International Journal of Control and Automation
Vol.9, No.1 (2016), pp.191-200
http://dx.doi.org/10.14257/ijca.2016.9.1.17

The Research of Enterprise Security Search Engine

Xiaopu Ma[1,*], Xing Chen[1], Ning Cheng[1] and Ruixuan Li[2]

[1]School of Computer and Information Technology, Nanyang Normal University
[2]School of Computer Science and Technology, Huazhong University of Science and Technology
mapxiao@nynu.edu.cn

Abstract

With the rapid development of computer and information technology, people are focusing on how to find valuable information from the vast contents quickly and efficiently. As a result, the search engine system is a solution to this problem. However, there are few relatively work done in access control for search engine system when security is a critical requirement for broader applications of the technology not only in the current but also in the future. In this paper, an architecture for enterprise security search engine based on role-based access control is presented, and based on this architecture the search engine system can not only crawl the documents from the controlled resource sites but also can return different searching results from the index database according to the user's different identities and permissions for the same keywords. We believe that the proposed architecture is realistic and secure.

Keywords*: security search engine, role-based access control, single sign on*

1. Introduction

Nowadays, with the rapid development of computer and information technology, especially the revolution of Internet technique, people are focusing on how to find valuable information from the vast contents fast and efficiently [1]. As a result, the search engine system is introduced to meet with this demand for the people. Despite the evolution of search engine system to more perfect, there are few relatively work done in access control for search engine system, especially the security demands for search engine system in enterprise products and enterprise management systems.

1) Nowadays, the strong demand of people for information is specialization, refinement, non-commercial, the strong demand makes the emergence of security search engines in professional fields. However, in the approach to search valuable information from the vast contents through the resource sites, a key challenge that has not been adequately addressed so far is how to generate different results according to the user's different permissions or identities. In other words, most of the existing search engine system did not consider the different nature and importance of each user, or treated each user evenly. But this is not always the case. For example, the user has the permission "read" of the patient's personal information may be more important than the user has the permission "write" to the patient's personal information. This is because the "read" permission usually leads to more information leakage, but the traditional search engine system simply ignores this different, hence all the users can search and read the patient's personal information if they use the same requirement keywords to search the patient's personal information based on the traditional search engine system not matter how permissions they have. In another case, as we know, most of the administrative operations are more sensitive and important than the ordinary ones of general users [2]. Hence, when the administrator and the ordinary users search the valuable information through the

ISSN: 2005-4297 IJCA

search engine system based on the same requirement keywords, the search engine systems may be only show a small number of results to the general users, while the administrator are usually associated with huge amounts of results in order to providing the safety mechanism for the enterprise information. However, the traditional search engine system cannot do like this because they cannot gain the user's identities or permissions. According to these situations, it has been suggested that the future development of search engine system will largely depend on the availability of novel methods for ensuring the people obtain different resources according to their different permissions or identities when they use the search engine system.

2) On the other hand, the traditional search engine system also cannot crawl the documents from the controlled resource sites, especially when the resources are in the enterprise interior, Internal campus network and so on because different enterprises or organizations need to assign different access permissions for the different user groups according to their own interests or other resources on the site in the implementation of access control. In this situation, the original resources cannot be crawled for the traditional search engine system. Therefore, it is very necessary to design a comprehensive access control mechanism that is general and flexible enough to reflect and cope with the special access control requirements associated with the search engine system, especially for specific applications where security is a critical requirement, such as the applications in insurance company and bank or in enterprise interior.

To this aim, in this paper, we introduce a role-based access control (RBAC) [3,4] architecture for enterprise security search engine system which uses single sign on [5] technique and role based access control policies to provide scalable and efficient access control services for controlled resource sites and different users in order to find different searching results according to the users' different identities or permissions. The remainder of this paper is organized as follows. We discuss related work in Section 2. The limitations in existing application of search engine system drive our motivation and Section 3 presents the role-based access control architecture for the enterprise security search engine system, and discusses why introduce the single sign on strategy and how to integrate role-based access control policies into the enterprise security search engine architecture. Finally, Section 4 provides some insight into our ongoing and future work.

2. Related Work

A search engine system is an information retrieval system designed to help people find valuable information from the vast contents that were stored on the resource sites or on the Internet fast and efficiently [6]. The search results are usually presented in a page hits list. A web search engine system is the most public, visible form of a search engine system which searches for information on the World Wide Web [7]. In this web search engine system, a crawler will fetch as many documents as possible from the resource sites firstly. Secondly another indexer program will read these fetching documents and create an index based on the words contained in each document which will be stored into an index database in order to provide for the searcher to search results according to the user's requirement keywords. Each search engine system uses a proprietary algorithm to create its indices such that, ideally, only meaningful results are returned through searcher for each query. According to the way of search engine system, the web search engine can be divided into three categories:

1. Selection-based Search System [8]: In the selection-based search engine system, user only uses the mouse to invoke a search query in order to search the valuable information from the resource sites. In this selection-based search engine system, it allows the user to search the internet for more information about any requirement keywords or phrase contained within a document or webpage in any software application on his desktop computer through the mouse. For example, people can find some shortcuts information such as images, videos, local, shopping, audio and so on from Yahoo shortcuts which is a selection-based search engine system that was first developed by Reiner Kraft in 2007.

