Software Design Document

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

The purpose of this Software Design Document is providing the details of project entitled as "Monitoring and finding inappropriate contents on Web sites of Çankaya University".

The purpose of this project is to monitor, find and report of inappropriate contents on the web sites which have "cankaya. edu. Tr" extension by using various tools such as apache nutch, web crawler, apache tika, apache lucene, vaadin tools.

For this purpose, it could be considered to use general-purpose search engines. However, they do not allow making some queries with morphologically rich languages. In recent years, Google has made significant efforts to improve its search process and to solve above mentioned problems. This is an important development for such searches, but there is still a big problem. This big problem is that users have no control over this process.

In this project, it is aimed that to focus on solving the above-mentioned problems and to monitor inappropriate contents of web pages on the web sites of Çankaya University. For this purpose, one of the target of this project is to improve formulating queries. More complex context of information will be described by this formulation. By using this tool, the access to the content on the Web will be possible in the shortest time interval. For formulating search queries, it is proposed to apply finite state

automata. Users will be able to describe very complex contexts, illegal words and phrases they want to find on web pages by this finite state automata.

The proposed software system with its own crawling sub-system will work as an application on a user's computer. By using this sub-system a user able to set a seed URL and a level of crawling. The graphs representing finite state automata or transducers will be used to describe information which to be wanted to find by the users. Those graphs will be used as a query for the search. When some information occurs on the monitored web page, users will be able to warn during the interval of repeated checks.

1.2 Product Scope

Nowadays internet usage is growing continuously. As a result of this growing, problems for people increase because of the cyber-crimes, malware and inappropriate contents. Therefore, monitoring and analyzing web sites for detecting inappropriate words and blocking the malware is very important. The malware can be detected using Signature-based, Statistical anomaly-based and Honeypot-based methods. If it is detected they can be blocked by further actions. 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

1.4 Motivation

As a senior student in computer engineering department, I am interested in developing applications web monitoring system to detect and report illegal and inappropriate contents of web pages on web sites of Çankaya University. Moreover, I always think over the software solutions about the problems that are occurred in our environment. So, I intended to assemble our both practical and theoretical experience in this project. To be able to track illegal contents of web sites, I developed a web monitoring software together with Computer Center of Çankaya University.

2 Monitoring the Web Sites

Recent years different tools for searching and monitoring the web sites have been developed. These tools improve the process of searching for information on the WWW. Two different tool groups developed for this purpose depending on their architecture, functionality, and the problems they focus on [1]. The first group tools focus on monitoring the web. They are all designed only to notify if a web page has been changed. Queries based on keywords can be searched but complex queries cannot be searched by these tools. A user can set the downloads frequency. Some of the tools such as online systems ChangeDetect3 and WebSite Watcher4 use regular expressions [1]. The ChangeDetect system is an online tool for monitoring web pages. Several parameters such as frequency of downloads, regular expression filtering, events causing alerting the user can be set by the users for this system. Changes occurring on the page in less than 12 hours may not be noticed by the system.

WebSite-Watcher software is designed to track changes on any number of web pages [1]. With this tool it is possible to monitor all formats of electronic text. Even password-protected pages can also be monitored by tis tool. However, this tool also does not support setting up complex queries. WebSite-Watcher also uses regular expressions. A user may search for the occurrence of a keyword or a phrase. The second group tools focus on making queries for the search. They are often linguistically oriented tools. One of the best known second group tools is WebCorp system [2]. WebCorp is designed by the Department of Research and Development of English Language at the University of Birmingham. A user can search for a specific word, a phrase, or a pattern using this tool. By applying this tool complex phrases or patterns can be described. However, much information on the web still cannot be accessible by this tool. A free online concordance service, GlossaNet [3], crawler regularly visits sources and greatly improves the process of search.

3 Web Monitoring Software System

Finite state automata and transducers are used in many fields of computational linguistics. These are adequate for describing relevant local phenomena in language research. Modeling of natural language may be carried out by these automata and transducers. Modelling of phonology, morphology, or syntax of natural language may be possible from the linguistics point of view. As a computer science the use of finite state machines have higher time and space efficiency. They need big graphs. Space efficiency is achieved by minimizing deterministic machines. So, instead of one big graph, sub graphs collections are used. There are several computer tools for linguistic research. Detailed review of theoretical and practical use of finite state transducers in natural language processing is given in various literatures [4], [5].

3.1 The Concept of Monitoring the Web

The concept of web monitoring is to be alerted to the users if there are changes on a particular web page or site. In practice, when a user visits a web site he/she expects that some event occurred on it,

without being interested in the rest of the content of the web site. Examples of such events are announcement of the results of some examination, electronic message with the specific content, message from a specific person, the news about a particular topic, and so on. In such cases, the user regularly visits the web site of interest, looking for the event at an optimal or possible time. This searching process is carried out automatically by the software for web monitoring. The software simulates the actions that the human would take.

