Software Requirements Specification

Monitoring and finding inappropriate contents on Web sites of Çankaya University

Neşe Türker

Department of Computer Engineering, Çankaya University, Ankara

Table of Content

1 Introduction
1.1 Purpose
1.2 Product Scope
1.3 Glossary
2 Overall Description
2.1 Product Perspective
2.2 Development Methodology
2.3 User Characteristic
2.4 Product Function
3 Structure of Web Monitoring System
4 Example for Application
5 Conclusion
6 References

1 Introduction

1.1 Purpose

The purpose of this project is to monitor, find and report of inappropriate contents on the web sites of the cankaya. edu. tr extension by using apache nutch, web crawler, apache tika, apache lucene, vaadin tools.

1.2 Product Scope

As Internet usage is growing daily it has also managed to create problems for people because of the increase in cyber-crimes and inappropriate contents. Therefore, there is a need for monitoring and analyzing web sites for detecting inappropriate words and blocking the malware. Using Signature-based, Statistical anomaly-based and Honeypot-based detection methods, the malware is detected and then further actions are taken to block them. This project scopes to detect, to monitor and to report the inappropriate contents on web pages with cankaya.edu.tr extension.

1.3 Glossary

Term	**Definition**
Inappropriate content	Illegal word, sentence or picture
Crawler program	a program that systematically browses the World Wide Web in order to create an index of data
Search engine crawlers	A tool crawl web pages periodically for updated contents

2 Overall Description

2.1 Product Perspective

The monitoring and tracking system can be thought of as a program that basically detects inappropriate contents on the WEB pages of Çankaya University. However, it is now known that many tools utilize for detecting, monitoring and tracking web pages. This and similar systems are used in many areas such as tracking, detection and prevention of website attacks. This idea has been tried to be used in Çankaya University.

2.2. Development Methodology

In this study an iterative development model is used to develop our system. Software was developed by applying six steps such as planning, requirements, analysis design, implementation, test and evaluation. First four steps create a loop. This loop continues until the software has reached the required threshold. After completed these four steps, the test and evaluation steps were carried out [1]. Thus, if a problem is encountered at evaluation step, first four steps are overviewed again. In order to monitor inappropriate contents of web pages the tools such as apache nutch, web crawler, apache tika, apache lucene, vaadin were tested.

2.3 User Characteristic

Administrator, manager and officers who be employee at Çankaya University Computer Center. Manager and officers must have knowledge of software. Teachers who are lecturer at Computer Engineering, Çankaya University may also be user.

2.4 Product Functions

The developed web monitoring software system may overcome problems in search process. Java programming language was used for this software. This software system consists of several sub-systems: (1) querying system (2) the management system (3) the crawler (4) the system for text post-processing (5) the alarming system and (6) the graphical user interface.

The management system, the crawler, the alarming system and the graphical user interface were developed for the purpose of the web monitoring system. The Unitex software and some of its components was used for querying and post-processing the text. The findings and pages were analyzed and processed to find patterns corresponding to the graph. When a pattern is found (and event occurred), the system notifies the user about this result.

3. Structure of Web Monitoring System

The Management System. The management is the central point of the overall system. Web monitoring software consists of several Java classes. More than one independent monitoring process may be run by users. Every monitoring process consists of the followings:
(1) URL (a web page URL from which the crawl and the search start);
(2) graph (a location of .fst2 file which describes the searched phrases);
(3) levels (an integer that defines the depth of crawl);
(4) alarm (a string attribute that defines the way a user should be alarmed if the event occurred).

When the application is started, the main window of the management system will open. A user can select some process from the table to view or change its characteristics. From this window the user can also delete, make a new, start or stop a process. After the user has started a monitoring process, the crawler starts crawling process that depends on the value of the parameter. If the value of the parameter is 1, the system should take only this shown page. Otherwise, when one page is downloaded from the Internet, it is analyzed to find other hyperlinks on it. The crawl process is stopped when all the pages with the level greater than 0 have been processed. After adjusting the parameters, the crawler for the given URL is started. The text found on every web page in this process is sent to the post- processing system for further analysis, i.e. for the graph search.

When a web page is downloaded from the Internet, first it is necessary to prepare the text it contains, and then to search for the appropriate event defined by the search graph. This task is performed by the system for post-processing. Since the Unitex's external program Locate has the central place in the search process, the text from the page is prepared in accordance with the requirements of this program. The text found on the web page is saved in a text file and stored in a temporary directory. This file is the starting file in the text processing. Although text on web pages is coded differently, most websites nowadays use UTF-8 encoding. UTF-8 (8 bit Unicode Transformation Format) encodes each Unicode character as a variable number of 1 to 4 octets, where the number of octets depends on the integer value assigned to the Unicode character. Since each character in the range of U+0000 through U+007F is represented as a single octet, UTF-8 is a very efficient encoding schema of text documents in which most characters are US-ASCII. This is also the reason why this encoding became dominant for electronic mail and web documents, and therefore web monitoring system

4 Example for Application

The web monitoring software system can be used for different tasks, such as press clipping, detecting spam messages by monitoring electronic mailboxes, management of various documents collections, and so on. A user wishes to find all inappropriate articles, which are placed on Web pages of Çankaya University.

Step 1. Defining the event to be searched for:

The event to be searched for is an occurrence of inappropriate contents. One of the contents (in Turkish language) is: "küfür".

Step 2. Describing the event by a graph:

The user uses Unitex and creates a graph that describes the defined event. This task can be performed in many different ways, and it depends on user's skills and available resources.

Step 3. Choosing the content to monitor:

The user chooses web pages or web sites he/she wishes to monitor. Having in mind that the user wishes to find news articles, he/she chooses official web sites of Çankaya University (http://www.cankaya.edu.tr, http://bim.cankaya.edu.tr/) as starting points of the monitoring process.

Step 4. Creating monitoring processes. In the web monitoring system, the user creates a process for each web site he/she wishes to monitor. For each process the user sets URL, number of levels for the crawl, location of the graph describing the event, the way of alarming, and the interval for repeating the process.

Step 5. Starting the Processes. The user selects the process in the table showing processes and starts it by choosing the appropriate button. The crawling and monitoring process starts and works in the background. The user can see the progress by choosing the button ("Report"). If the phrase that matches the graph is found on some page, the user is alerted either by e-mail or the page is saved locally on the user's computer.

5 Conclusions

In this project a web monitoring system has been developed to improve search process in terms of making more complex queries and access to content of web pages in a short period after their posting on the web. As a solution for complex querying, it is proposed that to use finite state machines. The software system Unitex was used for making queries and for the post- processing of the text. A new web monitoring system was designed and developed. It has integrated a subsystem for crawling web pages. By using this system, users can do their own crawl and search web pages. They do not need to use the common search engines. Furthermore, by applying this system the user creates, maintains and controls the processes of monitoring web pages or sites. The system also allows for looking up for some event such as a phrase occurrence on a page. The web monitoring software system may be used for specific types of text or for special purposes, such as monitoring electronic mailboxes, or search for a specific product in a database accessible from the Internet.

6 References

[1] 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