International Journal of Science and Research (IJSR)
ISSN (Online): 2319-7064
Impact Factor (2012): 3.358

An Offline Filtering Agent for Website Analysis and Content Rating: A Review

Vrushali Sanjay Kharad[1], Prof. S. S. Kulkarni[2]

[1]Master of Engineering, Department of Information Technology, Prof. Ram Meghe Institute of Technology & Research, Badnera-Amravati (444701), Maharashtra, India
[2]Associate Professor, Department of Information Technology, Prof. Ram Meghe Institute of Technology & Research, Badnera-Amravati (444701), Maharashtra, India

Abstract: ***In this 22nd centaury which is a world of vast Internet and Electronic media, information available on web does not necessarily to be correct or valid, Especially for children under eighteen. Many children uses internet daily for unlimited purpose and while browsing the web they may accidently encounter to such sites which includes articles, images or videos that are harmful to them. Content filtering is aimed at blocking out such undesirable material and other things reaching to end user. Most existing software content filters make use an access control list which involves some sort of manual search, gathering and classification of undesirable web sites so that the software filter can block the access of these URLs. In this paper we are describing offline filtering system in terms of its two main modules: First is automated web page crawling and second is intelligent classification module.***

Keywords: ACL, WebCrawler, WebParser, Self Organizing Feature Map (SOFM)

1. Introduction

Software's that uses content filter makes use of Access control list that involves looking out manually, collection and classify the undesirable and offensive websites in order that software package content filter will prohibit the access of those URL`s. however once we area unit talking concerning offline filtering system then some style of intelligence should be enclosed in it in order that there will be no like of manual search.

A.C.M. Fong, S.C. Hui and G.Y. Hong had worked and found out three major approaches adopted for analysis and filtering of offensive and objectionable online page. These embrace techniques supported keyword analysis, packet analysis and URL analysis. Among all the on top of accessible strategies we are using content analysis technique in Offline filtering system using some intelligence which will have a neural network. The scene behind the content analysis technique is that, once user enters any string or word in search engine then it is checked against the keywords within the database of the end user`s pc. information are going to be mechanically loaded to user`s pc once he install the package. Database can contain the maximum amount as potential words like porn and people words that are harmful to youngsters. once user open any web site whose one in all the word matches to the word within the database then site containing those words are going to be blocked. The advantage of content filtering is it adds little process overhead. An access control list (ACL), with reference to a file system, may be a list of permissions connected the object. An ACL specifies which users or system processes are granted access to objects, as well as what operations are allowed on given objects. Each entry in a typical ACL specifies a subject and an operation. It also refers to rules that are applied to port numbers or IP Addresses that are available on a host or other layer 3, each with a list of hosts and networks permitted to use the service. As mentioned above offline filtering agent comprises two major modules: Webparser and WebCrawler [1].

WebCrawler provides the automatic online page assortment mechanism, URL`s of websites are collected by querying three search engines which can come the list of URL`s of the net pages that matches the given keyword. A focused crawler is associate agent that targets a selected topic and visits and gathers solely relevant websites; It is enforced using Breadth-first search technique. A typical web crawler starts by parsing such internet page: noting any hyper text links on that page that point to alternative websites. The Crawler then parses those pages for new links, and so on, recursively. A crawler may be a code or script or automatic program that resides on one machine. The crawler merely sends communications protocol requests for documents to alternative machines on the web, even as an online browser will once the user clicks on links [1].

Once a page has been fetched, we want to dissect its content to extract data which will feed and probably guide the longer term path of the crawler. Parsing could imply easy hyperlink/URL extraction or it should involve the additional complicated method of tidying up the markup language content so as to investigate the markup language tag tree. The operating of this method is shown in fig 1.1 and it's clear that initial user can enter any uniform resource locator in parser he need to enter, then WebCrawler can do crawl, its job is that to search the pages so uniform resource locator can bear the WebParser to dissect the string and therefore the output of this parsing progressing to be a keyword who is going to be match the keyword that is within the database of end user`s system. Once keyword is matched then access control list can check the rights associate with that word and can take action consequently that's it'll block the web site. Parser can dissect string that may be a uniform resource locator word by word so it provides output a word. For the higher crawl and parsing there are numerous

techniques, software and algorithms to implement them [1].

