Dynamic Web Content Filtering based on User's Knowledge

N. Churcharoenkrung, Y. S. Kim and B. H. Kang

School of Computing, University of Tasmania
Hobart, Tasmania, 7001, Australia
{ nc4, yangsokk, bhkang }@utas.edu.au

Abstract

This paper focuses on the development of a maintainable information filtering system. The simple and efficient solution to this problem is to block the Web sites by URL, including IP address. However, it is not efficient for unknown Web sites and it is difficult to obtain complete block list. Content based filtering is suggested to overcome this problem as an additional strategy of URL filtering. The manual rule based method is widely applied in current content filtering systems, but they overlook the knowledge acquisition bottleneck problems. To solve this problem, we employed the Multiple Classification Ripple-Down Rules (MCRDR) knowledge acquisition method, which allows the domain expert to maintain the knowledge base without the help of knowledge engineers. Throughout this study, we will prove the MCRDR based information filtering system can easily prevent unknown Web information from being delivered and easily maintain the knowledge base for the filtering system.

1. Introduction

As the volume of information available on the internet increases, the Web becomes a main source of personal or organizational knowledge. However, not all information on the Web is useful or relevant to users. For example, it is required to protect some information such as pornographic or criminal information, which is very harmful to children and users. Lots of irrelevant information is also presented to them. For example, only small numbers of search results are viewed by users when search engines provide outcomes according to the user's query. Without an appropriate filtering mechanism, users are overwhelmed with information and become frustrated.

Filtering systems are proposed to overcome these problems. In the research area, the content based approach and the collaboration based approach, are the two main approaches for information filtering. Whereas the former uses contents, such as keywords of the Web document, to filter out unneeded information, the latter uses other users' judgment against the contents, such as a rating scale. In the commercial area, URL filtering and IP systems are extensively used to eliminate irrelevant information because they are very easy to implement, work very fast, and produce acceptable success rates. However, the performance of this kind of system entirely depends on the exactness of registered URLs and IPs. If the filtering systems have incomplete blocking URLs and IPs list, the efficiency of the filtering system is quickly degraded. Nevertheless, it is very difficult to keep up with all relevant URLs and IPs because the Web is continually changing with no notification. The filtering systems will deteriorate without appropriate acquisition of new filtering knowledge. The content based filtering system and the collaboration based filtering system can be employed to enhance URL and IP based filtering systems.

In this research, we focus on development of the content based filtering system. Traditional content based filtering systems use content characteristics of message, which are usually represented by single tokens (words) or multiple tokens (phrases). Filtering knowledge is usually acquired by using rule based systems or machine learning based systems. One critical issue in the content based filtering systems is the knowledge acquisition or the learning mechanism. The filtering systems should be adaptable to the content or domain knowledge changes, because the topics of messages are not fixed in the real world. For this reason, our system uses an incremental knowledge acquisition method, called the MCRDR (Multiple Classification Ripple-Down Rules), which was introduced in 1990, and has often been used in the development of commercial knowledge acquisition systems.

This paper is structured as follows. Section 2 investigates related filtering research results. Section 3 provides explanations about the MCRDR method and the MCRDR based filtering system. Section 4 explains the experiment and its results for the efficiency test of the MCRDR based filtering system.

Conclusions and further work will be discussed in Section 5.

2. Literature Review

Hanini et al. [1] propose a classification of information filtering systems (IFS), which can perform their functions actively and passively. Whereas active IFS autonomously seek relevant information for users, passive IFS eliminate irrelevant information from incoming streams of data items. The IFS can be located at the information source, at a filtering server, or at the user site. Our system, which is located at the user site, aims to filter out irrelevant information passively.

Cognitive filtering and social filtering are major approaches of filtering tasks. Cognitive filtering, also called the content-based filtering, 'characterizes the contents of the message and the information needs of potential message recipients and then using these representations to intelligently match messages to receivers' [2]. Research based systems, such as SIFT [3], and the commercial systems (e.g., DansGuardian, iProtectYou, Parental Filter, Symatec Web Security, and We-Blocker) usually use URL or keywords to characterize content, while the social filtering systems use personal organizational interrelationships of individuals in a community. Some researchers interpret this as a collaborative filtering approach, which is now being referred to as 'recommendation systems' [4]. The cognitive and the social filtering approach have their own benefits and drawbacks but are regarded as complementary approaches [5]. For this reason, some researchers try to integrate these two approaches [5, 6]. However, our current research only focuses on content based filtering and our system uses keywords to represent messages.

