Monitoring Web Browsing Habits of User Using Web Log Analysis and Role-Based Web Accessing Control

Phudinan Singkhamfu, Parinya Suwanasrikham

Chiang Mai University, Thailand

0659

The Asian Conference on Society, Education and Technology 2013

Official Conference Proceedings 2013

Abstract

Computer network in campus is an essential facility for education environment. The university network has provided knowledge accessibility to students. Students in undergraduate school sometimes get involve with many studying distractions, and use campus network in the wrong direction. Redundant connections in campus network as the main problem are challenging to overcome. The aim of this research is to measure the unfavorable web access habit of undergraduate student in campus computer laboratory room by capturing, and identify unwanted web connection from every connection from the room and to use the result of the first part to develop web access filtering algorithm to control the campus outgoing web connection. Role-based web accessing control (RBAC)[6] model was used to filter the system and content-based analysis was used to measure the quality of each web. Firstly, the web quality keywords are set and extracted from web log by text mining algorithm. Then, the ontology of web quality keywords is created. The ontology of keywords makes more accurate controlling web access of students. We implemented and embedded algorithm into a web browser and running the experiment on nursing and software engineering students of Chiang Mai University. Students are assigned role for RBAC and task to search on specific contents. The embedded algorithm monitored their browsing and blocked unrelated web. It is found that the algorithm enables to check accessing web more accurate than keywords without ontology. The web, not related on given content, was blocked or limited accessing. In this research, we combine role-based web accessing control and content base monitoring to control web accessing of student in campus. The ontology of content keywords increases accurate blocking unrelated web. Ontology adds more semantic relation to keywords.

1. INTRODUCTION

Content browsing control in university's network is challenging in term of automate system design. It determines the browsing habit of students as target group. Inappropriate the Internet using mostly from this user group and the network traffic is dramatically rising. The web can be useful tool for gathering educational resources; although using the university limited network traffic resource in wrong direction will cause many consequent effects for instant, more network maintenance tasks, and more expensive operation cost to consider.

The major concern is to know what the user had been browsed under the university network, and how to determine which is good or bad website. There is difficulty for staff to square all computer log files around the faculty's lab machines. The age gap between staff and student has raised the other problematic issue as the browsing habit of student age is significantly different with matured staff.

The web monitoring system can be used to assist staff in gathering the browsing log from each machine, and analyze what type of the content of each page is. Staff can easily monitor student browsing habit by setting up the content category function to create a list of good website, or irreverent website, and use the list to design web filtering system as a final solution.

Role based access control (RBAC) [1] is access model, which use role as a bridge between subjects and privileges. It capable to managing access control up to large number of principals.

This paper wills present benefit of RBAC, and an implementation on the web applies to work with the web monitoring system. We designed the web monitoring system by use context-based classification algorithm to categorize web content from the log using content keyword matching, the content will be provided from BHO's (Browser Helper Object) browser extension program to gather, and indexing the web log from target browser program on every computer laboratory machines, and stored it in system database. The result of this implementation is unwanted web black list, which ready for applying in the future design faculty's web filtering system.

2. RESEARCH METHOD

2.1 Web Content Categorization

The Web content categorization is based on content-based monitoring algorithm. The algorithm provides content analysis from monitored URL from target browser, using keyword-based technique analysis, the algorithm scans for matching with pre-defined keyword rules with content in each page. This project aims to design the categorization system in limitation of scope, which are as follows:

- Understand quality of content in each page
- Categorize types of content
- Propose flexible keyword, and rule manipulation
- Return final result as "Student", "Staff & Lecturer", and "Visitor"

Matching content with pre-defined group of website was the main propose of content analysis as illustrated in table 1 below. When the process completed the system is able to return result of content which already categorized to each group.

Pre-define Group	Good	Modulate	Bad
Game			✓
Entertainment News		✓	
Pornography			✓
Movie		✓	
General News	✓		
Wikipedia Content	✓		
Technology News	✓		
Travel Information		✓	
Facebook		✓	
Bit torrent			✓

Table 1. Example of Pre-define content group which might gain access from student in campus network

The group is instantly justified by rating marked as good, modulate, and bad in the first stage of analysis to understand and provide the weight score of each user activity; for example, page which has gaming content would be classified as "game" group and the game group has mark as "should be student" who using the browser in that moment. The "modulate" group refers to medium weight score and mark as "should

be anyone". The result of this table will analyzed again with deeper layer of keyword to justify the final result in the next step.