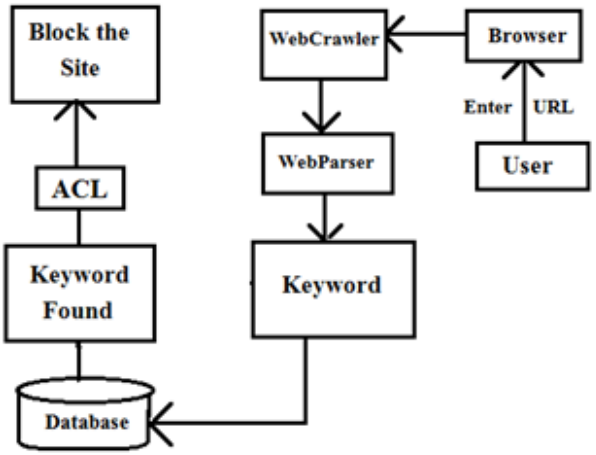

Figure 1.1: Flowchart of system

2. Literature Survey

Along with the ever-growing internet comes the proliferation of objectionable content, like sex, violence, racism, etc. we'd like economical tools for classifying and filtering undesirable online page. Mohamed Hammami, Youssef Chahir, and Liming Chen investigated this downside and represented WebGuard, an automatic machine learning-based pornographic web site classification and filtering system. Not like most commercial filtering product, that is mainly based on textual content-based analysis like indicative keywords detection or manually collected black list checking, WebGuard depends on many major data processing techniques related to matter, structural content primarily based analysis. Experiments conducted on a work of four hundred websites together with two hundred adult sites and two hundred nonpornographic ones showed internet Guard's filtering effectiveness, reaching a 97.4 percent classification accuracy rate once textual and structural content-based analysis was combined with visual content-based analysis. In providing a large assortment of hyperlinked transmission documents, the net has become a serious supply of knowledge in our way of life. Effective web site classification and filtering solutions are essential for preventing socio-cultural issues. as an example, as a number of the foremost prolific transmission content on the net, pornography is additionally thought-about as one of the foremost harmful, particularly for youngsters who every day have easier access to the net [2].

A large range of illegal and harmful content carried over the net may adversely have an effect on the national cultural security, the people's read on living and therefore the teenager's healthy growth, and should be supervised effectively. Researching and developing the net content rating implementation system is one amongst the possible means that to accomplish web content rating, observation and filtering. Firstly, Yan Liu, Gang Zhao and Tai Wang analyzed the most issues regarding the internet content rating technology. Secondly, they had a tendency to create a comprehensive analysis regarding the present web content rating implementation systems in step with their basic ideas, technical characteristics and shortcomings. Finally, they projected a brand new system style theme regarding the internet content rating. During this theme, they failed to solely use the present technologies, like the third-party rating, the digital signature, however additionally designed some new technical schemes, together with the intelligent analysis for the net content. They represented a whole, dynamic, self-adaptive and distributed web content rating implementation system that supervises the contents everywhere the net effectively [3]. At an equivalent time, individuals will acquire these contents a lot of and a lot of handily. On the one hand, this broadens the people's life and will increase the potency of their study and work. On the opposite hand, an outsized range of harmful content carried over the net may adversely have an effect on the national cultural security, the people's read on living and therefore the teenager's healthy growth. Thus it becomes a lot of pressing to supervise the net contents and construct the healthy network surroundings [4]. So as to supervise the net content effectively, some nations have established a series of laws and rules. Although it's useful and effectual for the net content management to promulgate these laws and rules, however we have a tendency to still need to use some technical means that to manage the net content in step with their ratings. Thus individuals will only read the particular web contents that suit them, and be separated effectively from the illegal and harmful contents [5]. The present different systems that carry out the net content rating are supported third-party rating, self-rating, or social network technical. As a result of these systems cannot resolve all the 5 key issues regarding the net content rating, they fail to be applied in practice. They proposed a brand new architecture design to accomplish the net content rating.