2. Web Search Engine System [9]: The web search engine system is designed to search for information on the World Wide Web according to the user's requirement keywords. The results of the web search engine system will be presented by a line of results pages which may be a specialist in the web pages such as images, information and other types of files. Some search engine systems also can mine available information from databases or open directories. Unlike web dictionaries, which are maintained only by human editors, search engine system also maintains real-time information by running a web crawler algorithm.

3. Meta-search Engine System [10]: The meta-search engine system is a search tool that makes use of other search engine systems. In this search engine system, it gets the users' requirement keywords firstly and send users requests to several other search engine systems or other databases. Finally, the search engine system will aggregate the results into a single list or displays them according to their source from the different search engine system.

While using the search engine system, the goal is to find valuable results exactly equal to the users' requested information, but it is restrained when the search engine system is used in the region of Intranet because of confidential need. Firstly, the search engine system cannot grasp documents from Intranet for the safety measures. For example, the traditional search engine system cannot grasp the resource from the resource sites where there is access control information on it. Secondly, the popular search engine system treats every user equally that cannot return different results according to the user's different identities and permissions. Hence, in order to overcome these disadvantages, we develop the enterprise security search engine system based on role-based access control to meet with the special needs in the controlled resources or for the different users' identities or permissions.

In the security area, RBAC is the most popular access control model at present, and is widely used as an alternative to the traditional discretionary access control (DAC) and the mandatory access control (MAC)in enterprise security management and enterprise management products[11-12]. In RBAC, a set of permissions is assigned to users through roles. This change on how to assign the permissions often reduces the complexity of access control because the number of users is generally much larger than that of roles in an organization. The most distinctive and important feature of the RBAC is the desire to specify and enforce enterprise-specific security policies in a way that maps naturally to an organizations structure. Its emphasis on controlling who has access to operations on what objects is fundamentally different from information flow security in multi-level secure systems, and therefore can support three well-known security principles: least privilege[13-14], separation of duties[15]and data abstraction. As a result, RBAC has been implemented successfully in a variety of commercial systems, such as insurance company and bank, and has become the norm in many applications. Hence, we can introduce RBAC architecture for enterprise security search engine system which uses single sign on technique and RBAC policies to provide scalable and efficient access control services for different users in order to match the safety management of enterprise on the controlled resource sites.

3. Preliminaries

We develop the material in this paper in the context of the NIST standard, the most widely known RBAC model [16]. It consists of RBAC0, RBAC1, RBAC2 and RBAC3 as shown in Figure 1. The last two incorporate separation of duty constraints. For the sake of simplicity, we do not consider sessions or separation of duties constraints in this paper.

Definition 1. The RBAC model contains the following components:

- *U*, *R*, *P*, the set of users, roles and permissions respectively;
- $PA \subseteq R \times P$, a many-to-many mapping of permission to role assignments;
- $UA \subseteq U \times R$, a many-to-many mapping of user to role assignment relationships;
- $UPA \subseteq U \times P$, a many-to-many mapping of user to permission assignments;
- $auth\_perm(R) = \{p \in P \mid (p, R) \in PA\}$,the mapping of roles onto a set of permissions;
- $RH \subseteq R \times R$ is a partial order of *R* called the role hierarchy or role dominance relation, also written as $\succcurlyeq$, where $r_1 \succcurlyeq r_2$ only if all permissions of r_2 are also permissions of r_1, and all users of r_1 are also users of r_2, i.e., $r_1 \succcurlyeq r_2 \Rightarrow auth\_perm\ (r_2) \subseteq auth\_\ perm\ (r_1)$.

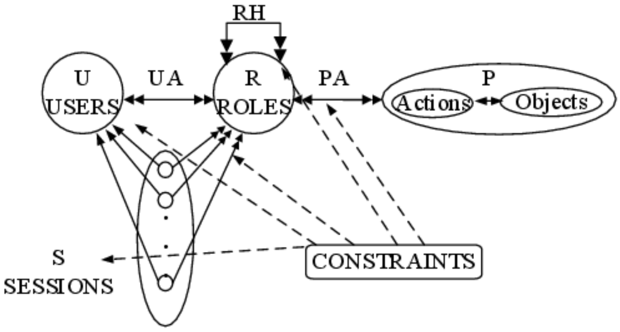

Figure 1. The Components of RBAC Model

Figure 2 depicts the architecture for the enterprise security search engine based on RBAC. We use the dashed rectangular frame to denote the traditional structure of the search engine system, the major RBAC model and single sing on model of the enterprise security search engine system are represented in real line rectangular frame. The RBAC architecture for the enterprise security search engine contains the following components:

