Webology, Volume 18, Number 4, 2021
ISSN: 1735-188X
DOI: 10.29121/WEB/V18I4/124

A METHOD FOR HANDLING CLOUD COMPUTING USING INTERNET CRAWLERS AND DATA MINING

Prabhdeep Singh[1], Vikas Tripathi[2], Dibyahash Bordoloi[3]

[1]Department of Computer Science & Engineering, Graphic Era Deemed to be University, Dehradun, Uttarakhand India
[2]Department of Computer Science & Engineering, Graphic Era Deemed to be University, Dehradun, Uttarakhand India
[3]Head of the Department, Department of Computer Science & Engineering,Graphic Era Hill University, Dehradun, Uttarakhand India

ABSTRACT

The internet becomes a crucial component of organisation. The massive usage data produced by user and Internet interaction can be collected and used as knowledge for a variety of applications. As internet browsing becomes a more common activity for more people worldwide, business and research communities are paying more attention to the analysis of web site regularities and patterns in user navigation. When it comes to quantity, complexity, semantics, distribution, and processing costs, cloud computing, internet-based computing, cognitive informatics, and computational intelligence represent extraordinarily huge scales of data. Even more information has been gathered about Internet and mobile device users. To enable the capability to make sense of and optimize the use of such large amounts of online data for extracting knowledge and decision-making, we need new tools for such a massive online data mining. It has been demonstrated that visualisation is a useful technique for gaining understanding about cloud computing. Visualization-based data discovery tools are centred on the front end of cloud computing to let organisations more easily and thoroughly explore the data, while technologies like Apache Hadoop and others are developing to solve back-end challenges like storage and processing.

Keywords: Internet crawlers, Visualization, data, mining, Visual data, Internet, Hadoop, Mining

INTRODUCTION

Internet mining was the use of data mining methods to glean knowledge from internet data, such as web documents, linkages between papers, website usage records, and other types of internet data. This method enables a person or corporation to market their products or services while comprehending the dynamics of the market and current online promotions. The World Wide Web Data Mining may encompass content mining, hyperlink structure mining, and use mining due to the abundance of data present in internet pages. All of these methods make an effort to get information from the Internet, use that information to generate some helpful outcomes, and then apply those outcomes to specific real-world issues. Because there are so many FAQs and reviews available, you may forecast your inventory needs using the purchasing trends revealed by online data mining.

Webology, Volume 18, Number 4, 2021
ISSN: 1735-188X
DOI: 10.29121/WEB/V18I4/124

Information retrieval IR systems, including search engines, trafficking controls, where traffic is tracked and observed, use internet mining.

The use of internet data mining technology is expanding opportunities for data collection, but it is also generating significant security-related problems. Internet data mining has contributed to keeping the idea that personal information needs to be secured in the forefront because there is a tonne of personal information available online. In the digital world, data is a collection of information from the grids of internet servers that are typically in a disorganised manner. Today's world has data that was created within the last two years to a degree of about 90% [2]. The vast majority of the material that is accessible on the internet is produced during a specific time period by either individuals, groups, or organisations. As more people use the internet and the internet becomes a more integral component of human activity, the volume of data grows daily. The growth of this data prompts the development of new technologies like cloud computing, which may be used to process, manage, and store very massive datasets. There are many different sources and forms of data, including semi-structured and unstructured data in addition to structured standard relational data. The application software, Apache Hadoop, is explored and reviewed below. Internet log mining has been the subject of numerous research projects. proposed the smart miner framework, which uses Hadoop Map reduce to extract user behaviour from internet records.

DataMining Process:

As a result, it is frequently used to refer to it as a basic component of knowledge discovery. So that the pertinent data may be recovered, the data is combined and cleansed. Data mining delivers uncovered information that is understandable to domain specialists as well as data mining analysts, who may utilise it to generate recommendations that can be put into practise. Analysis of genetic patterns, graph mining in finance, expert systems to obtain appropriate advice, and customer behaviour in marketing are examples of successful data mining applications. Traditional data mining makes use of tabularized relational tables, spreadsheets, or flat files that contain structured data.

Internet Mining

Internet mining is the technique of automatically locating and extracting useful information from documents and services on the World Wide Web using data mining technologies. Internet mining is the practise of using data mining techniques to draw knowledge from internet information, including documents, links between papers, records of website activity, etc. Internet mining can be broadly divided into three sorts: internet activity mining, internet structural mining, and online entertainment mining. This is based on the types of data that need to be mined.