Vic Grout and John N. Davies projected that, among the various choices for implementing web packet filters within the style of Access control Lists (ACLs), is that the intuitive but likely crude methodology of process the ACL rules in successive order. Although such an approach finally ends up in variable method times for each packet matched against the ACL, it in addition offers the possibility to reduce this time by rearrangement its rules in response to dynamic traffic characteristics. form of heuristics exist for optimizing rule order in consecutive processed ACLs and also the most efficient of these is shown to possess a useful result throughout a majority of cases and for ACLs with relatively small numbers of rules. They presented an improvement to this formula by reducing a section of its complexity. Although the simplification involved finally ends up in an instant lack of accuracy, the long trade-off between method speed and performance is seen, through experimentation, to be positive. This improvement, though tiny, is consistent and worthy and could be observed among the majority of cases. Implementation as trees or tries: The conception of composing ACL rules as a searchable tree structure (binary or otherwise) is also a reasonably obvious one. However, in observe, rules square measure higher organized as tries. A trie (from 'retrieval') is really a tree with an array of pointers at each node, indicating subtries.

There is a pointer at each node for each attainable worth. The bits of each rule are thus keeping it up the branches of the trie, not the nodes. Rule look-up is performed a lot of faster on tries than trees. In each case the time taken to appear for the primary matching rule are (different) constant [6].

As the variety of net users and also the variety of accessible websites grows, it's becoming difficult for users to hunt out documents that are relevant to their express desires. Users ought to either flick through a large hierarchy of ideas to seek out the information that they are making an attempt or submit a question to a in public available program and struggle through several results, most of them are irrelevant. Web crawlers are one of the foremost crucial components in search engines and their improvement would have a good result on up the trying efficiency. P Dahiwale introduced an intelligent web crawler that uses ontological engineering ideas for up its crawling performance. Intelligent crawler estimates the best path for crawling. This can be often the first crawler that acts intelligently with none connection feedback or training. The requirement of an internet Crawler that downloads most relevant pages continues to be a significant challenge within the field of information Retrieval Systems. the use of link analysis algorithms like page rank and totally different Importance-metrics have shed a replacement approach in prioritizing the address queue for downloading higher relevant pages. The intelligent crawler has been projected. The Intelligent crawler formula estimates the linguistics content of the address supported the domain dependent philosophy that in turn strengthens the metric that is used for prioritizing the address queue. In addition, once downloading the page, the information path plays necessary role in estimating the connection of the links in that page. The projected new rule will solve the foremost necessary disadvantage of finding the connection of the pages before the method of crawling, to a best level. Today, we tend to expect of information superhighway as a typical hypertext system: a widely distributed collection of documents, connected by links occurring among the body of each document. Among the course of reading a hypertext document the reader can choose to follow the embedded links to totally different connected documents. We have got a tendency to use the Crawler programs to follow these links. Crawler, that may be a main element of a look engine, is also a program that retrieves websites or an internet cache [7]. An internet crawler is one kind of bot or software agent. In general, it starts with a listing of URLs to go to, known as the seeds. as a result of the crawler visits these URLs, it identifies all the hyperlinks among the page and adds them to the list of URLs to travel to i.e. URL Queue, referred to as the crawl frontier. URLs from the frontier are recursively visited according to a set of policies.

P. Baldi, P. Frasconi, and P. Smyth planned that, There are a number of crawling techniques used by net Crawlers, the important ones are highlighted here, that are: General Purpose Crawling: A general purpose net Crawler gathers as several pages because it will from a specific set of URL's and their links. In this, the crawler is in a position to fetch a large variety of pages from totally different locations. General purpose crawling will prevent the speed and network bandwidth because it is fetching all the pages. Focused Crawling: A focused crawler is meant to assemble documents solely on a particular topic, so reducing the number of network traffic and downloads. The goal of the focused crawler selectively seeks out pages that are relevant to a pre-defined set of topics [8]. It crawl solely the relevant regions of the online and ends up in vital savings in hardware and network resources. The foremost crucial analysis of focused crawling is to measure the harvest ratio which is rate at that relevant pages are acquired and irrelevant pages are effectively filtered off from the crawl. This harvest ratio should be high, otherwise the focused crawler would pay plenty of your time simply eliminating irrelevant pages, and it should be higher to use a standard crawler instead [9]. PyBot could be a net Crawler developed in Python to crawl the web using Breadth first Search (BFS). The success of the planet Wide net (WWW), that itself built on the open web, has modified the manner however human share and exchange info and ideas. With the explosive growth of the dynamic net, users ought to pay much time simply to retrieve a small portion of the knowledge from the online. The birth of the search engines have created human lives easier by merely taking them to the resources they need. Net crawler could be a program employed by search engines to retrieve info from the planet Wide net in an automated manner. This paper focuses on the essential ideas of an internet crawler and provides an algorithm for web crawling. To begin our analysis, we enforced a straightforward BFS based Crawler referred to as PyBot and crawled in our University web site and collected the data.

