NIET Journal of Engineering & Technology (NIETJET)
Volume 8, Issue Summer 2020 ISSN: 2229-5828 (Print)

Process of discovering information on the internet.

Mohammad Shahid
Senior Lecturer,
Department of Information & Communication Technology,
ISBAT University, Kampla, Uganda,

Abstract: *Web mining is the process of automatically discovering and extracting information from web documents using data mining tools. The numerous forms the present state of web mining state for resent are summarised in this study.*

Keywords: World Wide mining , Browser Exploration, Mining the Structure of the Internet, and Usage of Web Usage Analysis.

Introduction

The amount of information available on the World Wide Web has increased dramatically during the previous decade so now into a significant way of data. As volume for material on the online and the number of persons consuming it grows fast, the web introduces new obstacles for information retrieval. It's nearly not possible for browse with the massive storage of data a user requires. As a result, a search engine is required. Crawlers are used by search engines to collect data, which is then stored in a database on the search engine's side. The search engine scans the local database for a particular user's query and presents the results very quickly.

Related Work

The majority of consumers nowadays use search search engines to locate them desired data. Every search engine has set of features & uses distinct methods with position, level , & display internet content. However, so every of These are the search engines dependent at final index word common & their query languages are artificial in nature, with limited syntax and vocabulary, there are some constraints that search engines cannot overcome, such as narrowing the search area. outcome meaningful information on internet is a difficult task. When communicating with the web, the user may run into the following issues[6].

- Low precision: Due to the irrelevant nature of many search results, today's search technologies have low precision. As a result, acquiring meaningful data becomes more difficult.
- Low recall is caused by the inability to index all of the content on the internet. As a result, finding relevant unindexed material becomes harder.
- Finding new knowledge from the information accessible on the internet is difficult: Because the internet is so large, diversified, and dynamic, it creates scalability, multimedia data, and temporal concerns.
- Personalization of information: Because people's preferences for content and display change as they engage with the web, this problem is typically tied to the kind and presentation of information.
- Learning about users: It's difficult to determine what each user's interests are.
- It is not possible find for graphic information.

Because internet are so large, various & versatile, it creates difficulties of scalability, heterogeneity, and dynamism. We are now drowning in information yet ravenous for knowledge as a result of these traits, making

Publisher: Noida Institute of Engineering & Technology,
19, Knowledge Park-II, Institutional Area, Greater Noida (UP), India.

the web a fruitful field for data mining study due to the vast volume of data available online. In today's world of ever-increasingly large databases, data mining has evolved as a new field. The technique the process of collecting or mining information from data is known as mining of DATA. Data mining LIKELY AS a more significant method for turning data into knowledge. Data Mining is a synonym for Knowledge Discovery from Data, or KDD.

Web Mining

The World Wide Web is a key source of information, and as the volume of information on the web grows dramatically, it poses new issues in terms of retrieval. Web mining [1] is the automated discovery and extraction of information from web publications and services using Data Mining methods. Oren Etzioni was the first person to create the phrase "web mining." Initially, there were two methods of define Web mining is the study of the internet. The first was a "process-centric approach," in which Web Mining was defined in terms of the kind of data being mined [1], while the second was a "data-centric view," in which Web Mining was defined in terms of the type of data being mined mined [2].

The Web Mining Methodology

The following subtasks can be divided into web mining:

The process of retrieving online resources is known as resource discovery. Pre-processing of information is the process of transforming the results of a resource search. Automatically extracting relevant information from freshly found Web sites is known as information extraction. Generalization is the process of identifying common patterns on specific Web sites as well as across several sites[3].

