WEBSITE CATEGORIZATION USING DATA MINING

[1]Abhishek Venkateswaran, [2]Harshal Solanki, [3]Kaustubh Rasam, [4]Tushar Shinde, [5]Aarti Puthran
[1]Student, [2]Student, [3] Student, [4] Student, [5]Assistant Professor
Department of Information Technology,
Shree L R Tiwari College of Engineering, Thane, India

Abstract: The number of websites as well as the categories of the websites has increased day by day. The content and category of information are overflowing the Internet channel. These problems arise because there is lots of website information generated. This problem is faced by the user to find out some information or buy some product through the online. Finding the right information from almost a billion of websites is considerably hard, but finding the perfect and appropriate one is even harder. Due to anonymity of URLs it makes it difficult for the user/customer to identify the category of the website as well as predict the authenticity of the website without visiting the URL. This makes the user vulnerable to various kinds of attacks and spams thus causing substantial harm to the user and their credentials. The solution of the given problem is "Website Categorization ". This technique will provide the efficient result to the user. By using website categorization, it will provide the efficient or accurate result about the URL's to the user in small amount of time.

Index Terms **- Data mining, Categorization, Clustering, Data Scrapping, Data Crawling.**

I. INTRODUCTION

What is Website Categorization?

The Internet is a massive place. At any given time, the indexed Internet contains around 4.5 billion unique web pages and billions more subpages. Website Categorization quite simply places this extensive number of websites into appropriate categories. For example, Facebook.com would be placed into the category Social Networking and Bovada.com would be categorized as a gambling website. This can be immensely helpful for companies, as managing categories can not only increase employee productivity but also help detect and prevent insider threats.

Why is Website Categorization Valuable?

1. Managing User Browsing Most organizations have an acceptable use policy in place when it comes to employees and web browsing. Business purposes is a very vague term, and, as you can imagine, some websites are not always what they seem. Some organizations are very strict and may block access to certain websites using a web content filter. Other companies may be very loose and allow users to use their best judgment. In either practice, security or IT teams often need the ability to detect when a user is going to a website not related to business needs. It would be a nearly impossible task for teams to categorize every website in creation, so these teams really need services and products to do it for them.

2. Detecting & Preventing Insider Threats While some companies choose to block sites such as social media or job searching sites, there are other sites on the Internet that companies need to block for security reasons. Those often include malicious, adult content sites – a common criminal practice to obtain sensitive information by tricking employees into becoming an insider threat. There are numerous web pages available on internet. More and more are becoming available every day. Such web pages represent a massive amount of information that is easily crawl able. Seeking category in this huge collection requires much of the work which can be automated through classification technique. The accuracy and our understanding of such systems greatly influence their usefulness. The task of data mining is to automatically classify documents into predefined classes based on their content. With the existing algorithms, a number of newly established processes are involving in the automation of text classification. The most common techniques used for this purpose include Association Rule Mining, Implementation of TF-IDF Classifier. Association rule mining finds interesting association or correlation relationships among a large set of data items. The discovery of these relationships among huge amounts of transaction records can help in many decision making process. Although the TF-IDF works well in many studies, it requires a large number of training documents for learning accurately.

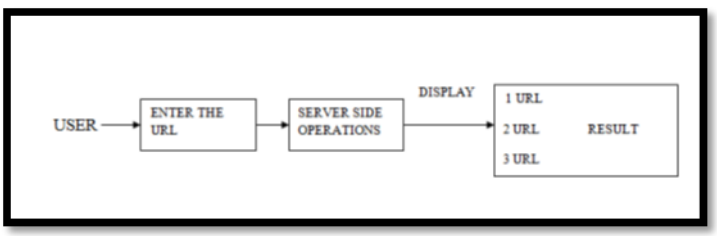


Fig.1 Flow of Website Categorization using DM

II. LITERATURE SURVEY

Classification of Websites into predefined categories has always been considered as a vital method to manage and process a vast amount of documents in digital forms that are widespread and continuously increasing. People largely access information from these online sources rather than being limited to paper sources like books, magazines, newspapers etc. But the main problem

