET Bahal, Haryana, India*

***Abstract*— World Wide Web has a vast source of information related to both educational and business domains. Data on web consists of both Structured and Unstructured type. For knowledge discovery and representation this raw data needs to be fetched to local system (crawling), mined (data extraction), extracted records needs further processing (cleaning, transformation, normalization) before it can be analysed by a data analyst (or interpreted by any other software application). This process is referred as ETL (Extract, Transform and Load) in Data Warehousing terminology. Information Extracted from web can then be used to make competitive business decisions by an Organisation.**

In this review paper we focus on the use of web data mining in Business Intelligence operations.

***Keywords*— Information Retrieval (IR), Information Extraction (IE), Deep Web, Crawling, Extraction, Web data mining.**

I. INTRODUCTION

A. Information Retrieval and Information Extraction:

Information Retrieval (IR) recovers a subset of documents from a source (such as World Wide Web) that match an end-user's query, while Information Extraction (IE) recovers individual facts from those documents. The difference between IR and IE is one of granularity regarding information access. IR is document retrieval whereas IE is fact retrieval.

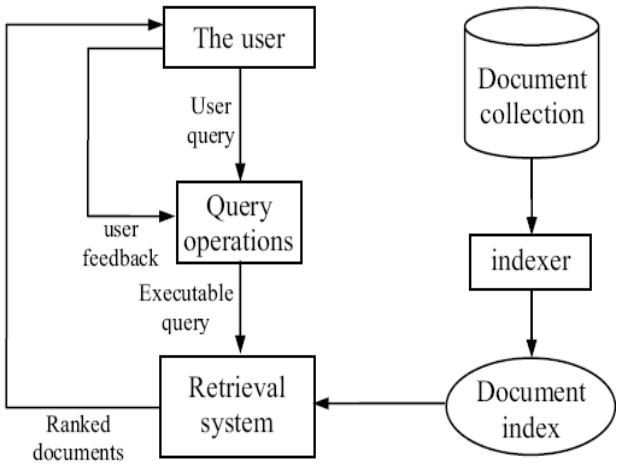

Fig 1. A general IR system architecture.

B. Deep Web(Hidden Web):

The deep web, invisible web, or hidden web is a part of the World Wide Web whose contents are not indexed by standard search engines for any reason. The deep web is opposite to the surface web.

Content types of Deep Web:

Methods which prevent web pages from being indexed by traditional search engines may be categorized as one or more of the following:

- *Contextual Web*: pages with content varying for different access contexts (e.g., ranges of client IP addresses or previous navigation sequence).
- *Dynamic content*: dynamic pages which are returned in response to a submitted query or accessed only through a form, especially if open-domain input elements (such as text fields) are used; such fields are hard to navigate without domain knowledge.
- *Limited access content*: sites that limit access to their pages in a technical way (e.g., using the Robots Exclusion Standard or CAPTCHAs, or no-store directive which prohibit search engines from browsing them and creating cached copies).
- *Non-HTML/text content*: textual content encoded in multimedia (image or video) files or specific file formats not handled by search engines.
- *Private Web*: sites that require registration and login (password-protected resources).
- *Scripted content*: pages that are only accessible through links produced by JavaScript as well as content dynamically downloaded from Web servers via Flash or Ajax solutions.
- *Software*: certain content is intentionally hidden from the regular Internet, accessible only with special software, such as Tor, BitComet I2P or other darknet software.
- *Unlinked content*: pages which are not linked to other web pages, which may prevent web crawling

programs from accessing the content. This content is referred to as pages without back- links.

- *Web archives*: Web archival services such as the Wayback Machine enable users to see archived versions of web pages across time, including websites which have become inaccessible, and are not indexed by search engines such as Google.

  The "Hidden Web" also refers to web pages that are dynamically generated from databases. Web technology has shifted away from putting content into static pages to towards placing information in relational databases.

  Based on a user query, content is pulled from databases and placed in a template ("on-the-fly") and delivered to the end-user. Conventional search engines cannot index the "hidden web."

C. Wrappers

Wrapper in data mining is a program that extracts content of a particular information source and translates it into a relational form.

Many web pages present *Structured Data* – such as telephone directories, product catalogue information etc. formatted for human browsing using HTML language. Structured data are typically descriptions of objects retrieved from underlying databases and displayed in Web pages following some fixed templates. Extracting such data allows one to provide value added services, e.g. comparative shopping, and meta-search. Software systems using such resources must translate HTML content into a relational form. Wrappers are commonly used as such translators.

Wrappers demonstrate that extensive linguistic knowledge is not necessary for successful Information Extraction (IE). Instead, shallow pattern-matching techniques (use of Regular Expressions) can be very effective. Information can be extracted from texts based on document formats rather than what the sentences "actually means." This type of IE analysis is ideally suited to the Web because online information is a combination of text and document structure. Almost all documents located on Web servers offer clues to their meaning in the form of textual formatting.

