World Academy of Science, Engineering and Technology
International Journal of Computer and Information Engineering
Vol:1, No:7, 2007

# Specialized Web Robot for Objectionable Web Content Classification

SuGil Choi, SeungWan Han, Chi-Yoon Jeong, and TaekYong Nam

International Science Index, Computer and Information Engineering Vol:1, No:7, 2007 waset.org/Publication/411

***Abstract*****—**This paper proposes a specialized Web robot to automatically collect objectionable Web contents for use in an objectionable Web content classification system, which creates the URL database of objectionable Web contents. It aims at shortening the update period of the DB, increasing the number of URLs in the DB, and enhancing the accuracy of the information in the DB.

***Keywords*****—**Web robot, objectionable Web content classification, URL database, URL rating

## I. Introduction

INTERNET users are easily exposed to hate literature, pornography, pedophiles, and other inappropriate information. This is becoming serious social issue and parents are much concerned about the access to this kind of information by children. Now, we see many parents installing objectionable Web content filtering software in PC.

Objectionable Web content filtering solution can be built on URL database method or dynamic method, or both. Most of the filtering solutions block connection to URLs which are in URL database. This is known to be more effective and efficient. But, the generation and maintenance of the URL database by human is difficult and slow. As the result, we need a system for the generation and maintenance of the URL database by automatically collecting and classifying Web content.

The collection of Web content in the system is the role of Web robot. However, it is inappropriate to employ a general-purpose Web robot for objectionable Web content classification, because general-purpose Web crawling quickly loses its way and starts to gather harmless pages, even though it starts from the URLs of porn sites. As the gathering process goes to wrong direction, objectionable Web content classification takes longer and the URL database refreshes much less frequently, which mean it is not likely to filter out new objectionable Web contents. In addition, general-purpose Web robot does not feed any information to the classification module except the downloaded pages, so the classification module can miss important hints for deciding if the Web content is harmful or not.

In this paper, we describe the Web robot specially designed for objectionable Web content classification. The design principles are mainly three: acquiring relevant Web contents, achieving respectable coverage of objectionable Web contents, and supplying helpful information to the classification module. We designed a Web robot which embodies above three principles. We are sure that this helps keep the URL database more up-to-date, achieve respectable coverage of harmful URLs, and make the DB more accurate. The proposed robot is just right fit in the objectionable Web content classification system.

The rest of the paper is organized as follows. In Section 2, we provide some background information on Web robot and objectionable Web content classification system. The problem of employing a general-purpose Web robot for objectionable Web content classification is given in Section 3. The proposed Web robot is presented in Section 4. Conclusion is in Section 5.

## II. Preliminaries

### A. Web Robot

A Web robot is a program that automatically traverses the Web's hypertext structure by retrieving a document, and recursively retrieving all documents that are referenced. Note that "recursive" here doesn't limit the definition to any specific traversal algorithm; even if a robot applies some heuristic to the selection and order of documents to visit and spaces out requests over a long space of time, it is still a robot. Normal Web browsers are not robots, because they are operated by a human, and don't automatically retrieve referenced documents (other than inline images). Web robots are sometimes referred to as Web wanderers, Web robots, or spiders. These names are a bit misleading as they give the impression the software itself moves between sites like a virus; this not the case, a robot simply visits sites by requesting documents from them.

Manuscript received July 26, 2005.

SuGil Choi is with the Electronics and Telecommunication Research Institute, Daejeon, Korea (phone: 82-42-860-1367; fax: 82-42-860-5611; e-mail: sooguri@etri.re.kr).

SeungWan Han is with the Electronics and Telecommunication Research Institute, Daejeon, Korea (phone: 82-42-860-4942; fax: 82-42-860-5611; e-mail: hansw@etri.re.kr).

Chi-Yoon Jeong is with the Electronics and Telecommunication Research Institute, Daejeon, Korea (phone: 82-42-860-4937; fax: 82-42-860-5611; e-mail: iamready@etri.re.kr).

TaekYong Nam is with the Electronics and Telecommunication Research Institute, Daejeon, Korea (phone: 82-42-860-6781; fax: 82-42-860-5611; e-mail: tynam@etri.re.kr).