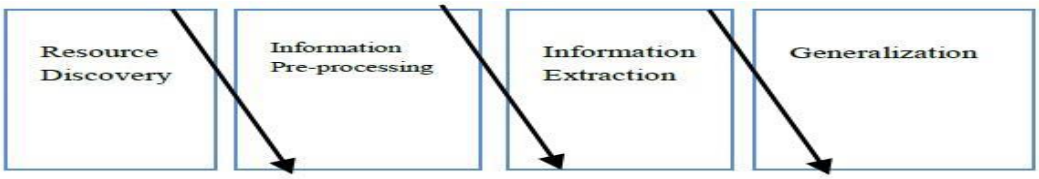

Fig 2.1: Web mining Process

Taxonomy for Web Mining

The web has several features, each of which requires a distinct strategy to mining:

Text makes up web pages. Hyperlinks are used to connect web sites. Web server logs may be used to track user activities. Web content mining, web structure mining, and web use mining are all examples of web mining. are the three categories that result from these three facets [4-7].

Fig 2.2: Web mining Taxonomy

TABLE 2.1 WEB MINING CATEGORIES

	Web Mining			
	Web Content Mining		Web Structure Mining	Web usage Mining
	IR View	DB View		
View of Data	Unstructured Semistructured	Semistructured Web Site	Link Structure	Interactivity
Main Data	Text Documents Hypertext Documents	Hypertext Documents	Graph	Server logs Browser logs
Representation	Bag of Words, n-grams Terms, phrases Concepts or ontology Relational	Edge labeled graph Relational	Link Structure	Relational table Graph
Method	TFIDF Machine Learning Statistical	Proprietary algorithms ILP Association Rules	Proprietary algorithms	Machine Learning Statistical Association Rules
Application Categories	Categorization Clustering Finding extraction rules Finding patterns in text User Modeling	Finding frequent substructures Web site schema discovery	Categorization Clustering	Site construction Adaption and management Marketing User modeling

WCM Exploration of Web Content

Publisher: Noida Institute of Engineering & Technology,
19, Knowledge Park-II, Institutional Area, Greater Noida (UP), India.

The technique of obtaining usable information from the contents of web documents is known as web content mining. The Lists and tables are examples of facts that may be found on a web page, is known as content data. The use of text mining on the internet material received the greatest attention.

Internet & World Wide Web search and indexing tools such as Lycos, Alta Vista, WebCrawler, MetaCrawler, and others give some comfort to users, but they do not provide structural information, nor do they categorise, filter, or interpret texts. These issues have encouraged academics to build more sophisticated information retrieval techniques in recent years.

Intelligent online agents, as well as growing database and data mining tools to provide semi-structured content on the internet with a better level of organisation. Web content mining is the practise of extracting useful information from web content such as text, images, audio, and video. The finding of web resources is a subset of web content mining research, document classification and clustering, and web page information extraction. The study of the web's hyperlink structure is known as web structure mining. It mainly entails analysing a web page's in-links and out-links, and it's been used to rank search engine results. The goal of web use mining is to uncover interesting patterns in search logs or other activity records.

The creation of advanced AI systems that can discover and organise web-based material autonomously or semi-autonomously on behalf of a specific user are becoming more common central to the agent-based approach to web mining. The two perspectives on online content mining are Information Retrieval View and Database View. From the perspective of information retrieval, [6] detailed the study effort done for unstructured and semi-structured data. It indicates that the majority of studies employ a bag of words to represent unstructured text and use single words obtained in the training corpus as features, All of the works use HTML structures within the papers for semi-structured data, and some use a hyperlink structure between the documents.

In terms of the database perspective, in order to improve information management and querying on the web, mining attempts to deduce the structure of a web site in order to convert it into a database. Documents can be represented in a variety of ways.

Model in Vector Space

Clustering is a data mining method used in Online Text Mining to arrange web pages into clusters so that items in the same cluster are similar and objects in other clusters are distinct. The most often used model for representing documents is the model vector document is A vector of extracted keywords from the document, with associated weights representing the significance of the keywords in the text and throughout the whole document collection, is used to represent the theme. A query is represented as a list of keywords with associated weights reflecting the significance of the keywords in the statistically based vector-space technique. A phrase's weight in a document vector can be calculated in a number of ways.