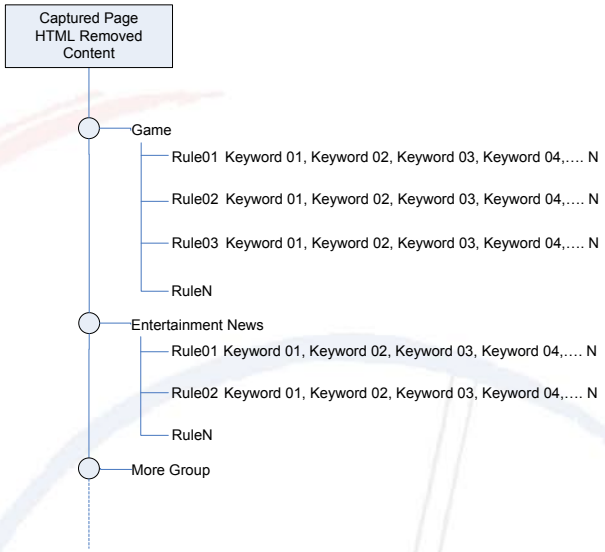

Fig. 1. Rule management for group categorization

Fig. 1 illustrates keyword in structure of categories, which separates and describes as a rule. User may need to select category of classification content to gain more accuracy of classification. The system is working more accurate with user browse to search for specific content, and obtusely keyword defines in to the classification rules. Some faculties, like nursing school and engineering school, are easily distinguished between relates and non-related content. The browsing pattern of student in school or campus of the mentioned schools would be easily predicted by using our technique. Furthermore, some rules can also design to capture fault statement that might appear in the same page, such as student searching for the operating system installation tips on one of famous computer forum. Some forum topics may have a heading looks involving with searching topic; however, the content may promote the products instead of providing right information. In this case we could use logic operator to control each rule and also can be able to control between keyword in the rule.

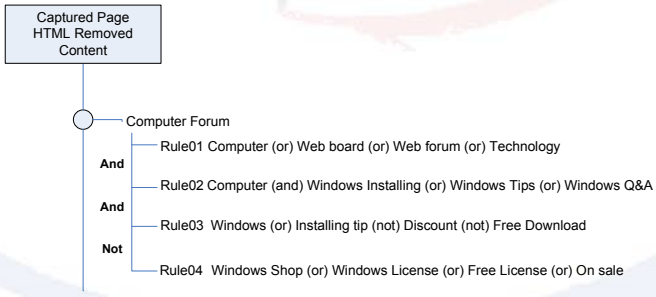

Fig. 2. Assigning logic operator between rule and inside rule

Using logical operator logic control between rules, the system has more flexibility, and increase more precise result[3]. *Fig. 2* shows rule creation possibilities when applying logical operator to control action between, and in inside the rules. We can

design various complex rules as user can predict what the captured page would be involved.

2.2 Rule-Based Access Control and Context Rule-Based Access Control

RBAC is model to control accessing to data or content for authorized users. RBAC can be categorized in many types. RBAC supports many constraint policies. There are different dimension of constraints in RBAC. The examples of constraint categories are [6]

- Static constraints that are constraint that evaluated on design state of RBAC model.
- Dynamic constraints that are constraint that tested on working state of RBAC model.

Not even static/dynamic constraints, they can be categorized as

- Authorization constraints are constraints which add more details for access control process. If users are granted
- Assignment constraints are constraints which control the election of permissions and roles to users.

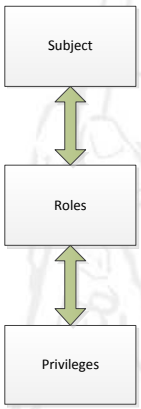

Figure 3. RBAC Relation [6]

RBAC provides a powerful mechanism for reducing the complexity, cost, and potential for error of assigning users permissions. Also, many policies can't be used by standard RBAC. Normally, access control base on identity of user or subject [5]. However, there is a dramatic change in technologies and knowledge.

Roles base server provides the privilege to categorize user as a ticket to verify at the web filtering system in final stage. The privilege from Roles base server was contained in secession, which will come along with user with in limited of session time.

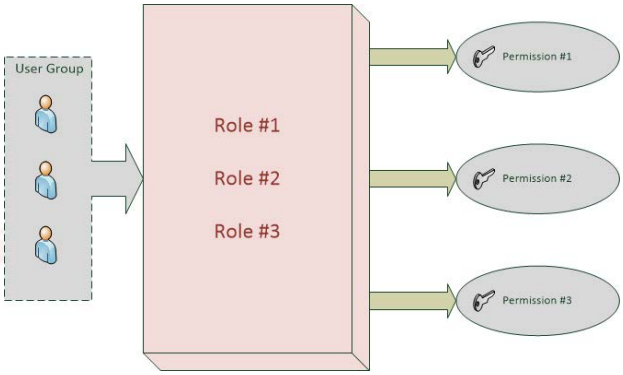

Figure 4. User and role to permission accessing in RBAC

Fig. 4 shows relation between user and permission, which assigned by role in RBAC [7], user is input as subject from *Fig. 1*. Acquiring role type from role server or service to gain permission, user group is also required to classify into the mentioned three different groups from the content-based classification system.