II. Web Content Mining

Web content mining extracts or mines useful information Or knowledge from Web page contents.

Basic steps of a Web content mining model can be described as follows:

A. Crawling:

A Web crawler is a program that automatically traverses the Web's hyperlink structure and downloads each linked page to a local storage. Crawling is often the first step of Web mining process.

There are two types of crawlers: universal crawlers and topic crawlers. A universal crawler downloads all pages irrespective of their contents, while a topic crawler downloads only pages of certain topics (relevant to user's search query).

B. Page Chunking :

Once the web pages (or documents of our interest) have been crawled to our local system (or LAN server hard disk) the next step is to logically divide the web pages (or data source) into separate blocks/sections of interest (based on information contained in those logical blocks), filter out any irrelevant information(like banner- adds, navigation menu , header-footer etc). The process may also refer to as Pre-processing of document before data extraction.

C. Data Extraction:

Data extraction is the act or process of retrieving data out of (usually unstructured or poorly structured) data sources for further data processing or data storage.

There are two main approaches to data extraction.

One is the supervised approach, which uses supervised learning to learn data extraction rules. The other is the unsupervised pattern discovery approach, which finds repeated patterns (hidden templates) in web pages for data extraction.

Regular Expressions (Regex) pattern matching can be effectively used for web content mining.

D. Data Cleaning, Transformation & Normalization:

Data cleansing, data cleaning or data scrubbing is the process of detecting and correcting (or removing) corrupt or inaccurate records from a record set, table, or database.

Once the raw data has been extracted from a unstructured source it needs to be cleansed and transformed, the various tasks of data post processing (after data extraction) may involve removal of unwanted html tags (like <b>, <i>, etc), removal of whitespaces, duplicate records in data set, text replacement(of special Unicode characters, currency symbols or html entities), conversion to uppercase (or lowercase) characters, data transformation to desired date/time formats (MM/DD/YYYY hh:mm:ss) according to the target database requirements.

Normalization is the process of organizing the columns (attributes) and tables (relations) of a relational database to minimize data redundancy.

E. Loading into Target System :

Normalized data is then uploaded to the target system, the target may be a client interface (client's portal), a data warehouse, a data analyst's database server or any other software system consuming that data as input for further processing.

III. Applications In Pricing Intelligence

Information thus gathered through the web content mining process can be applied to make competitive or pricing variance decisions by an organization as per directives of the top-level management (or data analysts).

