Literature Review

Monitoring and finding inappropriate contents on Web sites of Çankaya University

Neşe Türker

Department of Computer Engineering, Çankaya University, Ankara

Abstract

Web content is very important in organizations like Çankaya University. As a result, there is no single page. There are a lot of page owner. Checking of all pages for inappropriate contents is difficult, but monitoring them may be an option. For this reason, it is important to have a tool to monitor the contents and report them. This is the target of the proposed project. In this project, it is aimed that to find and report of inappropriate contents on the web sites of the Çankaya. Edu. tr extension by using apache nutch, web crawler, apache tika, apache lucene, vaadin tools. This project is being carried out together with Çankaya University Computer Center. As a result, the name of our school will be prevented from being mentioned with inappropriate words.

1 Introduction

To monitor the contents of publically accessible websites is very important to prevent organization from being mentioned with bad words. Therefore, publically accessible, illegal websites are checked by control agencies for criminal point of view [1]. Westlake et al. examine the child exploitation (CE) websites using a repeated measures design, over a period of 60 weeks [1]. In that study a custom-designed web-crawler was used for obtaining network data. It is demonstrated that increased volumes of CE code words and images are associated with premature failure.

Pajic et al. published a paper related to a software system called WebMonitoring [2]. In that paper, they designed a system in order to solve certain problems in the information search process on the web. To obtain further development they evaluate the WebMonitoring system and suggest directions. There is a problem about accessing to web pages. The reason for this problem is mainly limitations of search engine's crawling or time difference between the moment a web page is set up on the Internet and the moment the crawler. They achieve to solve a problem about accessing to web page content that is inaccessible by common search engines.

In the order to overcome problems in search process defined above the WebMonitoring software system is developed by Pajic et al. Java programming language was used to write this software. It

consists of several sub-systems. These sub-systems are (I) the system for making queries based on the Unitex software system, (ii) the management system, (iii) the crawler, (iv) the system for text post-processing, (v) the alarming system and (vi) the graphical user interface.

In recent years monitoring methods of Web pages contents come increasingly in the interest of governments [3]. By searching the constituents of Web pages may give symptoms of criminal activities. Some information is hidden by embedding in multimedia materials. Finding hidden information in indexing systems is a large challenge of contemporary criminal analysis. Turek et al. described a Web crawling system with a multimedia materials analysis algorithms [3]. The proposed The Web crawling system provides a mechanism for plugin inclusion by searching a few hundred pages per second. Processed resources can be analyzed by a plugin and detected references to multimedia materials. In that study, some implementation assumptions are presented as an example. It is also described that several approaches to the integration.

Navrat et al. proposed an approach and a system for tracing web pages by trying to find the most relevant pages with most recent news on it [4]. The proposed system is based on a bee hive model. It may be difficult due to changing information rapidly. The purpose of this study is to give a procedure in order to keep track of a developing story. It is not possible to download all the pages and revisit all pages to see whether new information was added or not. Therefore, they suggest to use a focused crawler for downloading the pages. They made a case study to demonstrate that the proposed system is suitable to collect related pages and to monitor story added.

In another paper, an Internet Public Opinion Monitoring System (IPOMS) is proposed [5] to collect web pages with some certain key words from various sources such as Internet news, topics on forum. With this system, the progress of one event can also be carried out the by using the function of automatically tracking. In this system web crawler, html parser and topic detection and tracking tools were used and the technologies of web page cleansing and k-d tree algorithm in topic tracking have been adopted. IPOMS is improved due to the existence of numerous data in web pages.

Nowadays, an automated tools of checking for correctness and consistency of data is essential. Typical Web sites publish more and more information from various data sources which are also updated very frequently. The eShopmonitor was proposed to tracked, monitors the items on the Web pages for users to specify items of interest [6]. This system also reports on any changes observed. In that system to achieve the above functionality, there are a crawler, a miner, a reporter, and a user component that work together. The items of interest are located on a class of pages by a miner. The user supply just one sample by using the user interface (UI). The learning algorithm is based on the XPaths of the Document Object Model (DOM) of the page.

Yuan et al. [7] proposed an efficient scheme to remove crawler traffic from the Internet. According to their findings approximately 40% of current Internet traffic is web crawler is estimated due to Web crawlers retrieving pages for indexing. This problem was solved by introducing an efficient indexing system based on active networks by employing active routers. These active routers monitor passing Internet traffic constantly, analyze it, and then transmit the index data to a dedicated back-end repository. It is demonstrated that the proposed simulations have shown that active indexing is up to 30% more efficient than the current crawler-based techniques.

2 Tools for monitoring Web pages

In order to monitor and to detect inappropriate contents by using such as apache nutch, web crawler, apache tika, apache lucene, vaadin tools.

2.1 Apache Tika

Apache Tika is a detection and analysis tool for web content. It is written in Java system but it is also widely used from other languages. It can detect data and text from different file types (such as PPT, XLS, and PDF) and then extract them. Tika can use for identification approximately above thousand file types from the Internet Assigned Numbers Authority taxonomy of MIME types. Content extraction, metadata extraction and language identification may be carried out by Tika. Several common file formats such as PowerPoint .ppt and Word .doc formats are actually held within a common container format OLE2 document. Another format is Apple iWork formats. They are actually a series of XML files within a Zip file.