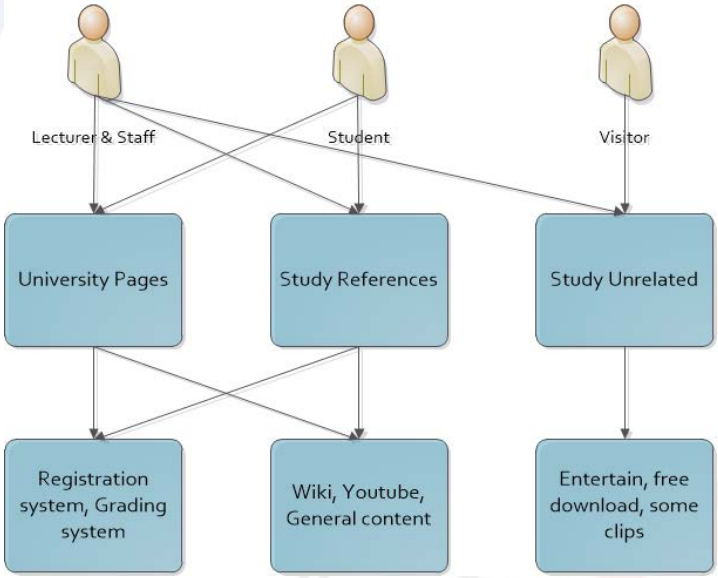

Figure 5. RBAC permission design for each user roles

The privilege assessment system performs in *Fig. 5* shows example roles in each user, and accessing designs for suitable with each user group. Obviously, the most accessibility in the system is Lecturer & Staff group due to, this user group required wide area of work, and research, and on the other hand, student user group is only required to access into studying related activities, and some few of University web based study assistance pages. Entertainment and many download centers may filter for this user group, due to bandwidth consumption overload problem preventing.

3. RESULTS AND ANALYSIS

This paper was aimed to use this algorithm to learn about behavior of user who use computer network in campus. The model can predict different the Internet browsing behavior of users. It was intent to use the system with three different groups, as shown on table 2. The system considers individual user from start using the Internet and browsing page normally. URL that has passed through student's browser from schools proxy sync student log in for the first time, is recorded and analyzed.

Pre-define Group	Student	Lecturer & Staff	Visitor
Game	✓		✓
Entertainment News	✓	✓	✓
Pornography	✓		✓
Movie	✓		✓
General News	✓		✓
Wikipedia Content	✓	✓	
Technology News	✓	✓	✓
Office web pages		✓	
Facebook	✓	✓	✓
Bit torrent	✓		✓

Table 2. Final stage of user classification mapping result

The classification mapping table shows the way of defining content to user group based on role based control. The challenging of this experiment is context-based classification system, which shows more accurate result from distinguishing keyword. For example, game keyword can address clearly the same as pornography which have many specific keyword and easy to create classification rule, then passed on to role base system to assign role to each user group to handle in filtering system in the next step.

4. CONCLUSION

From all observation and experiments, it is very clear that the privilege assessment system perform its jobs to control access to group of web content-related to user group. This conclusion supports the idea to use the algorithm's result to control user-browsing habit. The study shows some major difficulties of process in the privilege handling, which attached with user from the authentication and capture all of required attributes to filtering system for acquired the appropriate accessing role.

However, it is not a complete experiment. Under simulation scenario, this study acquires briefly results within limited period of time. This experiment is able to

expand to bigger scale with implementing the filtering system which will applied the assigned roles from incoming request by reading the design secession. It is advisable to develop new techniques to reduce error close to zero as much as possible. Regarding to the framework, it is intended to membered???? this algorithm with bowser by using BHO (Browser Helper Object) to transfer captured data from user instead of our current method which need permission from network administrator, which is one of our research difficulties.

Finally we intend to use this experiment result as one of school web filter block list setting up criteria, which will be able to act in many blocking roles depending on user group. This would help network administrator to determine the unrelated and inappropriate web page more efficiently.

REFERENCES

[1] Nicola Guarino,, "Content Base Access to the Web," *IEEE Transactions on Industrial Electronics,* vol. 99, pp. 70-80, 1999.

[2] Jiang Su, and Harry Zhang., "A Fast Decision Tree Learning Algorithm," American Association, 2006.

[3] P. Y. Lee, and S. C. Hui., "An Intelligent Categorization Engine for Bilingual Web Content Filtering", IEEE, 2005

[4] Xiaolong Zhang, Na Zeng., J Tang, and Xin Xu, "User Behavior Monitoring for Campus Networks," IEEE, 2010

[5] Gustaf Neumann, Mark Strembeck., "An Approach to Engineer and Enforce Context Constraints in an RBAC Environment" *SACMAT'03*, *June 2–3, 2003, Como, Italy.*

[6] JOON S. PARK, RAVI SANDHU., "Role-Based Access Control on the Web" *ACM Transactions on Information and System Security,*vol 7, pp 37-71

[7] Mohammad A. Al-Kahtani, Ravi Sandhu., "Rule-Based RBAC with Negative Authorization" *Proceedings of the 20th Annual Computer Security Applications Conference (ACSAC'04)*

iafor