JETIR.ORG ISSN: 2349-5162 | ESTD Year : 2014 | Monthly Issue
JOURNAL OF EMERGING TECHNOLOGIES AND INNOVATIVE RESEARCH (JETIR)
An International Scholarly Open Access, Peer-reviewed, Refereed Journal

EXPLORING THE POTENTIAL OF WEB MINING BASED INFOTAINMENT SYSTEMS IN VANETS: CHALLENGES AND OPPORTUNITIES

[1] Mr. K. Senthilkumar, [2]Dr. S. Sathiyabama
[1]Research Scholar, [2]Assistant Professor
[1]Department of Computer Science,[2]Department of Computer Science
[1]Thiruvalluvar Government Arts College,[2]Thiruvalluvar Government Arts College
[1,2]Rasipuram, Namakkal Dt, India

Abstract: The use of web mining techniques to gather information from the internet in order to improve the traffic updates, to gain insights into traffic patterns, road conditions route recommendations, emergency services, driver assistance and informational experience of passengers in VANET-connected vehicles. Data of Interest can be achieved by analyzing various online sources such as web pages and social media to obtain information on destinations, current events, and other relevant data which can be utilized to offer personalized and current information to passengers. Using Web crawling, Web scraping methods to create more advanced infotainment systems that enhance the overall passenger experience while traveling. Moreover, by analyzing data obtained from VANETs, infotainment systems can also offer other benefits such as entertainment and informational.

Keywords—Web mining, VANETs, Web crawling, Web scraping, Deep Crawling, Infotainment systems, Web-based personalization.

I.INTRODUCTION

Web mining [1] is the practice of collecting and analysing data from the internet to gain valuable knowledge. Web mining involves a variety of techniques and methods that are used to gather, process and make sense of the massive amount of information available on the web. It is usually divided into three main categories:

Web Content Mining: This involves extracting useful information from web pages such as text, images and videos which can be used for tasks such as text summarization, sentiment analysis, and topic modelling.

Web Structure Mining: This involves extracting information about the structure of the web such as links between web pages and the organization of information on a website, which can be used for tasks such as web page classification, link analysis, and web page clustering.

Web Usage Mining: This involves extracting information about how people interact with the web, such as their browsing habits, search queries, and clickstream data, which can be used for tasks such as web personalization, recommendation systems, and user behaviour analysis[2].

Web Scraping and web crawling are used to find the specific, personalised information from different sources in internet. Both web scraping and web crawling are used to gather information from the internet, and they are considered as techniques of web mining[3]. Figure 1 shows the Web mining categories.

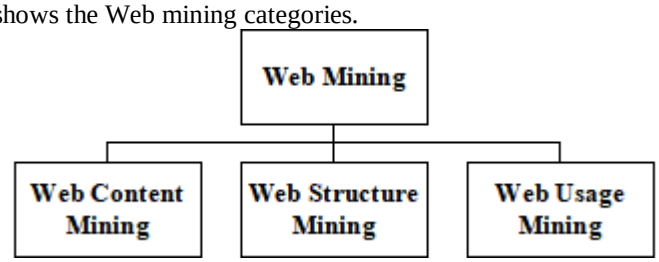

Figure 1. Web mining Categories

VANETs, or Vehicular Ad-hoc Networks, [4] are a type of mobile ad-hoc network (MANET) that allow vehicles to communicate with each other and with roadside infrastructure through wireless communication technologies. VANETs have been gaining increasing attention in recent years for their potential to improve road safety, traffic efficiency, and the overall driving experience.

One of the main advantages of VANETs is their ability to provide real-time traffic information and alerts to drivers. This can help drivers avoid accidents and congestion, and can also improve traffic flow by allowing vehicles to adjust their speed and route in response

to changing conditions. Additionally, VANETs can also provide other services such as navigation, weather updates, and emergency notification.

Another important aspect of VANETs is their potential to improve road safety[5]. VANETs can provide early warning of potential hazards and can also facilitate communication between vehicles, which can help to reduce the risk of accidents. Additionally, VANETs can also provide other safety features such as automatic emergency braking and collision avoidance systems.