is that this enormous information lacks organization which makes it difficult to manage. Web classification is recognized as one of the key techniques used for organizing such kind of digital data. We have studied the existing work in the area of Web classification which will allow us to have a fair evaluation of the progress made in this field till date. **Arul Prakash Asirvatham, Kranthi Kumar. Ravi, Web Page Categorization based on Document Structure, 2000. (3)** Reviewed web categorization algorithms. Major classification has been divided into five classes: Supervised classification -This is useful when classes have been predefined. Semi-Supervised Classification-Semi-supervised learning is a class of supervised learning techniques that also make use of unlabeled data for training – typically a small amount of labeled data with a large amount of unlabeled data. Unsupervised Classification - Clustering algorithms can group web documents without any pre-defined framework, or background information. Meta tags based categorization. - Using Meta tag attributes for web documents classification. The assumption that author of document will use correct keywords in the Meta tags is not always true. Text content based categorization.-A database of keywords in a category is prepared and commonly occurring words are removed from this list. The remaining words can be used for classification. The web is highly dynamic; lots of pages are included, redesigned and expelled each day and it handles tremendous arrangement of data subsequently there is an entry of numerous number of issues or problems. Typically, web information is high dimensional, restricted query interface, keyword arranged search and constrained customization to individual clients. Because of this, it is extremely hard to locate the important data from the web which may make new issues. Web mining systems are Association Rules, Clustering and Classification which are utilized to comprehend the client behavior, assess a specific site by utilizing conventional data mining parameters.

III. PROPOSED SYSTEM

The main idea behind this project is to develop an interface where users can enter their desired URL for checking their respective categories. This will thus avoid users from directly visiting a particular unknown website. Thus the user will get a brief idea about the website prior visiting it directly. This will avoid and prevent the user from visiting a malicious or spam websites.

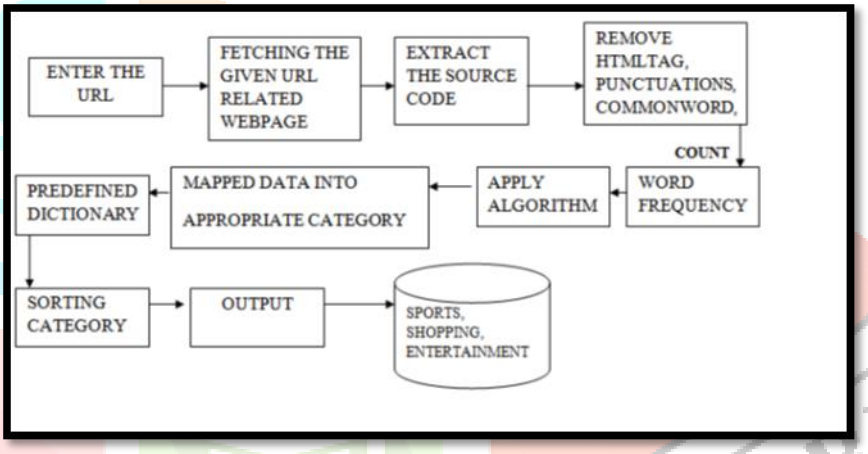


Fig.2. System Architecture

In above system overflow, when user enters an URL and searches then it fetches, crawls and scraps the content and after clustering the user is provided with the appropriate category for the respective website or URL searched.

A. Web crawler

A web crawler (also known as a web spider or web robot) is a program or automated script which browses the World Wide Web in a methodical, automated manner. This process is called Web crawling or spidering. Many legitimate sites, in particular search engines, use spidering as a means of providing up-to-date data. Web crawlers are mainly used to create a copy of all the visited pages for later processing by a search engine that will index the downloaded pages to provide fast searches. Crawlers can also be used for automating maintenance tasks on a Web site, such as checking links or validating HTML code. Also, crawlers can be used to gather specific types of information from Web pages, such as harvesting e-mail addresses (usually for spam). A Web crawler starts with a list of URLs to visit, called the seeds. As the crawler visits these URLs, it identifies all the hyperlinks in the page and adds them to the list of URLs to visit, called the crawl frontier. URLs from the frontier are recursively visited according to a set of policies. If the crawler is performing archiving of websites it copies and saves the information as it goes. The archives are usually stored in such a way they can be viewed, read and navigated as they were on the live web, but are preserved as „snapshots'. The archive is known as the repository and is designed to store and manage the collection of web pages. The repository only stores HTML pages and these pages are stored as distinct files.