The weight of a phrase is determined by two factors: how frequently the term j occurs in the document I (the term frequency tfi,j) and how often it occurs in the whole document collection (the term frequency tfi,j) (the document frequency dfj). Wi,j= tfi,j * idfj = tfi,j * log N/dfj = tfi,j * log N/dfj = tfi,j * log N/dfj = tfi,j * log N/dfj = tfi,j * log N/dfj = tfi,j * log N/dfj = tfi

The number of documents in the document collection is N, and the inverse document frequency is idf. This strategy gives phrases that appear often in a limited number of papers in the document collection a lot of weight. After determining the word weights, we must compare the document vectors for similarity. The angle between the document vectors and when they are represented in a V-dimensional Euclidean space, where V is the vocabulary size, is determined using a standard similarity measure called as the cosine measure. A document's resemblance to another document Dq is defined as

$$sim(D_q, D_i) = \frac{\sum_{j=1}^{V} w_{q,j} * w_{i,j}}{\sqrt{\sum_{j=1}^{V} w_{q,j}^2 * \sum_{j=1}^{V} w_{i,j}^2}}$$

Publisher: Noida Institute of Engineering & Technology,
19, Knowledge Park-II, Institutional Area, Greater Noida (UP), India.

Where wq,j is the query's weight of word j, and is defined similarly to wi,j. The normalisation factor in the denominator of this equation eliminates the influence of document lengths on document scores. As a result, a document having x, y, z will have the same score as a document containing x, x, y, y, z, z, z, z, z, z, z, z, z, z, We've presented a series of increasingly simpler approximations because the exact vector-space model is difficult to construct.

Mining the Structure of the Internet

Online pages serve as nodes, while hyperlinks serve as edges linking related sites in a typical web graph. The method of finding structural information from the web is known as web structure mining. Web pages may express more information than typical papers because to link architecture. The amount of links pointing to a page indicates its popularity, whereas links pointing to a page indicate the page's subjects or content richness. A frequently mentioned page might be an essential one. As a result, we can mine the web using link architectures. Page Rank and CLEVER are two examples. The core approach for web structure mining is HITS (Hyperlink Induced Topic Search). HITS is broken down into sub-steps: putting together a sub-graph of WWW and computing hubs.

Data Preparation

Preprocessing of data is required in order to transform raw data for future data mining processing. It is divided into sections as follows.Content Preparation: The process of transforming text, images, scripts, and other material into formats that may be used by use mining is known as content preprocessing. It uses many methods, such as stemming and deleting non-relevant terms from online material, such as a, an, and the.Preprocessing of the structure: The linkages between page views constitute the structure of a website. Structure preprocessing can be approached in the same way as content preparation. Each server session, on the other hand, may be necessary to create a site structure that is separate from the others. Web server logs, referral logs, registration files, index server logs, and even data from a previous study are all possible sources of information. all possible inputs for the preparation step. The user session file, transaction file, site topology, and page classifications are the outputs.

Pattern Recognition

most important aspect of web mining. Data mining, machine learning, statistics, and pattern recognition are all examples of data mining techniques all include methods and approaches that may be used in pattern finding. It is divided into sections as follows.

Statistical Analysis: When evaluating the session file, statistical analysts may do several descriptive statistical studies based on various factors. The extracted report can be potentially beneficial for increasing system performance, strengthening system security, facilitating site update tasks, and giving support for marketing decisions by evaluating the statistical information contained in the periodic web system report.

Important Applications Last several years, the industry has built online applications at a considerably quicker rate than web-related technological research the concept of web mining is at the heart of many of these. Here are a few of the most popular apps.

1. Search the Internet

One of the most prominent and commonly used search engines is Google. It allows users to access information from over 10 billion web pages indexed on its server. It is the most successful search engine due to the quality and speed of the search function. Previously, search engines relied solely on online content to deliver relevant pages in response to a query. Google was the first to emphasise the relevance of link structure in extracting data from the internet. PageRank is the core technology of all Google search products, and it leverages structural information from the web graph to produce high-quality results [1]. PageRank quantifies the significance of a website.