1. Internet Content Mining

The process of removing valuable information from the text of web sites and documents is known as "internet content mining." The group of information that makes up a web page's content data. Multimedia data is data that includes text, graphics, audio, video, and structured records like lists and tables. The use of text mining for internet material has been the subject of the most study. Text mining addresses a number of issues, such as subject finding and tracking, association pattern extraction, online document clustering, and web page classification.

Webology, Volume 18, Number 4, 2021
ISSN: 1735-188X
DOI: 10.29121/WEB/V18I4/124

2. Internet Structure Mining Internet pages serve as the nodes in a typical internet graph and hyperlinks serve as the edges linking related pages. Finding structure-related information from the internet is a process called "internet structure mining." The websites themselves, the links that connect them, and the route that internet users follow to get to a specific website are all included in the internet resources that have been evaluated [6]. In addition, the various HTML and XML tags on an Internet page can be used to organise the content of the page in a tree-structured format. 3. Usage data records both an internet user's identification or place of origin and their online surfing habits. Depending on the type of usage data taken into account, internet usage mining can be divided into several categories: The content of the real information used for internet usage mining and the anticipated information that will be obtained from it both provide unique challenges. [7]

CHARACTERISTICS OF INTERNET DATA

The aspects of internet data that make it challenging to mine include the following: 1. The majority of the information on the Internet is in heterogeneous form. This increases the difficulty of integrating data from numerous pages. 2. A large amount of related information is available on the Internet. Within a site and between sites, there are links between Internet pages. When present across distinct sites, hyperlinks reflect an implicit delegation of authority to the destination pages. Within a single site, hyperlinks serve as means for organising content.

3. The internet was a noisy place for information. There are two basic origins of this noise. First off, a normal Internet page includes a variety of pieces of data, such as the website's primary content, navigation links, adverts, copyright notices, privacy rules, etc. But only a portion of the information is helpful for a specific application. The remainder is regarded as noise. The noise should have to be eliminated in order to undertake fine-grain Internet data analysis and data mining. Second, because there is no information quality control on the Internet, meaning that anybody can post nearly whatever they want, a lot of the information there is of poor quality, inaccurate, or even deceptive.

4. The Internet offers services as well. The majority of commercial Internet sites enable users to carry out practical tasks on their websites, including as making purchases, paying bills, and completing forms that transmit sensitive personal data from one location on the internet to another.
5. Because it is constantly changing, the material on the Internet is dynamic. Keeping up with the shift and tracking the change are significant problems for many applications.

6. A virtual community exists on the Internet. The Internet is about interactions between people, companies, and automated systems in addition to data, information, and services. One can instantaneously and easily contact with anyone anywhere in the world, at any time, and share their opinions on anything.

All of these qualities create opportunities and problems for information mining and discovery on the Internet. As a result of the vast amount of data that cloud computing requires, we are primarily concentrating on internet data mining approaches in this study for mining of photos, d audios, anvideos. It's necessary in today's world to mine and manage so much cloud computing. METHODS FOR HANDLING BIG INTERNET DATA Visual Internet Mining Architecture Figure1 below depicts the architecture for implementing visual internet mining. For examination, we choose one or

Webology, Volume 18, Number 4, 2021
ISSN: 1735-188X
DOI: 10.29121/WEB/V18I4/124

more internet sites. Internet sites and internet server log files are the system's input. The local file system or a remote internet server can be used to access the internet log. The pages of the website are retrieved using an internet robot (internetbot). In the background, a sessionizer downloads Internet Server Log files, processes them, and creates a LOGML file.

• The Integration Engine is a collection of applications for the extraction, cleaning, transformation, and integration of data. loaded into the database after which XGML graphs are produced. The engine works with both open-source and proprietary data formats, using XQuery, XSLT, and Regular Expressions.

• The integration engine's and the database's transformation system are being worked on to improve performance. By removing user sessions from internet records, this produces data that is essentially related to a particular user. Then, using cSPADE, user sessions are translated into a unique format for Sequence Mining.

• With a predetermined minimum support, outputs are frequently contiguous sequences. Non-maximal frequent sequences are eliminated when these are loaded into a database, so we only take into account the maximal frequent contiguous sequences. Later, other queries are run on this data based on various criteria, such as the support for each pattern, the length of the patterns, etc.