A repository is similar to any other system that stores data, like a modern day database. The only difference is that a repository does not need all the functionality offered by a database system. The repository stores the most recent version of the web page retrieved by the crawler. Any labels that humans can generate, any outcomes you care about and which correlate to data, can be used to train a neural network.

B. Web scraping:

Web scraping, web harvesting, or web data extraction is data scraping used for extracting data from websites. Web scraping software may access the World Wide Web directly using the Hypertext Transfer Protocol, or through a web browser. While web scraping can be done manually by a software user, the term typically refers to automate processes implemented using

a bot or web crawler. It is a form of copying, in which specific data is gathered and copied from the web, typically into a central local database or spreadsheet, for later retrieval or analysis.

Web scraping a web page involves fetching it and extracting from it. Fetching is the downloading of a page (which a browser does when you view the page). Therefore, web crawling is a main component of web scraping, to fetch pages for later processing. Once fetched, then extraction can take place. The content of a page may be parsed, searched, reformatted, its data copied into a spreadsheet, and so on. Web scrapers typically take something out of a page, to make use of it for another purpose somewhere else. An example would be to find and copy names and phone numbers, or companies and their URLs, to a list (contact scraping).

Web scraping is used for contact scraping, and as a component of applications used for web indexing, web mining and data mining, online price change monitoring and price comparison, product review scraping (to watch the competition), gathering real estate listings, weather data monitoring, website change detection, research, tracking online presence and reputation, web mashup and, web data integration.

C. Term frequency

Suppose we have a set of English text documents and wish to rank which document is most relevant to the query, "the brown cow". A simple way to start out is by eliminating documents that do not contain all three words "the", "brown", and "cow", but this still leaves many documents. To further distinguish them, we might count the number of times each term occurs in each document; the number of times a term occurs in a document is called its term frequency. However, in the case where the length of documents varies greatly, adjustments are often made (see definition below). The first form of term weighting is due to Hans Peter Luhn (1957) which may be summarized as:

The weight of a term that occurs in a document is simply proportional to the term frequency. In the case of the term frequency tf (t,d), the simplest choice is to use the raw count of a term in a document, i.e. the number of times that term t occurs in document d. If we denote the raw count by ft, d, then the simplest tf scheme is tf (t,d) = ft,d. Other possibilities include

- Boolean "frequencies": tf(t,d) = 1 if t occurs in d and 0 otherwise;
- term frequency adjusted for document length : ft,d ÷ (number of words in d)
- logarithmically scaled frequency: tf(t,d) = log (1 + ft,d);[6]
- augmented frequency, to prevent a bias towards longer documents, e.g. raw frequency divided by the raw frequency of the most occurring term in the document:

The first part of the formula tf (t,d) is simply to calculate the number of times each word appeared in each document. Of course, as with common text mining methods: stop words like “a”, “the”, and punctuation marks will be removed beforehand and words will all be converted to lower cases.

IV. SYSTEM ELEMENTS

A. Functional Requirements

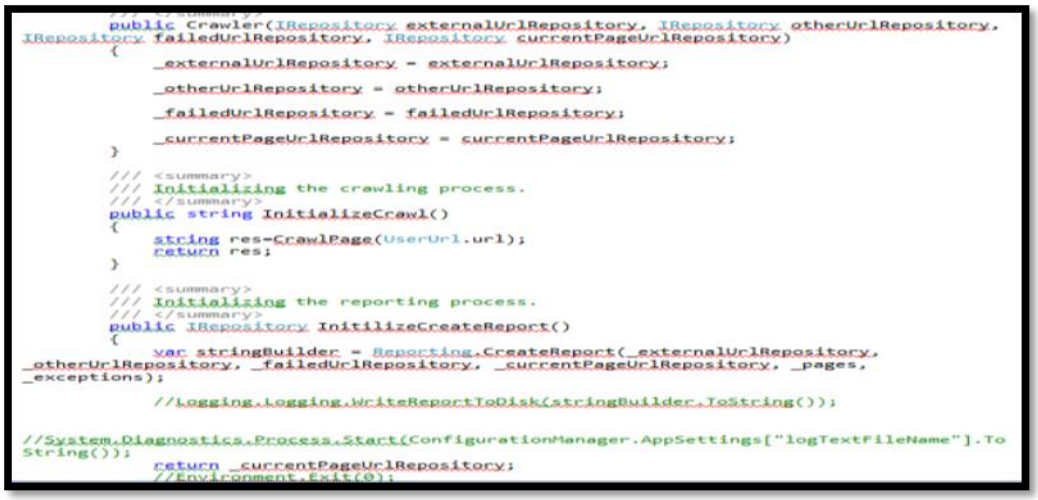
This section includes the requirements that specify all the fundamental actions of the Software system.