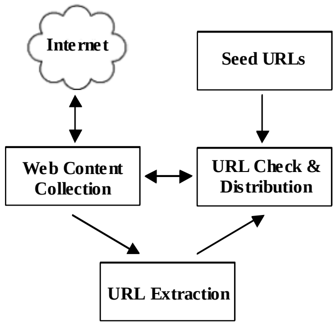

Fig. 1 Operation of the General-purpose Web Robot

A robot begins crawling from Seed URLs which are manually generated for certain purpose. For example, if a robot is meant to collect travel information, the Seed URLs may contain the URLs of travel portals. Only the Seed URLs which passed URL check procedure are actually used. URL check procedure sees if the URL is reachable, it hasn't been collected, and etc. The Web content corresponding to the Seed URLs are downloaded and saved in the local storage. URLs linked from Web pages downloaded are extracted and new URLs out of those URLs are used for continuing crawling process [1].

### B. *Objectionable Web Content Classification System*

Objectionable Web content classification system creates the URL database of objectionable Web contents. This database is critical part of objectionable Web content filtering solution. Objectionable Web content filtering solution can be built on URL database method or dynamic method, or both. The URL database method is based on white and black lists of classified URLs generated and periodically maintained by specialized employees. For example, when the URL of www.persiankitty.com is in the URL database, the connection attempt to the URL is blocked. Dynamic methods are based on algorithms that analyze texts or images on the fly. Many algorithms for analyzing Web contents on the fly were proposed, but nothing guarantees respectable accuracy and efficiency. Most of the filtering solutions are employing URL database method, which is known to be more effective and efficient. But, new objectionable Web sites are appearing every day, the URLs of the existing Web page are changed, which make the generation and maintenance of the URL database by human difficult and slow. As the result, an automated system for the generation of the URL database was proposed, which is called objectionable Web content classification system. This system consists of mainly two parts: Web content collection module and Web content classification module. The collection of Web contents in the system is the role of Web robot. [2].

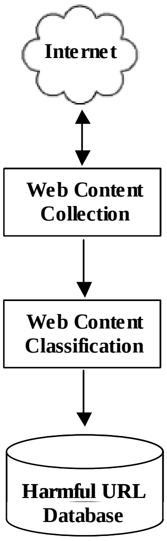

Fig. 2 Objectionable Web Content Classification System

## III. Problem Overview

Many Web robots are now working for search engines, e.g. yahoo, google, and etc., and this kind of Web robot is general-purpose. However, it is inappropriate to employ a general-purpose Web robot for objectionable Web content classification, because general-purpose Web crawling quickly loses its way and starts to gather harmless pages, even though it begins crawling from the URLs of porn sites. For example, as shown in Fig 3, a porn site has a link to www.google.com and www.sologirls.com. Once a Web robot moves to google.com and starts to gather Web contents from it, the robot will spend most of time crawling the regions of harmless Web contents.

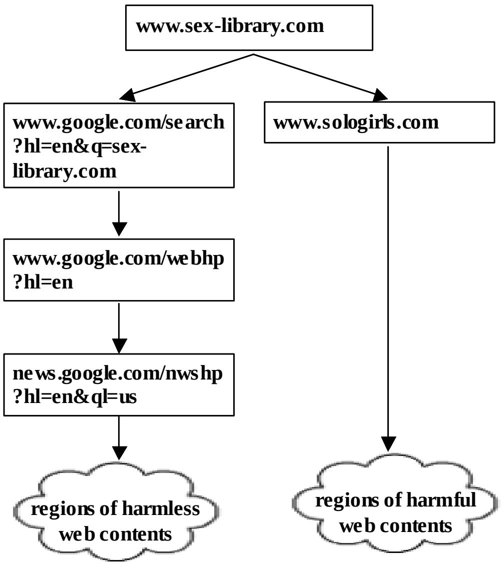
www.sex-library.com
www.google.com/search ?hl=en&q=sex-library.com
www.sologirls.com
www.google.com/webhp ?hl=en
news.google.com/nwshp ?hl=en&ql=us
regions of harmless web contents
regions of harmful web contents

Fig. 3 Example of Web Robot Crawling

As a robot is not intelligent at all, it collects all the Web contents, both text and image, which lead to slow crawling and huge storage consumption. A robot can't frequently refresh and further explore relevant regions of the Web. As the crawling goes slow, objectionable Web content classification takes longer and the URL database refreshes much less frequently, which mean it is not likely to filter out new objectionable Web contents. In addition, general-purpose Web robot does not feed any information to the classification module except the downloaded pages, so the classifier can miss important hints for deciding if the Web content is harmful or not. In short, the reason why general-purposed Web robot is not appropriate is as follows:

- collection of irrelevant Web contents
- very slow crawling
- huge storage consumption
- bypassing important information

Focused crawling methods have appeared on literatures in the past few years, but the proposed techniques are insufficient to give expected result, as they are not targeted especially on objectionable Web content collection [3] [4]. Some focused Web robots, such as porn info seeker [5], were developed. The porn info seeker searches out porn sites on the World Wide Web, starting from some of the major entry points for porn sites, then retrieves them and stores them. But, the porn info seeker determines the relevance only by the presence of dirty keywords in Web pages. This approach can give just partial improvement. We need more advanced methods.