Filtering knowledge can be obtained implicitly or explicitly [1]. The implicit approach utilizes the user's reaction to incoming data in order to learn from it. For example, time for reading messages, hyperlink clicking, document printing, and scrolling, are regarded as the user's interest expression [7-9]. The explicit approach requires users to fill out a form describing their areas of interest or relevant parameters. There are many rule based filtering systems that support users to construct more flexible filters, for example Lens [2] and ISCREEN [10].

Knowledge acquisition bottleneck is the main problem in rule based systems. Acquiring human knowledge is very difficult because knowledge is incrementally extracted from the domain expert and is differently expressed according to situations. As the size of the knowledge base increases, it is very difficult to create new knowledge without conflict with the existing knowledge. In the traditional knowledge based system, this problem becomes worse if the system user is not a system expert but a domain expert or a naïve user. He/she can not properly maintain the knowledge base without help from knowledge engineers. The Multiple Classification Ripple-Down Rules (MCRDR) - an incremental knowledge acquisition method - is proposed to overcome this problem and is based on the maintenance experience of a real world medical expert system, called GARVAN-ES1 [11, 12]. We use the MCRDR method to construct our filtering system and in the next section we provide a detailed explanation.

3. MCRDR Method and Filtering System Implementation

In the MCRDR system, knowledge is regarded as temporal, which means current knowledge is true only if the new situation is consistent with the old situation. New rules are usually added as exceptions to the current rule in the MCRDR system. This approach makes the validation process easier than the traditional approach, because the validation process only takes place in relation to the current rule and its children rules[11]. The MCRDR system has another knowledge acquisition facilitator, known as 'cornerstone cases'. Though cases have context information, they are usually discarded in the traditional rule based system. However, the MCRDR systems save this information as a cornerstone case while a new rule is created and used in the further knowledge acquisition process. This approach is beneficial because the user can more easily understand the knowledge acquisition context by employing a cornerstone case instead of an abstract rule. The MCRDR system also provides a difference list, which shows differences between current case, which become a cornerstone case of a new rule, and validation cornerstone cases, which are cornerstone cases of current rule and its children rules. This approach allows even a naïve domain expert to maintain a very complicated knowledge base without a knowledge engineer's help [13, 14]. For this reason, we employed the MCRDR method to construct a content based information filtering system because the main users of the filtering systems are not knowledge engineers but naïve domain users.

A content based filtering system is implemented with C++ programming language and the MCRDR knowledge acquisition method. Rules in the system are represented in an n-ary tree and each rule has condition keyword / keywords, class (pass or block), and cornerstone case. There are three kinds of rules in the system. The ground-breaking rule makes a new branch of rule groups under the root node. The refining rule makes an exception rule to the current rule. For example, when a current rule has condition 'gambling' and conclusion 'pass', a user can add a new condition such as 'gambler behavior research' and make a different conclusion 'block'. The stopping rule is similar to refining rule, but it has no conclusion.

Knowledge acquisition takes place when the filtering system suggests an incorrect inference result or no inference result. When a user initiates a knowledge acquisition process, the system questions whether the user wants to classify this case into 'pass' or 'block'. Once the user selects a class, the system shows all cornerstone cases that have the same class. If the user selects a case or several cases from the list, the system generates a difference list, which is an unduplicated words list between current case and selected cornerstone case or cases. The user can select conditions from this list and then the system generates case lists that will be reclassified by this rule. If the user wants to reclassify all these cases, he/she confirms reclassification, but if he/she does not want to reclassify some of these cases, he/she should add new conditions from the difference lists. The user can make a new consistent rule by following these procedures.

4. Experiments and Results

4.1 Data Sets

We randomly collected 283 web pages, which were reviewed to make a standard classification. To simplify the experiment, this study only classified Web pages into two categories – 'gambling' and 'non-gambling'. The gambling Web page displays gambling information such as casino advertisements, rules of play, etc. However, this research excluded information such as how to help people addicted to gambling, researches on the effects of gambling and other useful information that does not persuade Web site viewer to gamble. Table 1 summarizes the data. There were 146 Web pages classified as not gambling related and 137 classified as gambling related.