A.Guerriero and F. Ragni, C. Martines planned that, a web crawler could be a comparatively easy automatic program or script that methodically scans or "crawls" through web pages to retrieval info from knowledge. Various names for an online crawler include web spider, web robot, bot, crawler, and automatic indexer. There are many alternative uses for an internet crawler. Their primary purpose is to gather knowledge so once web surfers enter a search term on their web site, they can quickly give the surfer with relevant websites. During this work we propose the model of a low cost web crawler for distributed environments supported an efficient url assignment algorithm. The function of each module of the crawler is analyzed and main rules that crawlers should follow to keep up load balancing and robustness of system when they are looking on the online simultaneously, are discussed. They planned a dynamic address assignment method, based on grid computing technology and dynamic clustering, results efficient increasing web crawler performance. They have described the model and design of a distributed crawler system and also the basic rules that ought to be used to obtain a load balanced system. We given a replacement technique for address assignment during a parallel net crawler that with restricted resources can do a very efficient crawler [10].

Dr. Shruti Kohli and Sandeep Kaur planned that Content is most significant part on any web site. Keyword

primarily based reports, obtained from internet analytics tools, provide insight to content usability of web site. during this paper a tool 'Keyword similarly measure Tool'(KSMT) is presented which may be used to optimize the Keyword primarily based report by combining the similar keyword that have relevant which means and obtain a more in-depth and concise image. The aim is to boost the information accuracy and overcome limitation of comparable keywords being vastly separated within the report. This manner the methodology additionally provides holistic read of the information for similar keywords, by combining the matrices like bounce-rate, visits for the similar keywords and thus aim to supply a collective read and content analysis. The methodology additionally provides some way to match & analyze Keywords with 'Suggested Keyword' provided by the user. The KSMT tool is developed using Perl, Apache and flex. Keyword is important measuring of traffic metric visitors use to find you once they use search engines. Though internet analytics tool provides completely different reports supported different metrics, however these reports are very lengthy mainly because of the character of the information. We have a tendency to attempt to consolidate the reports by combining the similar keywords i.e. those have similar which means by using KSMT procedure. The combining method is based on the issue or association rule that what percentages of keywords are similar. We are presently considering 90.9% similarity in keywords to consolidate the reports. With the assistance of this consolidated report directors ought to improve the accessibility or readability so users can visit at affordable frequency and reside such pages for applicable length of your time. The continuous improvement in analytic tools is nice; however a tool on its own is not any answer. It's like saying a hammer will build you a home. Till you try that hammer with a master carpenter and a blueprint, you will only get a sore thumb. Similarly, you wish to match the increasing talents of your analytic tools with an analyst and measurable business goals before you get any worth out of it [11].