## IV. Proposed Web Robot for Objectionable Web Content Classification

In this paper, we propose a Web robot specially designed for objectionable Web content classification. The design principles are mainly three: acquiring relevant Web contents, achieving respectable coverage of objectionable Web contents, and supplying helpful information to the classification module. These principles are explained and the corresponding building blocks embodying these principles are shown. Fig 4 depicts the overall operation of the proposed Web robot. Harmful URL MetaSearch block, Non-harmful Image Filter block, Monitoring in Web Content Collection & Monitoring block, and Harmless URL Check in URL Extraction & Harmless URL Check block differentiate the proposed Web robot from general-purpose Web robot.

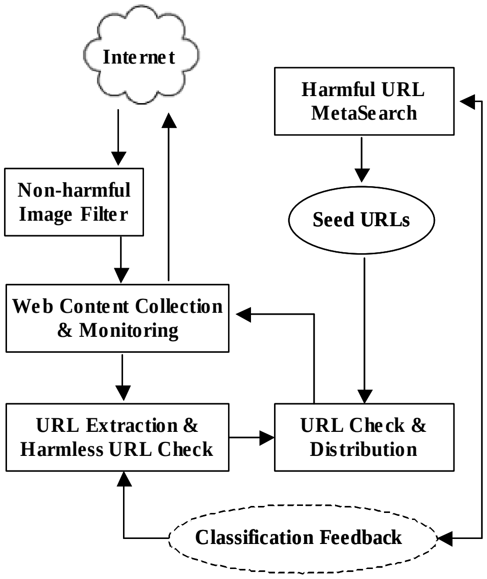

Fig. 4 Architecture of the Proposed Web Robot

**Acquiring relevant Web contents:** It is with the intention of refreshing the URL database more frequently. This can be achieved by raising the ratio of objectionable Web contents to the all contents downloaded by the robot. URL Extraction & Harmless URL Check block and Non-harmful Image Filter block in Fig 4 are employed to achieve this goal.

The proposed robot collects Web contents that are likely to be most relevant and tries to stay away from irrelevant regions of the Web. In other words, it attempts to download Web contents that are likely to be harmful. As the result, the amount of contents to be analyzed by the classification module is dramatically reduced, so URL database can be refreshed more frequently. For this, the robot filters out irrelevant contents based on white list of classified URLs and harmless top-level domain names, e.g. edu, gov, and org. Harmless URL Filter in Fig 5 plays this role.

It is also guided by an objectionable Web content classification module which evaluates the relevance of a content with respect to the harmfulness. The classification module informs the robot of URLs of irrelevant contents and the robot stops retrieving links branched out from the irrelevant URLs. URL Link Relation Management module in Fig 5 keeps link relation showing a certain URL is referenced from which URLs. When it receives a classification feedback suggesting a certain URL is non-harmful, it retrieves all the child URLs branched out of the URL and does not collect Web contents corresponding to those child URLs assuming those are non-harmful.

As opposed to the robots for search engines which usually collect only text from the Web, this robot gathers both text and image because objectionable Web content classifier analyzes both. The downloading of image files severely damage the crawling speed as image data is bigger than that of text data and there are several images in one hypertext document. Therefore, we analyzed the characteristics of irrelevant images such as menus and lines, and created a profile for the images. Non-harmful Image Filter in Fig 5 determines if an image is likely to be non-harmful and, if so, it discards the image data.