3.2 The System for Making Queries

The WebMonitoring system developed by Pajic et al. [1] uses graphs produced by the Unitex software system [6] as search queries. Unitex is a collection of programs. It is developed for analyzing text written in natural languages. It is also applied to different linguistic resources and tools to the text. It is an open source software having graphical interface. This graphical interface is user-friendly and has very good functionality. The main two advantages of the Unitex software are (1) its well-designed graphical user interface for creating graphs and (2) the possibility to use linguistic resources, such as electronic dictionaries and grammars. Electronic dictionaries contain simple and compound words. Their lemmas and the set of grammatical code is also found in electronic dictionary. Electronic dictionaries in DELA format is used by Unitex software. In DELA format each entry is a line of text terminated by a new line. It conforms to the following syntax:

apples,apple.N+conc:p

“apples” is the first word which is an inflected form of the entry and it is mandatory. Second word is standard form (lemma) of the entry. This information may be removed if the canonical form is the same as the inflected form. The following sequence of codes (N+conc) gives the grammatical and semantic information about the entry. In this example, code N demonstrates for noun, and “conc” indicates that this noun designates a concrete object. The letter “p” indicates “plural” form. Unitex creates separate files with simple words, compound words, and unrecognized words after applying dictionaries and grammars to the text. Those files are used in the search process. For example, a user can use the query <be.V> that matches all entries having be as standard form and the grammatical code V. Thus all occurrences of the verb to be (am, is, being etc.) will be recognized by this query. A user can use lexical masks and/or morphological filters. For example, the filter <<ism$>> matches all words that end with “ism” (capitalism, racism, etc.). Users can make graphs that correspond to very complex queries by applying lexical resources to the text. He/she can combine lexical masks and morphological filters for this purpose. Those graphs in Unitex software system may have two formats, (1) for the design phase of graphs the format .grf, and (2) for further processing and applying to a text the format .fst2.

The content of the corresponding .fst2 file, which is used as a search query, is as follows:

```
0000000003
-1 orlovi :
-2 1 -3 1 1 1
t
f
-2 orloviPre
: 9 2 8 1 6 1
: 5 3 4 3 3 3 2 3
: 7 4 5 3 4 3 3 3 2 3
t
: 5 3 4 3
f
-3 orloviPost
: 5 1 4 1 3 1 2 1
: 11 3 10 2
t
: 12 2
f
%<E>
%<<^orlov>>
%<<^reprezenta>>
%<<^fudbaler>>
%<<^tim>>
%<<^ekip>>
%<<^srpsk>>
%<<^nacionaln>>
%<<^Antićev>>
%<<^naš>>
%<<^Srbij>>
%<<^Radomir>>
 %<<^Antić>>
f
```

The .fst2 format is strictly defined by the Unitex software. The first line represents the number of graphs that are encoded in the file. Lines containing the number and the name of the graph identify the beginning of each sub-graph. In the above file, the sub-graphs are the lines -1 orlovi, -2 orloviPre and -3 orloviPost. The following lines describe the states of the graph. If the state is final, the line starts with t character. For each state, the list of transitions is a sequence of pairs of integers. The first integer indicates the number of the label or sub-graph that corresponds to the transition. Labels are numbered starting from 0. Sub-graphs are represented by negative integers. The second integer represents the number of the result state after the transition. In each graph the states are numbered starting with 0. By convention, state 0 is the initial state. The Unitex system and its interface for creating graphs represent a system for making queries which describe an event and a user wants to be notified of. Using the Unitex software system, a user creates a graph that describes the event of interest. Then a

user compiles the graph into the .fst2 format. This file contains all the necessary information about the event of interest.

4 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

[2] Kehoe, A., Renouf, A., WebCorp: Applying the Web to Linguistics and Linguistics to the Web, WWW2002 Conference, Honolulu, Hawaii (2002).

[3] Fairon, C., GlossaNet: Parsing a web site as a corpus, Lingvisticae Investigationes, John Benjamins Publishing Company, Volume 22, Number 2, pp. 327-340(14) (2000)

[4] Casacuberta, F. Vidal, E. Picó, D. Inference of finite-state transducers from regular languages, Pattern Recognition, Volume 38, Issue 9, pp.1431-1443 (2005)

[5] Friburger, N. Maurel, D. Finite-state transducer cascades to extract named entities in texts, Theoretical Computer Science 313, pp 93 – 104 (2004)

[6] Paumier S., Unitex 1.2 User Manual, Université de Marne-la-Vallée. http://www- igm.univ-mlv.fr/˜unitex/UnitexManual.pdf (2006)