B. Document Collection:

This is first step of classification process in which we are collecting web content. In this phase, the dataset that we used as training and testing data were extracted over the internet. These data contain data about website categories available. The input for this step would be the huge dataset that we extracted from the internet.

C. Web Crawling:

A Web crawler starts with a list of URLs to visit, called the seeds. As the crawler visits these URLs, it identifies all the hyperlinks in the page and adds them to the list of URLs to visit, called the crawl frontier. URLs from the frontier are recursively visited according to a set of policies. If the crawler is performing archiving of websites it copies and saves the information as it goes. The archives are usually stored in such a way they can be viewed, read and navigated as they were on the live web, but are preserved as „snapshots'. The archive is known as the repository and is designed to store and manage the collection of web pages. The repository only stores HTML pages and these pages are stored as distinct files. The repository stores the most recent version of the web page retrieved by the crawler.



```
        public Crawler(IRepository externalUrlRepository, IRepository otherUrlRepository,
IRepository failedUrlRepository, IRepository currentPageUrlRepository)
        {
            _externalUrlRepository = externalUrlRepository;
            _otherUrlRepository = otherUrlRepository;
            _failedUrlRepository = failedUrlRepository;
            _currentPageUrlRepository = currentPageUrlRepository;
        }
        /// <summary>
        /// Initializing the crawling process.
        /// </summary>
        public string InitializeCrawl()
        {
            string res=CrawlPage(UserUrl.url);
            return res;
        }
        /// <summary>
        /// Initializing the reporting process.
        /// </summary>
        public IRepository InitilizeCreateReport()
        {
            var stringBuilder = Reporting.CreateReport(_externalUrlRepository,
_otherUrlRepository, _failedUrlRepository, _currentPageUrlRepository, _pages,
_exceptions);
            //Logging.Logging.WriteReportToDisk(stringBuilder.ToString());

//System.Diagnostics.Process.Start(ConfigurationManager.AppSettings["logTextFileName"].To
String());
            return _currentPageUrlRepository;
            //Environment.Exit(0);
```

Fig.3. Code Snippet of a Web Crawler

D. Pre-Processing:

The first step of pre-processing is used to present the text documents into clear word format. The documents prepared for next step are represented by a great amount of features. Commonly the steps taken are: Removing stop words: Stop words such as "the", "a", "and", etc are frequently occurring, so the insignificant words need to be removed.

E. Feature Selection:

After pre-processing the important step of text classification, is feature selection to construct vector space, which improves the scalability, efficiency and accuracy of a text classifier. The main idea of Feature Selection (FS) is to select subset of features from the original documents. FS is performed by keeping the words with highest score according to predetermined measure of the importance of the word.

F. Classification:

The automatic classification of documents into predefined categories has observed as an active attention, the documents can be classified by three ways, unsupervised, supervised and semi supervised methods.

G. Output:

This is the final step where after performing the above steps the URL or website will be classified into one of the categories or classes thus belonging to that group. This will suggest the user that the entered URL belongs to which category.

STEPS:

Step 1: Enter the URL.
Step 2: Click on Submit.
Step 3: We will get a list of URL"s.
Step 4: Click on Crawl.
Step 5: It will give the Categories for all URL"s.
Step 6: The most Matching category will be the output