Table 1. Web page classification

Review	*Web pages*	*Ratio*
Pass	146	51.59
Block	137	48.41
Total	283	100.00

4.2 Training MCRDR Filtering System

The MCRDR filtering system was trained with 197 Web pages. The training process in the MCRDR systems was not a batch process like machine learning systems, but an incremental process like a traditional rule based system. The structure of the knowledge base was in the form of an n-ary tree. The tree had 42 rule nodes and four level heights, with most nodes on the second level. The first level had four nodes, the second level 25 nodes, the third level 11 nodes, and the forth level two nodes. Figure 1 illustrates the relationship between rule creation and its level in the tree (root is counted as level 0). It is evident that whereas most first level rules are added during the early training, the higher level rules are added during the late training phase. This relationship depends on the characteristics of the examples. For instance, if the examples are varied, this graph will have many dots at first level and less at higher levels. Another factor is the user's classification strategy – breadth first or depth first. For example, if a user creates the general rule first, the first level rules appear at the early stage and other rules above second level appear at the late stage. However, the first level rule can appear at any stage because knowledge can be incrementally acquired in our system and the user can add new general knowledge at the later stage.

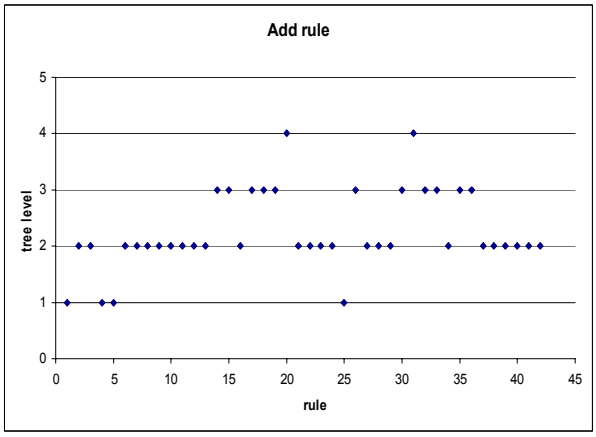
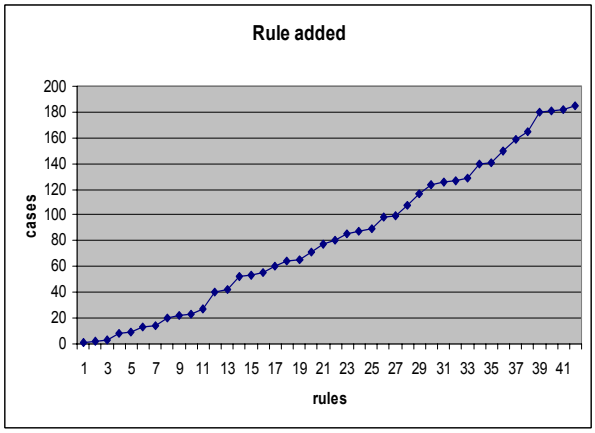
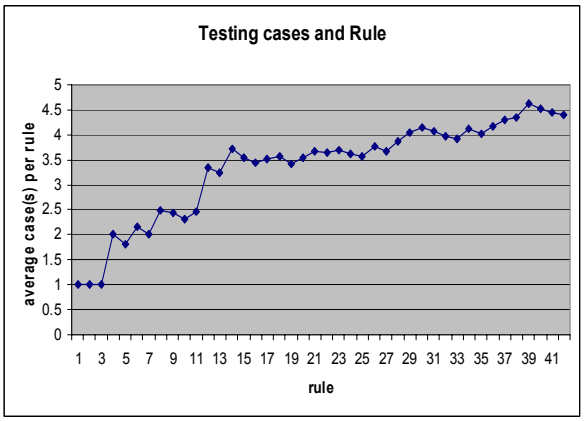


Figure 1. Rule creation and tree level.

The relationship between cases and rules is illustrated in Figure 2, which illustrates that as the number of rules grow steadily, the ratio of rule versus cases decrease (Figure 2 (a)). For example, whereas the first 10 rules were created with 20 articles, the last 10 rules were created with 60 articles. As the training of the filtering system becomes mature, the knowledge base is big enough to cover more cases, and the need for rule creation decrease. Figure 2 (b) illustrates the training cases used per the created rules. From this graph, it can be seen that when the first case enter the system, a rule was created, and it is similar to some cases, which came in during the early period of training.



(a) Rule and cases



(b) Cases per rule

Figure 2. Rule creation for training MCRDR filter

4.3 Testing MCRDR Filtering System

The MCRDR filtering system used 86 Web pages for testing. They were classified into four categories – passed (43 Web pages), blocked (39 Web pages), classified both passed and blocked (1 Web page) and unclassified (3 Web pages). The unclassified case or classified both categories case, occurs because the current knowledge base does not cover these cases, which means it is needed to acquire more knowledge. Figure 3 illustrates the correctness or incorrectness for each classification, gambling (passed) and non-gambling (blocked). The correctness of non-gambling is higher than the gambling, but is not much different. Both incorrect classifications are around 20 percent, which means the system does not under or over block.