Table 1List of Web scraping tools

Tool	Description
Scrapy	An open-source and collaborative web scraping framework for Python. It allows for easy data extraction and is suitable for large-scale scraping projects.
Beautiful Soup	A Python library for pulling data out of HTML and XML files. It allows for easy navigation and searching of the parse tree
Selenium	A browser automation tool that can be used for web scraping. It allows for the execution of JavaScript and is useful for scraping dynamic websites.
PyQuery	A Python library for parsing and manipulating HTML and XML documents. It is similar to jQuery and allows for easy traversal of the DOM.
Parsehub	A cloud-based web scraping tool that supports data extraction from dynamic websites and can handle JavaScript, AJAX, cookies and sessions.
Octoparse	A data scraping software that can extract data from websites automatically. It supports both static and dynamic websites, and can handle AJAX, JavaScript, cookies, and sessions.
WebHarvy	A visual web scraper that allows you to scrape data from websites without writing any code. It supports scraping from multiple pages, and can handle AJAX, JavaScript and CAPTCHAs.

II. WEB SCRAPING

Web scraping [7][10]is the process of automatically extracting information from a website. It involves making HTTP requests to a website's server and then parsing the HTML response to extract the data of interest. Table.1 lists various web scraping tools with a brief description

Web Scraping rule: A web scraping rule for retrieving traffic information from various sources could include the following instructions:

- Scrape traffic data from websites such as Maps, Openstreetmap.
- Extract information such as current traffic conditions, traffic incidents, and estimated travel times.
- Extract data in JSON or XML format.
- Use a specific set of CSS selectors or Xpaths to extract the relevant information from the website's HTML structure.
- Schedule the scraping to run at regular intervals (e.g. every 30 minutes) to ensure that the traffic information is up-to-date.
- Use a tool such as Scrapy, BeautifulSoup, or Selenium to automate the scraping process.
- Store the extracted data in a database such as MongoDB or Cassandra for further analysis.
- Use a library such as Pandas or NumPy to process and analyze the data.
- Use visualization libraries such as Matplotlib or Plotly to display the traffic information in a user-friendly format.
- It is also important to comply with the website's terms of service and robots.txt file to avoid any legal issues.
- The rule can be to scrape traffic information periodically and extract the current traffic conditions, incidents and estimated travel time, store the data in a MySQL database, and use the data to update a traffic report frequently.

Figure 2. Process of Web Crawler

III.WEB CRAWLING

Web crawling [6] is the process of automatically visiting web pages and following links to other pages. It is used to discover new pages and update existing pages, and it is a key component of web scraping and web mining.web crawler, also known as a spider or robot, is a program that performs the crawling process. Figure 2 depicts the process of Web Crawler.

Web crawling rule

A web crawling rule for retrieving specific traffic information from various sources could include the following:

- Start with a seed URL
- Crawl the website and follow all the links within the domain, but limit the crawl to a maximum depth of 2 levels.
- Extract information such as current traffic conditions, traffic incidents, and estimated travel times.
- Schedule the crawling to run at regular intervals to ensure that the traffic information is up-to-date.
- Use a tool such as Scrapy, BeautifulSoup, or Selenium to automate the crawling process.
- Use a distributed crawling system like Nutch or Heritrix to handle large-scale crawling.
- Store the extracted data in a database such as MongoDB or Cassandra for further analysis.
- Use a library such as Pandas or NumPy to process and analyze the data.
- Use visualization libraries such as Matplotlib or Plotly to display the traffic information in a user-friendly format.
- Process of web crawling for traffic related information
- The process of web crawling for traffic related information from social media and other sources can be broken down into the following steps:
- Start with seed URLs: The crawler starts with seed URLs, which can be the homepage of social media websites such as Twitter and Facebook, or websites of transportation agencies, news websites, and blogs that cover traffic-related information.
- Discover new links: The crawler visits the seed URLs and discovers new links by following the “href” attribute of the <a> tags, as well as by searching for specific keywords related to traffic information such as "traffic conditions", "traffic incidents", etc.
- Filter the links: The crawler filters the discovered links to include only those that are relevant to traffic information.
- Visit new links: The crawler visits the filtered links and repeats the process of discovering new links.
- Extract the data: As the crawler visits the links, it extracts the relevant data, such as traffic conditions, incidents, and estimated travel times, as well as information on the source (e.g. user name, location, timestamp)
- Parse the data: The data is parsed and cleaned to remove any irrelevant information, and it is stored in a structured format such as JSON or XML
- Store the data: The parsed data is stored in a database or file system for further analysis.
- Limit the crawl: The crawl can be limited by setting a maximum depth for the crawl, a maximum number of links to be visited, or a time limit for the crawl.
- Duplicate handling: The crawler should handle duplicates and not crawl the same page more than once.
- Robot.txt handling: The crawler should respect the website's terms of service and robots.txt file and not crawl the pages that are disallowed by the website.
- Loop until all pages are visited: The process of discovering new links, visiting new links, extracting data, parsing data, and storing data continues until all the links on the website have been visited or the crawl is stopped due to the limit set.