V. RESULT

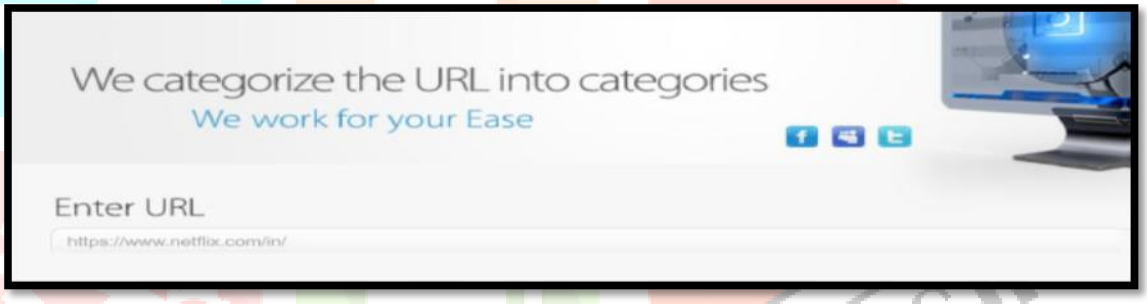


Fig.4 Input Value to Website categorization

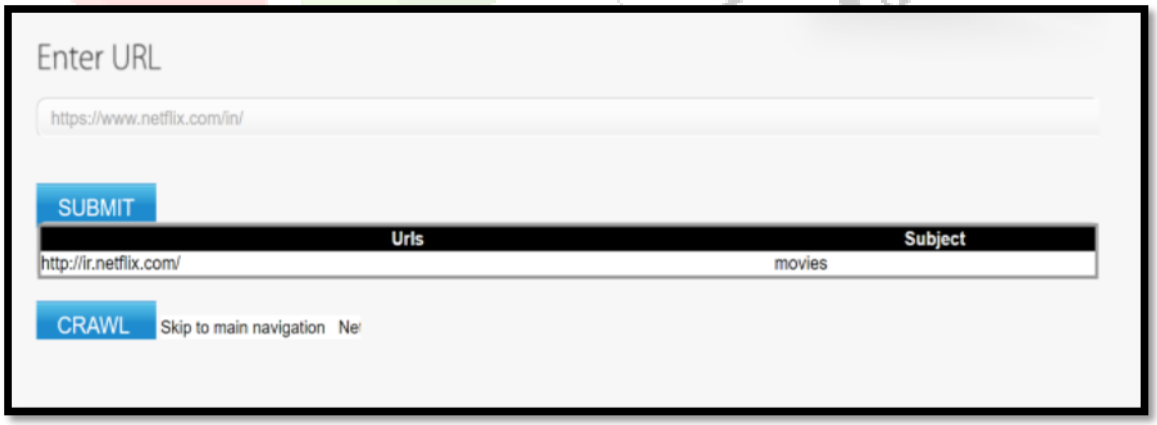


Fig.5 Result of Website categorization

VI. CONCLUSION

In this project a Website categorization System has been proposed. To improve the performance of classification TF-IDF classifiers are used. After reviewing Web classification research with respect to its features and algorithms, we conclude this by summarizing the lessons we have learnt from existing research and pointing out future opportunities in Web categorization.

VII. ACKNOWLEDGMENT

Authors would like to thanks to her guide Prof. Aarti Puthran, Assistant Professor of I.T department, SLRTCE, Thane for their valuable support and help.

VIII. REFERENCES

[1] Autonomous Website Categorization with Pre-Defined Dictionary -AdsadawutChanakitkarnchok, Kulit Na Nakorn, KultidaRojviboolchai Department of Computer Engineering Chulalongkorn University Bangkok, Thailand, adsadawut.ch@student.chula.ac.th, kulit.n@chula.ac.th, 2016

[2] Web Page Classification: Features and Algorithms ,XIAOGUANG QI & BRIAN D.DAVISON Lehigh University https://www.cs.ucf.edu/~dcm/Teaching/COT4810-Fall%202012/Literature/WebPageClassification.pdf

[3] Arul Prakash Asirvatham, Kranthi Kumar. Ravi, Web Page Categorization based on Document Structure, 2000.

[4] SoumenChakrabarti, Mining the Web, Morgan Kaufmann Publishers, 2003

[5] Web Crawling Algorithms, Aviral Nigam Computer Science and Engineering ,Department, National Institute of Technology - Calicut, Kozhikode, Kerala 673601, India, 2014

[6] A Literature Survey on Web Content Mining , International Journal on Recent and Innovation Trends in Computing and Communication ISSN: 2321-8169 Volume: 4 Issue: 10

[7] Web crawler https://en.wikipedia.org/wiki/Web_crawler

[8] Data Scraping https://en.wikipedia.org/wiki/Data_scraping

[9] TF-IDF https://en.wikipedia.org/wiki/Tf%E2%80%93idf