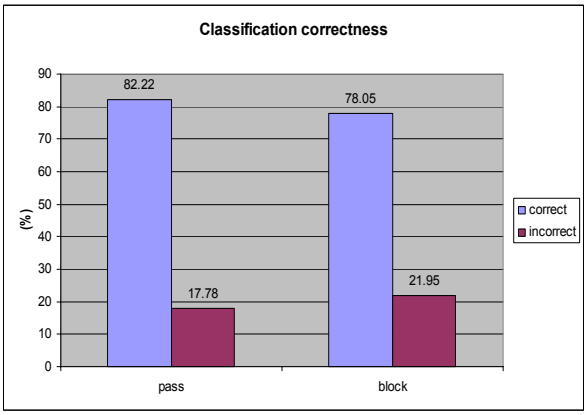
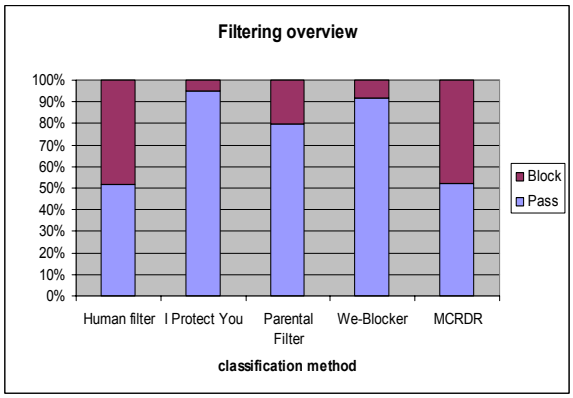
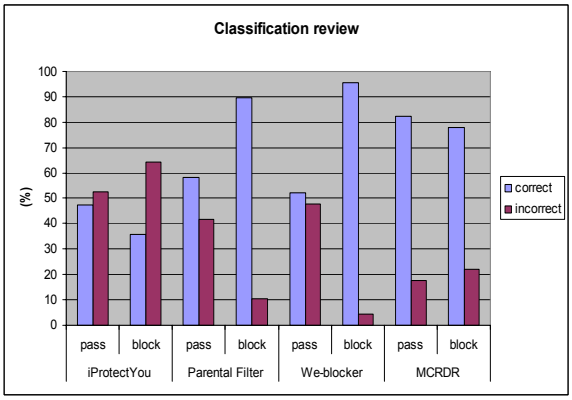


Figure 3. Filtering Correctness

We used three commercial filtering systems (iProtect, Parental Filter, and We-Blocker) as a comparison. Figure 4 illustrates percentage of Web pages that pass the filtering program. The human filter used human opinion for each page as to whether it should be passed (non-gambling) or blocked (gambling), and this opinion is the standard for this experiment. From Figure 4 (a), it can be seen that most testing filtering programs are under block, a high percentage of the passing Web page. The comparison is made as a percentage because the testing group of the MCRDR filtering system is smaller than others. Figure 4 (b) illustrates the correctness of the classification. The blocking correctness of Parental Filter and We-blocker is higher than the MCRDR filtering system. However, this can be explained because they rarely block the web pages (see the passing percentage in Figure 4 (a)) and therefore there are few incorrect blockings.



(a) Classification Results



(b) Classification Correctness

Figure 4. Classification Comparison

5. Conclusions and Future Work

On the Web, there is no centralized control of information, and the current research focuses on the active push style information delivery. Hence, filtering out unwanted information will be more important in the future. Applying content based filtering is necessary to protect users from unknown information. However, it is not easy to maintain the filtering knowledge in such systems. First of all, the knowledge for filtering can not be formalized and has to be updated whenever needed. Therefore, it is desirable to maintain the filtering knowledge without computer engineers or knowledge engineers.

Although there are some expert system development tools that provide the interface to maintain the knowledge base, the verification and validation of knowledge is left to the domain expert or end users. However, the domain expert is not good at representing their knowledge in the well organized structure. In many cases, it is sometimes impossible

because their knowledge is episodic and grown from many experiences.