Table 2. Data for Weather Condition and Location of Vehicles

Date	Time	Location	Type of Vehicle	No.of Vehicles	Weather Condition
01/02/2023	8:00 AM	Rajpath, New Delhi	Cars	50	Clear skies
01/02/2023	12:00 PM	Marine Drive, Mumbai	Trucks	15	Cloudy
01/02/2023	4:00 PM	MG Road, Bangalore	Bikes	25	Rainy
01/02/2023	7:00 PM	Park St, Kolkata	Cars	40	Snowy

Table .2 shows different types of Vehicles from different locations including the weather condition collected from various sources by web mining techniques and that information processed and analysed, then shared with VANETs to facilitate the drivers or passengers in various aspect related to their data of interest. Figure 3 represents the data shown in Table 2. This is with minimum attributes for sample dataset.

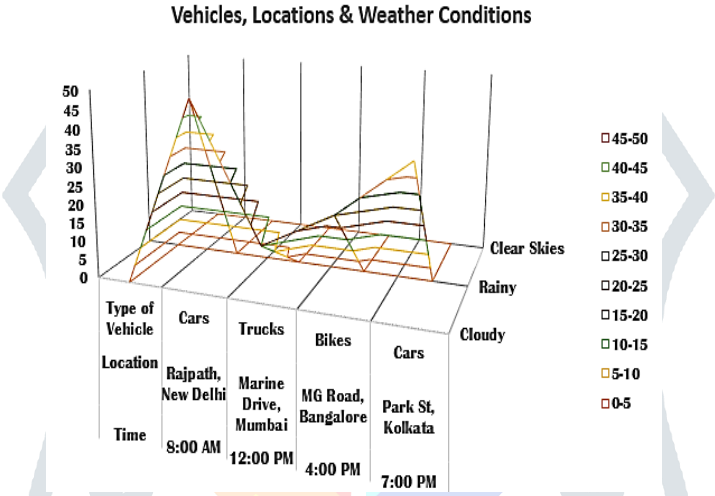

Figure 3 Identifying the Weather Condition

IV. OPPORTUNITIES

Edge and Fog computing: With the increasing amount of data generated by VANETs, edge and fog computing will play a vital role in processing and analysing this data in real-time[9]. This will enable new and enhanced VANET applications such as real-time traffic management and efficient fleet management.

Cyber-Physical Systems (CPS): Cyber-Physical Systems (CPS) will play an important role in VANETs, enabling vehicles to sense, process, and act on their environment. This will enable new and enhanced VANET applications such as advanced driver-assistance systems (ADAS) and autonomous driving

Cooperative Intelligent Transport Systems (C-ITS): Cooperative Intelligent Transport Systems (C-ITS) is a technology that enables vehicles to communicate with each other and with the infrastructure. This will enable new and enhanced VANET applications such as advanced driver-assistance systems (ADAS) and autonomous driving.

Web-based personalization: In the future, web mining will be used to create personalized web experiences[10] for users. This will involve the use of machine learning algorithms to predict user preferences, and then tailoring the web content to those preferences.

V. CHALLENGES

Security: Security is a major concern in VANETs as they are vulnerable to various types of attacks. One of the common techniques used for securing VANETs is the use of digital certificates and public key infrastructure (PKI) for authenticating vehicles and establishing secure communication channels. Another technique is the use of intrusion detection systems (IDS) to detect and prevent attacks.

Congestion Control: VANETs are subject to congestion due to the high mobility of vehicles, which can lead to packet losses and delays. One of the techniques used for congestion control in VANETs is the use of adaptive rate control, which adjusts the transmission rate of vehicles based on the current network conditions.

Positioning: VANETs use positioning techniques to determine the location of vehicles. GPS (Global Positioning System) is the most widely used positioning technique, but it is not reliable in urban areas due to the presence of tall buildings. Other techniques such as wireless local positioning systems (WLPS) and VANET-based positioning are used to overcome this limitation.