Handling Big Internet Data with Hadoop Map reduce

A typical database tool cannot handle the increasing and expanding dataset of cloud computing. The expansion of data is exceeding the expectations of the average internet user since internet use and the internet are becoming daily concerns for many people. Such a big amount of data also contributes to the cloud computing issue. Hadoop utilises cloud computing, where a vast amount of data is handled utilising a cluster of affordable hardware. Internet server logs are vast volumes of semi-structured flat text files produced by computers. Map reduction uses it effectively since it processes each line separately. Data processing platform Apache Hadoop MapReduce is utilised in fully distributed and pseudo-distributed modes. The framework successfully recognises the internet user's session in order to identify the distinct users and pages that the users have accessed. In order to provide a statistical report based on the overall daily visit count, the identified session is examined in R. The presented work has improved time efficiency, storage, and processing performance when compared to non-hadoop approaches in a java context.

On a network of inexpensive hardware, A versatile platform is offered by Hadoop for handling massive amounts of data and computing.

Applications may now access petabytes of data and thousands of computationally independent processors. Hadoop's basic tenet is that calculations should be moved along with the data, not the other way around. The massive amount of incoming data is divided up into smaller chunks using Hadoop so that each could be processed independently on various computers. Hadoop uses the MapReduce programming model to enable parallel processing. Large data sets can be processed and produced using the Map Reduce programming model and its related implementation..

Webology, Volume 18, Number 4, 2021
ISSN: 1735-188X
DOI: 10.29121/WEB/V18I4/124

CONCLUSION:

The internet has now become a crucial component of many businesses, organisations, and everyday people in the modern, technologically evolved world. We have researched the qualities of internet data because it comes in so many various formats. As mining specific data from the internet is crucial, we have researched two efficient methods to do so: the first uses Apache Hadoop Map Reduce, and the second uses Visual Internet Mining, a visualization-based method (VWM).

REFERENCE:

1. Ristevski, Blagoj, and Ming Chen. "Cloud computing in medicine and healthcare." Journal of integrative bioinformatics 15.3 (2018).
2. Majumdar, Jharna, Sneha Naraseeyappa, and Shilpa Ankalaki. "Analysis of agriculture data using data mining techniques: application of cloud computing." Journal of Cloud computing 4.1 (2017): 1-15.
3. Choi, Tsan-Ming, Stein W. Wallace, and Yulan Wang. "Cloud computing in operations management." Production and Operations Management 27.10 (2018): 1868-1883.
4. Kumar, Manish, Rajesh Bhatia, and Dhavleesh Rattan. "A survey of Internet crawlers for information retrieval." Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery 7.6 (2017): e1218.
5. Ghavami, Peter. Cloud computing methods: analytics techniques in data mining, deep learning and natural language processing. Walter de Gruyter GmbH & Co KG, 2019.
6. Hwang, Kai, and Min Chen. Big-data analytics for cloud, IoT and cognitive computing. John Wiley & Sons, 2017.
7. Ageed, Zainab Salih, Rowaida Khalil Ibrahim, and M. A. Sadeeq. "Unified ontology implementation of cloud computing for distributed systems." Current Journal of Applied Science and Technology 39 (2020): 82-97.
8. Stergiou, Christos L., et al. "Secure machine learning scenario from big data in cloud computing via internet of things network." Handbook of computer networks and cyber security. Springer, Cham, 2020. 525-554.
9. Hussain, Yasir, et al. "Context-aware trust and reputation model for fog-based IoT." IEEE Access 8 (2020): 31622-31632.
10. Mohammed, Fathey, et al. "Cloud computing services: taxonomy of discovery approaches and extraction solutions." Symmetry 12.8 (2020): 1354.
11. Tyagi, Himani, and Rajendra Kumar. "Cloud computing for iot." Internet of Things (IoT). Springer, Cham, 2020. 25-41.
12. Rekha, G., Amit Kumar Tyagi, and Nandula Anuradha. "Integration of fog computing and internet of things: an useful overview." Proceedings of ICRIC 2019. Springer, Cham, 2020. 91-102.
13. Deshpande, Prachi. "Cloud of everything (CLeT): the next-generation computing paradigm." Computing in Engineering and Technology. Springer, Singapore, 2020. 207-214.
14. Agarwal, Vidushi, Ashish K. Kaushal, and Lokesh Chouhan. "A survey on cloud computing security issues and cryptographic techniques." Social networking and computational intelligence. Springer, Singapore, 2020. 119-134.
15. Albdour, Layla, Saher Manaseer, and Ahmad Sharieh. "IoT crawler with behavior analyzer at fog layer for detecting malicious nodes." International Journal of Communication Networks and Information Security 12.1 (2020): 83-94.