Pooja Gupta and Mrs. Kalpana Johari, said that, the world Wide internet is associate interlinked collection of billions of documents formatted using html. Ironically the very size of this collection has become associate obstacle for info retrieval. The user must shift through uncountable pages to return upon the knowledge he/she needs. Internet crawlers are the center of search engines. Internet crawlers unendingly persevere travel the net and notice any new websites that are additional to the net, pages that are removed from the net. Because of growing and dynamic nature of the internet; it's become a challenge to traverse all URLs within the web documents and to handle these URLs. A targeted crawler is associate agent that targets a specific topic and visits and gathers solely relevant websites. During this thesis they had worked on style and dealing of internet crawler which will be used for copyright infringement. They took one seed uniform resource locator as input and search with a keyword, the looking out result is supported keyword and it will fetch the net pages wherever it will notice that keyword [12]. This targeted primarily based crawler approach retrieves documents that contain explicit keyword from the user's query; we have a tendency to area unit implementing this using breadth-first search. Now, after we retrieved the net pages we are going to apply pattern recognition over text. They offered one file as input and apply the pattern recognition algorithms. Here, pattern symbolizes text solely and check what quantity text is offered on the net page. A crawler may be a program that downloads and stores websites [13], usually for a web search engine. The rapid growth of World Wide internet poses challenges to go looking for the foremost applicable link. Targeted crawler is developed to extract solely the relevant websites of interested topic from the net. Till now, Allan Heydon and marc Najork described "Mercator: A scalable, extensible web Crawler". Mercators's main support is for extensibility ant customizability. The crawler uses pattern recognition algorithms and generates the quantity of times the input text exists within the text found on a link. The crawler performs the pattern recognition using 3 rules independently and generates the quantity of performed by every algorithm. The knowledge therefore generated offers associate insight within the efficiency of the pattern-matching rule. The internet crawler designed is using only one technique of text mining i.e. pattern recognition. The webnet crawler will further be extended to use different text mining techniques. Thereby creating a crawler additional intelligent and higher equipped in finding copyright infringement [14]

References

[1] "An intelligent offline filtering agent for website analysis and content rating" A.C.M. Fong, S.C. Hui and G.Y. Hong
[2] "WebGuard: A Web Filtering Engine Combining Textual, Structural, and Visual Content-Based Analysis", Mohamed Hammami, Youssef Chahir, and Liming Chen, IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, VOL. 18, NO. 2, FEBRUARY 2006
[3] "Architecture Research and Design for the Internet Content Rating Implementation System", Yan Liu, Gang Zhao and Tai Wang, Second International Workshop on Education Technology and Computer Science, 2010
[4] "The network culture and the culture safety in Our country", Shuhong Xu, Senlin Zhang, Northeast Normal University master's degree paper, December 2006.
[5] "Going Where No Search Engine Has Gone Before Sarkar", Federal Computer Week, December 2005. http://www.fcw.com/article88982-05-30- 05-Print.
[6] "A Simplified Method for Optimizing Sequentially Processed Access Control Lists, An efficient process for reordering rules in traffic packet filers", Vic Grout and John N. Davies, Sixth Advanced International Conference on Telecommunications, 2010.
[7] "Networking Algorithmics: An Interdisciplinary Approach to Designing Fast Networking Devices", G. Varghese and Morgan Kaufmann, 2005.
[8] "PyBot: An Algorithm for Web Crawling", Rajendra Kumar Dash Department, IEEE 978-1-4577-2037-6/11/$26.00, 2011.

[9] "Modeling the Internet and the Web: Probabilistic Methods and Algorithms", P. Baldi, P. Frasconi, and P. Smyth, John Wiley & Sons Publishers, 2003.
[10] "A dynamic URL assignment method for parallel web crawler", A.Guerriero DEE, Politecnico of Bari Bari, Italy, IEEE 978-1-4244-7230-7/10/$26.00, 2010
[11] "A Website Content Analysis Approach Based on Keyword Similarity Analysis" Dr.Shruti Kohli and Sandeep Kaur, IEEE/WIC/ACM International Conferences on Web Intelligence and Intelligent Agent Technology, 2012
[12] "Implementation of web crawler" Pooja gupta and Mrs. Kalpana Johari, Second International Conference on Emerging Trends in Engineering and Technology, ICETET, 2009
[13] "Searching the Web", Arvind Arasu, Junghoo Cho, Hector Garcia-Molina, Andreas Paepcke. Sriram Raghavan, Conference (WWW) Computer Science Department, Stanford University, May 2004
[14] "Mercator: A Scalable, Extensible Web Crawler", Allen Heydon and Mark Najork, Compaq Systems Research Center, 130 Lytton Ave, Palo Alto, CA 94301, 2001