2.2 Web Crawler

A Web crawler or only crawler (sometimes known a spider or spiderbot) is an Internet bot that systematically browses the World Wide Web for Web indexing. Web crawling software is used by Web search engines and some other sites in order to update their web contents. Web crawlers copy pages for processing by a search engine. Search engine indexes the downloaded pages to make easy and more efficient search for users.

2.3 Apache Lucene

Apache Lucene is a free tool for open-source information taking back software library. It is written completely in Java by Doug Cutting. However, it is supported by the Apache Software Foundation. It is released under the Apache Software License. It is suitable applications that requires full text indexing and searching. Lucene has been widely used in the implementation of Internet search engines and local, single-site searching. It has also been used to implement recommendation systems such as

'MoreLikeThis' Class. It can make recommendations for similar documents. The similarity approach of 'MoreLikeThis' measures Co-citation and Co-citation Proximity Analysis. Lucene's API is independent of the file format. All file formats such as text from PDFs, HTML, Microsoft Word, Mind Maps, OpenDocument documents and many others can be indexed and their textual information can be extracted. However images cannot be indexed by this system.

2.4 Apache Nutch

Apache Nutch is another highly extensible and scalable open source web crawler software project. Nutch is also coded entirely in the Java programming language like Apache Tika. However, the data has been written in language-independent formats. It Apache Nutch has a modular architecture. These property of Apache Nutch allows developers to create plug-ins for media-type parsing, data retrieval, querying and clustering. Nutch has many advantages. It is highly scalable and relatively feature rich crawler, polite which obeys robots.txt rules, robust and scalable and quality.

2.5 Vaadin

Vaadin is developed as an open-source platform for web applications. A set of web components, a Java web framework, and a set of tools and application starters present on Vaadin platform. Implementation of HTML5 web user interfaces is carried out by its flagship product, Vaadin Flow by using the Java Programming Language. Components of Vaadin are developed that can be used in web documents (without frameworks) and web frameworks compatible with Web Components. These components are a comprehensive set of Web Components for application developers and the core of Vaadin Flow.

3 Conclusions

To monitor, detect and report the inappropriate contents on the web sites of the Çankaya. edu. tr extension by using various tools such as apache nutch, web crawler, apache tika, apache lucene, vaadin. In this paper these tools are discussed. To remove of inappropriate contents is also one of the aim for the project. However, this goal is excluded for this project due to limited time of project. There are various tools to detect and to track contents of web sites introduced in this paper. The tools are mentioned in this paper with their advantages and disadvantages. Furthermore, the application areas with some studies are given as examples.

4 References

[1] Westlake, B.G., Bouchard, M., Criminal Careers in Cyberspace: Examining Website Failure within Child Exploitation Networks, Justice Quarterly, 33(7, 1154-1181, 2016. DOI: 10.1080/07418825.2015.1046393

[2] Pajic, V., Vitas, D., Lazetic, G.P., Pajic, M., WebMonitoring Software System: Finite State Machines for Monitoring the Web, Computer Science and Information Systems, 10(1), 1-23, 2013. DOI: 10.2298/CSIS110918036P

[3] Turek, W.,Opalinski, A., Kisiel-Dorohinicki, M., Extensible Web Crawler - Towards Multimedia Material Analysis, Multimedia Communications, Services, and Security, Edited by:Dziech, A; Czyzewski, A., Book Series: Communications in Computer and Information Science, 149, 183-190, 2011.

[4] Navrat, P., Jastrzembska, L., Jelinek, T., Bee Hive At Work: Story Tracking Case Study, 2009 IEEE/WIC/ACM INTERNATIONAL JOINT CONFERENCES ON WEB INTELLIGENCE (WI) AND INTELLIGENT AGENT TECHNOLOGIES (IAT), VOL 3, Edited by:BaezaYates, R; Berendt, B; Bertino, E; Lim, EP; Pasi, G., 117-120, 2009.

[5] Ding, J., Xu, J.G., POMS: an Internet Public Opinion Monitoring System, 2009 SECOND INTERNATIONAL CONFERENCE ON THE APPLICATIONS OF DIGITAL INFORMATION AND WEB TECHNOLOGIES (ICADIWT 2009), 433-437, 2009.

[6] Agrawal, N., Ananthanarayanan, R., Gupta, R., Joshi, S., Krishnapuram, R., Negi, S., The eShopmonitor: A comprehensive data extraction tool for monitoring Web sites, IBM JOURNAL OF RESEARCH AND DEVELOPMENT, 48(5-6), 679-692, 2004. DOI: 10.1147/rd.485.0679

[7] Yuan, X., MacGregor, M.H., Harms, J., An efficient scheme to remove crawler traffic from the Internet, ELEVENTH INTERNATIONAL CONFERENCE ON COMPUTER COMMUNICATIONS AND NETWORKS, PROCEEDINGS, Edited by:Luijten, R; Wong, E; Makki, K; Park, EK, 90-95, 2002. DOI: 10.1109/ICCCN.2002.1043051