Real-time crawling [8]: As the web is becoming more dynamic and the need for real-time data is increasing, web crawlers are being adapted to collect data in real-time. This allows for more timely insights and faster decision-making.

High mobility: Vehicles in VANETs are highly mobile and can change their position rapidly, which makes it difficult to maintain the connectivity and consistency of data.

Limited resources: Vehicles in VANETs have limited resources, such as battery power, storage, and processing capabilities, which can

limit the amount of data that can be collected and analyzed.

Scalability: VANETs can have a large number of vehicles, which can make it difficult to scale web mining techniques to handle the volume of data.

Network congestion: VANETs can experience network congestion due to high traffic density, which can reduce the efficiency of web mining techniques and make it difficult to collect and analyse data in real-time.

VI. CONCLUSION

In conclusion, the web mining techniques, web scraping and web crawling are used to gather information from the internet, to greatly improve the traffic updates, gain perceptions into traffic patterns, road conditions, route recommendations, emergency services, driver assistance, and informational experience of passengers in VANET-connected vehicles. By analysing various online sources such as web pages and social media, relevant and personalized information can be obtained and utilized to enhance the overall passenger experience while traveling. Furthermore, by analysing data obtained from VANETs, infotainment systems can also offer other benefits such as entertainment and additional information to passengers. Overall, web mining techniquescan play a crucial role by both overcoming challenges and providing opportunities for the future of transportation and the passenger experience in VANET-connected vehicles.

REFERENCES

[1] K.Jayamalini and M. Ponnavaikko, "Research on web data mining concepts, techniques and applications," 2017 International Conference on Algorithms, Methodology, Models and Applications in Emerging Technologies (ICAMMAET), Chennai, India, 2017, pp. 1-5, doi: 10.1109/ICAMMAET.2017.8186676.

[2] S. P. Singh and Meenu, "Analysis of web site using web log expert tool based on web data mining," *2017 International Conference on Innovations in Information, Embedded and Communication Systems (ICIIECS)*, Coimbatore, India, 2017, pp. 1-5, doi: 10.1109/ICIIECS.2017.8275961.

[3] V. Rana and G. Singh, "Analysis of web mining technology and their impact on semantic web," *2014 Innovative Applications of Computational Intelligence on Power, Energy and Controls with their impact on Humanity (CIPECH)*, Ghaziabad, India, 2014, pp. 5-11, doi: 10.1109/CIPECH.2014.7019035.

[4] P. Mutalik and V. C. Patil, "A survey on vehicular ad-hoc network [VANET's] protocols for improving safety in urban cities," *2017 International Conference On Smart Technologies For Smart Nation (SmartTechCon)*, Bengaluru, India, 2017, pp. 840-845, doi: 10.1109/SmartTechCon.2017.8358491.

[5] A. B. Prasetijo, S. S. Alwakeel and H. A. Altwaijry, "Effects of VANET's attributes on network performance," 2014 The 1st International Conference on Information Technology, Computer, and Electrical Engineering, Semarang, Indonesia, 2014, pp. 303-308, doi: 10.1109/ICITACEE.2014.7065761.

[6] I. G. S. Rahayuda and N. P. L. Santiari, "Crawling and cluster hidden web using crawler framework and fuzzy-KNN," *2017 5th International Conference on Cyber and IT Service Management (CITSM)*, Denpasar, Indonesia, 2017, pp. 1-7, doi: 10.1109/CITSM.2017.8089225.

[7] Đ. Petrović and I. Stanišević, "Web scrapping and storing data in a database, a case study of the used cars market," 2017 25th Telecommunication Forum (TELFOR), Belgrade, Serbia, 2017, pp. 1-4, doi: 10.1109/TELFOR.2017.8249451.

[8] M. S. Islam, K. S. Ali, and M. A,. Imran in “Real-time Web crawling: An overview" Journal of Computer Science, vol. 9, no. 1, 2013.

[9] H. Gao, X. Shen, and X. Lin "Security in VANETs: A comprehensive survey" in IEEE Communications Surveys & Tutorials, vol. 18, no. 4, 2016.

[10] H. Herlawati, R. TriasHandayanto, I. Ekawati, K. I. Meutia, J. Asian and U. Aditiawarman, "Twitter Scrapping for Profiling Education Staff," *2020 Fifth International Conference on Informatics and Computing (ICIC)*, Gorontalo, Indonesia, 2020, pp. 1-6, doi: 10.1109/ICIC50835.2020.9288607.