2. Understanding the Behavior of Users

Publisher: Noida Institute of Engineering & Technology,
19, Knowledge Park-II, Institutional Area, Greater Noida (UP), India.

Web mining knowledge is the core intelligence underlying Amazon's online book shop features like immediate recommendations, purchasing circles, and wish lists, among others.Amazon gathered this information through tracking individuals' browsing habits and interests.

Conclusion

Web mining is a relatively young and promising research topic that aims to assist consumers in obtaining relevant information from the internet. We offer a brief overview of Web Mining, including its definition, taxonomy, and applications, in this paper.

References

[1] Dr. Mohammad Shahid "KNOWLEDGE DISCOVERY ON THE INTERNET (WEB MINING TOOLAND TECHNIQUE" INDIAN JOURNAL OF RESEARCH(2012)6,
[2] ANVIKSHIKI ISSN 0973-9777 Advance Access publication 20 July. 2012
[3] Dr. Mohammad Shahid" Taxonomies, Challenge And Approaches To Automotive Web Query Classification "2 ND INTERNATIONAL COFFERENCE ON COMPUTER APPLICATION 2012 ICCA'12 PONDICHERRY, INDIA.
[4] RAYMOND KOSALA, HENDRIK BLOCKEEL, Web Mining Research: A Survey, Sigkdd Expirations, Acm Sigkdd, July 2000.
[5] M. KOSHER. ALIKE - Archie-Like Indexing In The Web. In Proc. 1st International Conference On The World Wide Web, Pages 91--100, May 1994.
[6] R. COOLEY, B. MOBASHER, AND J. SRIVASTAVA. Web Mining: Information And Pattern Discovery On The World WideWeb. In Proceedings Of The 9th Ieee International Conference On Tools With Artificial Intelligence (Ictai'97), 1997
[7] R. KOSALA, H. BLOCKEEL. Web Mining Research: A Survey Data & Knowledge Engineering, Volume 53, Issue 3, June 2005, Pages 225-241
[8] NASRAOUI, O. ET AL. , A Web Usage Mining Framework For Mining Evolving User Profiles In Dynamic Web Sites, IeeeSmita Mishra, Somesh Kumar; 'Classification of Bug Report Using Naïve Bayes Classifier with Gain Ratio',Volume No.3,Issue No.2,2015,PP.012-020,ISSN :2229-5828
[9] Sudhansu Aggarwal , Rajnish Kumar Pandey, Raman Chauan; 'Application of Laplace Decomposition Algorithm to Solve the System of Homogeneous Linear Partial Differential Equations',Volume No.3,Issue No.2,2015,PP.021-023,ISSN :2229-5828
[10] Sudhansu Aggarwal, Garima Bindal, Manoj Kumar Yadav; 'Application of Laplace Decomposition Algorithm to Solve the System of Weakly Singular Voltera Integral Equation',Volume No.3,Issue No.2,2015,PP.024-026,ISSN :2229-5828
[11] Sudhansu Aggarwal , Anjana Rani Gupta, Chetan Swarup; 'A New Application of Laplace Decompostion Algorithm for Handling Voltara Integral Equations',Volume No.3,Issue No.2,2015,PP.027-029,ISSN :2229-5828
[12] Deepak Kumar, Anjana Rani Gupta Somesh Kumar; 'Optimation of Bleach Plant in Paper Industry for waste Minimization',Volume No.3,Issue No.2,2015,PP.030-037,ISSN :2229-5828
[13] Ravin Kumar, C.S.Yadav; 'Planned Outline for Indian Sign Language Recogination',Volume No.3,Issue No.2,2015,PP.038-044,ISSN :2229-5828

Publisher: Noida Institute of Engineering & Technology,
19, Knowledge Park-II, Institutional Area, Greater Noida (UP), India.