This study proves that content based filtering is useful to protect people from unwanted information. In addition to this, we established that the MCRDR method can be used to maintain filtering knowledge. The advantage of the MCRDR method is that the system can easily acquire new knowledge without lots of the training cases when the knowledge in the system fails. In MCRDR, it is common that the knowledge in the system can be maintained by the domain expert without help from system engineers. The most interesting result of our study is that the performance of the system is similar to the human when it filters out the unknown information from the Web. In fact, MCRDR has been used in the development of many heuristic classification systems. We would like to prove that MCRDR can be used to maintain the heuristic filtering knowledge in the Web based information filtering systems.

In this study, we did not integrate URL based systems and content based systems. It is preferable to integrate two systems to measure the exact performance enhancement. The MCRDR filtering system should be extended to check other content such as image, HTML links, HTML tag information and various media types.

There are some pages that hardly contain text, displaying many pictures instead. If the MCRDR program could check for the picture's name, the performance might be better. However, checking the picture's name may not always work because many are named as sequence e.g. pic1.jpg. For that kind of title, this function will be ineffective.

HTML links are also an interesting component to check. There are some pages which give only links on their pages but will not cause a difficulty if they use texts as hyperlink. The problem will occur where only pictures and links are used. The current version of MCRDR filtering system ignored these two contents.

There are a few tags that might also be checked, such as META tag and script. Sometimes META tag contains quite useful information for filtering including Web page description, Web page keyword, etc. However, this information is given by the Web master and it has no standard, except for PICS, which may prove unreliable.

The MCRDR filtering system should recognize other pattern changes in the Web sites. They seem more interactive. To do the interaction, they use other components than HTML tags, such as Flash. For the moment, this kind of web site can not be checked. So the MCRDR system might need other methods to check this component. PDF is also a problem because it uses its own format and the data transmitted is not readable by humans.

6. References

[1] Hanani, U., B. Shapira, and P. Shoval, *Information Filtering: Overview of Issues, Research and Systems.* User Modeling and User-Adapted Interaction, 2001. 11(3): p. 203-259.

[2] Malone, T.W., et al., *Intelligent information-sharing systems.* Communications of the ACM, 1987. vol.30, no.5: p. 390-402.

[3] Yan, T.W. and H. Garcia-Molina, *SIFT-a tool for wide-area information dissemination.* Proceedings of the 1995 USENIX Technical Conference, 1995: p. 177-186.

[4] Resnick, P., et al. *GroupLens: an open architecture for collaborative filtering of netnews*. in *Computer Supported Cooperative Work*. 1994. Chapel Hill, North Carolina, United States: ACM Press New York, NY, USA.

[5] Balabanovic, M. and Y. Shoham, *Combining Content-Based and Collaborative Recommendation.* Communications of the ACM, 1997. 40(3).

[6] Claypool, M., et al. *Combining Content-Based and Collaborative Filters in an Online Newspaper*. in *ACM SIGIR Workshop on Recommender Systems*. 1999.

[7] Morita, M. and Y. Shinoda, *Information filtering based on user behavior analysis and best match text retrieval.* SIGIR '94. Proceedings of the Seventeenth Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval, 1994: p. 272-281.

[8] Konstan, J.A., et al., *GroupLens: applying collaborative filtering to Usenet news.* Communications of the ACM, 1997. vol.40, no.3: p. 77-87.

[9] Goecks, J. and J. Shavlik, *Learning users' interests by unobtrusively observing their normal behavior.* IUI 2000. 2000 International Conference on Intelligent User Interfaces, 2000: p. 129-132.

[10] Pollock, S., *A rule-based message filtering system.* ACM Transactions on Information Systems (TOIS), 1988. 6(3): p. 232-254.

[11] Kang, B., P. Compton, and P. Preston. *Multiple Classification Ripple Down Rules : Evaluation and Possibilities.* in *9th AAAI-Sponsored Banff Knowledge Acquisition for Knowledge-Based Systems Workshop*. 1995. , Banff, Canada, University of Calgary.

[12] Compton, P. and D. Richards, *Generalising ripple-down rules.* Knowledge Engineering and Knowledge Management Methods, Models, and Tools. 12th International Conference, EKAW 2000. Proceedings (Lecture Notes in Artificial Intelligence Vol.1937), 2000: p. 380-386.

[13] Compton, P. and R. Jansen, *A philosophical basis for knowledge acquisition.* Knowledge Acquisition, 1990. vol.2, no.3: p. 241-258.

[14] Compton, P., et al., *Knowledge acquisition without analysis.* Knowledge Acquisition for Knowledge-Based Systems. 7th European Workshop, EKAW '93 Proceedings, 1993: p. 277-299.