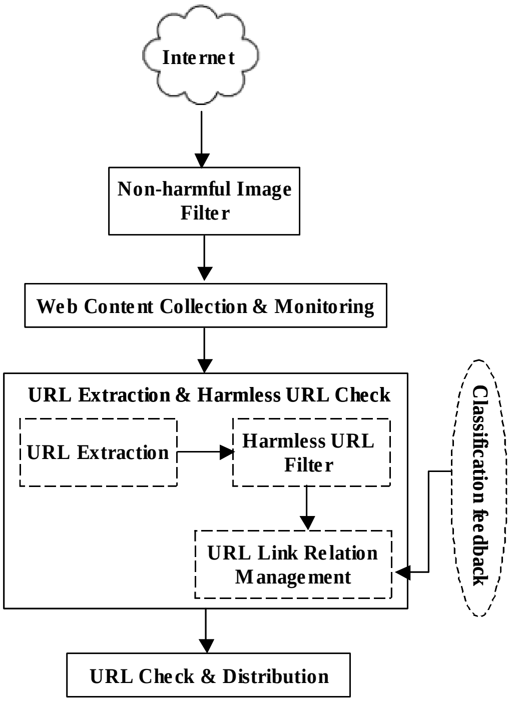

Fig. 5 Blocks for Acquiring Relevant Web Contents

**Achieving respectable coverage of objectionable Web contents:** The purpose is for the robot to cover large part of objectionable Web contents in order to increase the amount of URL database. If the crawl covers only 60% of the objectionable Web contents, URL database contains at best 60% of objectionable Web contents even with 100% correct classification module. In order to achieve high coverage, the seed URL for the crawl must include many URLs of objectionable Web sites. As new objectionable Web sites appear every day, it is important to include these in Seed URL. In an effort to discover new URLs of harmful sites, we employed meta search methods. We built a harmful keyword dictionary and executed meta search on several search engines with the keywords from the dictionary. As the result of the meta search, we can get the URLs which are likely to be harmful. By refining the URLs with white list, black list, and relevance feedback, URLs of new objectionable Web sites can be discovered. Fig 6 describes this process. We are sure that starting from the new set of URLs can expose the unknown part of an objectionable Web.

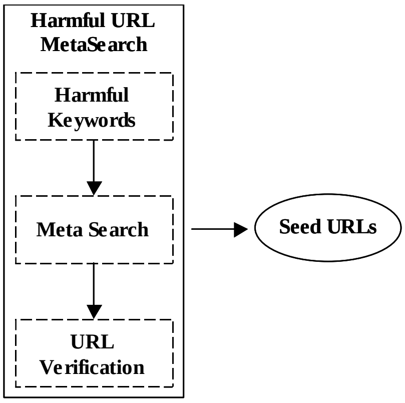

Fig. 6 Blocks for Achieving Respectable Coverage of Objectionable Web Contents

**Supplying helpful information to the classifier:** When a robot downloads texts and images, the objectionable Web content classifier takes only them as input. Therefore the classifier can miss important information for decision. For instance, Internet users might have experienced many redirections when attempted to connect to porn site, which is the characteristic of porn site. If the classifier can use this information in addition to text and image, the accuracy of detecting porn site can be increased. There are many characteristics that can be found by monitoring the session with objectionable Web site. This kind of information can be obtained only by a robot. Web Content Collection and Monitoring block gather this kind of information and classification accuracy is enhanced using this hint.

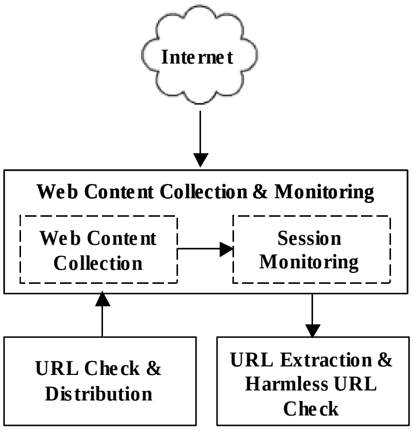

Fig. 7 Blocks for Supplying Helpful Information to the Classifier

## V. Conclusion

We showed the reason why general-purpose Web robot is not right for use in objectionable Web content classification system and defined three principles for building specialized Web robot. We also sketched the architecture and building blocks of the proposed Web robot. We are sure that this helps keep the URL database more up-to-date, achieve respectable coverage, and make the DB more accurate. The proposed robot is just right fit in the objectionable Web content classification system.

## References

[1] http://www.robotstxt.org/wc/faq.html#what

[2] SeungMin Lee, TaekYong Nam, JongSu Jang. http://kidbs.itfind.or.kr/WZIN/jugidong/1161/116101.htm. IITA itfind, 2004

[3] Soumen Chakrabarti, Martin van den Berg, and Byron Dom. Focused crawling: a new approach to topic-specific Web resource discovery, 8th International World Wide Web Conference, 1999.

[4] C. C. Aggarwal, F. Al-Garawi, P. Yu. Intelligent Crawling on the World Wide Web with Arbitrary Predicates, WWW Conference, 2001

[5] Porno Robot, http://www.allworldsoft.com/software/9-556-porno-robot.htm