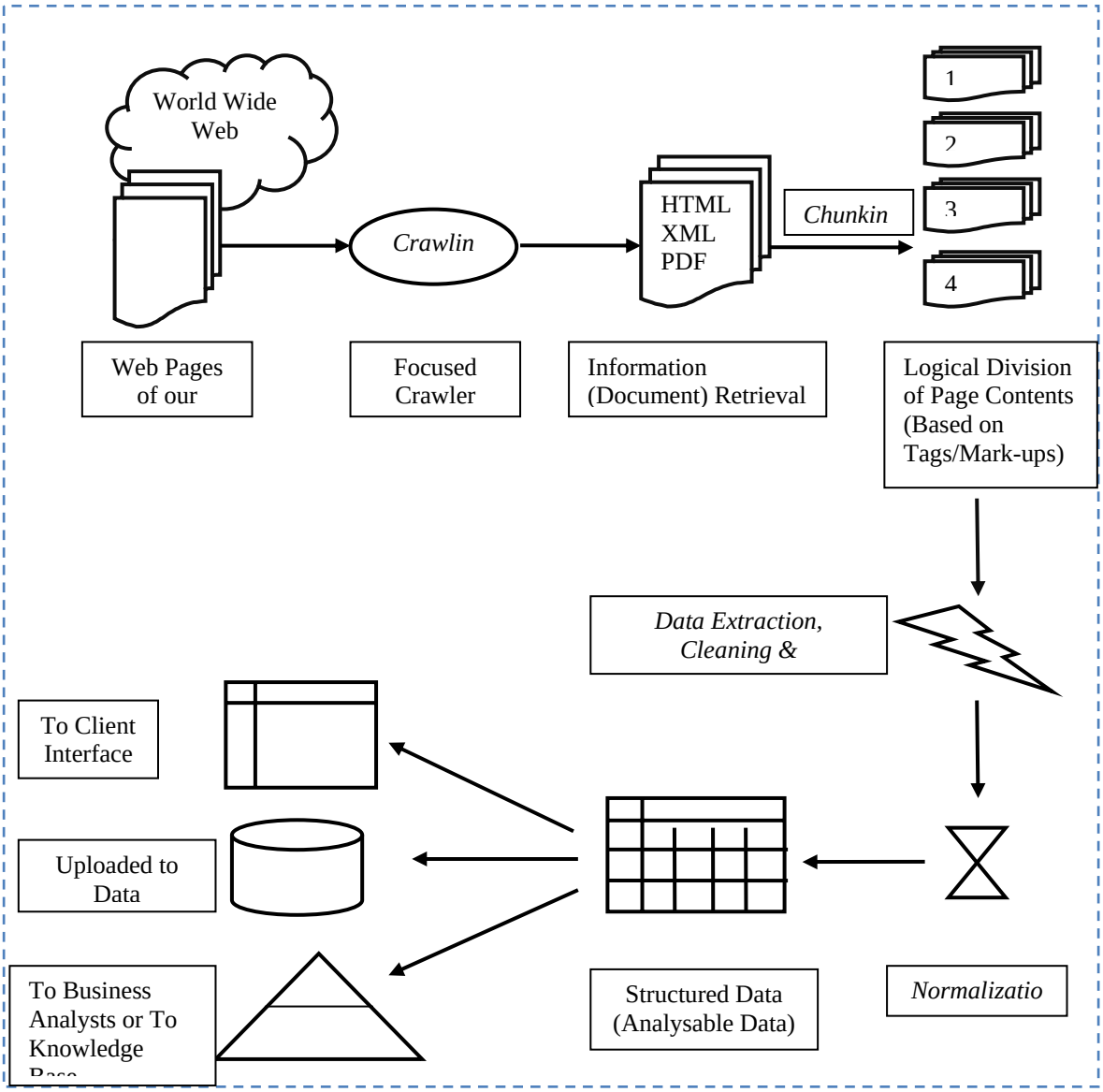

Fig 2. A general architecture for Web Content mining model

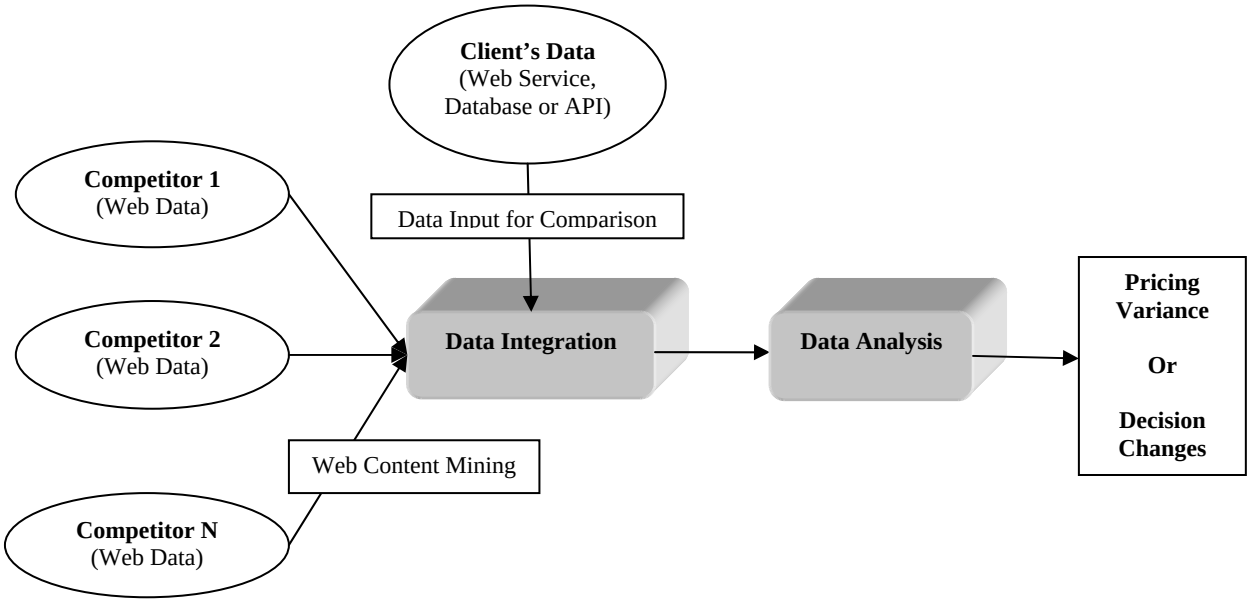

Fig 3. Pricing Variance Model

IV. CONCLUSIONS

Here in this review paper we have discussed the general web content mining model and its application in business pricing intelligence process.

Our future work involves an implementation of a general script based crawler and a Regex (regular expression) based data extraction engine that can implement the above discussed model.

REFERENCES

[1] Bing Liu, Web Data Mining – Exploring Hyperlinks, Contents and Usage Data, Springer 2007.

[2] Katherine C. Adams, The Web as Database: New Extraction Technologies and Content Management. ONLINE March 2001 – Information Today, Inc.

[3] Nicholas Kushmerick, Daniel S.Weld, Robert Doorenbos, "Wrapper Induction for Information Extraction", IJCAI-97.

[4] Wikipedia(www.wikipedia.com)
https://en.wikipedia.org/wiki/Web_crawler
https://en.wikipedia.org/wiki/Deep_web