1) Crawler Model: A crawler is a computer program that browses the information in a methodical, automated manner or in an orderly fashion from the resource sites. As the crawler visits these URLs, it identifies all the hyperlink in the page and adds them to the list of URLs to visit. In the traditional search engine system, crawler is mainly used to create a copy of all the visited pages or resources from the resource sites for later processing by a search engine system. When the search engine system obtains the crawled resources, it will index the download pages or resources to provide fast searching for the people obtaining valuable resources from the vast contents according to the requested keywords. However, with the rapid development of information technology, there is existing vast information deployed in enterprise management products. In this situation, security failures can disrupt an organization’s operations and can have financial, legal, human safety, personal privacy and public confidence impacts. Hence the enterprise needs to apply the access control mechanism to control the actions, functions, applications, and operations of legitimate users within an organization in order to protect the integrity of the information stored within the system. However, as organizations increase the functionality and information offered on internal and external networks, controlling access to information and other resources becomes more complex and costly. In these access control model, RBAC is a relatively new access control system that maps to organizational specific structures in a way that reduces direct and indirect administrative costs and improves security. Hence, we can introduce the RBAC model into the search engine system to protect the security of the information in the enterprise.

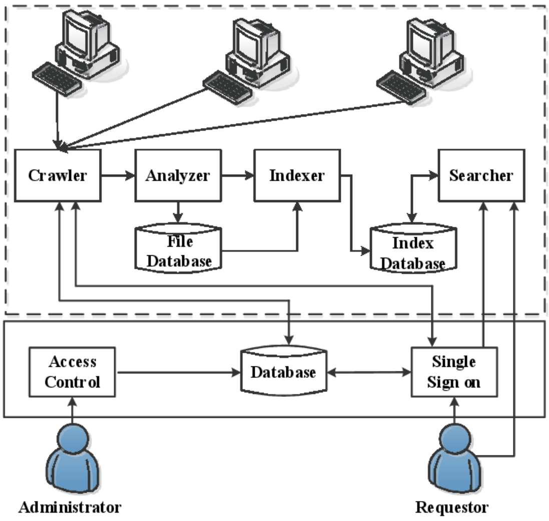

Figure 2. A Role-Based Access Control Architecture for Enterprise Security Search Engine System

According to this situation, the traditional crawler cannot gather documents from the controlled resource sites. Furthermore, the different users in the enterprise security management products will have different identities and permissions when they want to search information from the controlled resources. In other words, the different users need search different results according to their different identities and permissions. According to this, in RBAC architecture for the enterprise security search engine system, the crawler model needs to get the sites' different authentication information in order to login in the different resource sites to grasp the resources firstly. Furthermore, the crawler also needs to obtain the different user's identities in order to gather the different resources according to the users' identities and permissions. Here, the crawler will crawl the resource site from the home URL of resource website home page according to each resource site registration information from the database server. The process of crawler can be decomposed into two types:

- According to the public resources, crawler finds all the public resources from the initial URL link of the resource website home page, then add all the resources to the index file, store in a public document indexing, write all the crawled URL into the hash table in order to provide page analysis reference for controlled resources packet crawling. Hence, the public resources are no longer analyzed in the controlled resource grouping crawler.
- According to the controlled resources, crawler firstly finds the resource web site according to the description information of web site as described in Table 1, secondly finds the user group information according to the description information of user and user group as shown in Table 2 and Table 3 respectively. Finally, the crawler model sends the login parameters to the single on model in order to access the controlled resource sites. If the login successful, the crawler model can gather the different resources from the different controlled resources sites according to the different authentication information that it gets. In the establishment of the group index, packet crawling will set ID as a field into index (according to the public resources, the value of this field is public).

Table 1. The Description Information of Web Site

	Site ID	Site Name	Site URL	Login URL	Logout URL
Item$_1$					
…					
Item$_N$					

Table 2. The Description Information of User

	User ID	User Group ID	Site Group ID	Comment
Item$_1$				
…				
Item$_N$				

Table 3. The Description Information of User Group

	User ID	Password	User Web ID	User Rank	Comment
Item$_1$					
…					
Item$_N$					

2) Analyzer/Indexer Model: The purpose of storing and index of the search engine system is to optimize speed and performance in finding relevant documents according to the user's identities based on the search requirement keywords. Without an index, the search engine system would scan every document in the corpus, which would require considerable time and computing power. For example, while an index of 10,000 documents can be queried within milliseconds, a sequential scan of every word in 10,000 large documents could take hours. Hence, when the crawler obtains the different documents from the different controlled resource sites or public resource sites, the analyzer and indexer model will collect, parse, and construct the index on the resources according to the documents and the different user's identities information. Finally, the documents will be stored into the index database for searching according to the different user's authentication information. Thus, RBAC architecture for enterprise security search engine can provide different results according to the user's different identities and permissions. Furthermore, the sensitive information will not only be safety through this architecture but also can be crawled by the enterprise security search engine system.

3) Searcher Model: A search model is a query that a user enters into the enterprise security search engine system in order to satisfy the users' information demands. The steps of the enterprise security search model based on RBAC can be described as Figure 3.

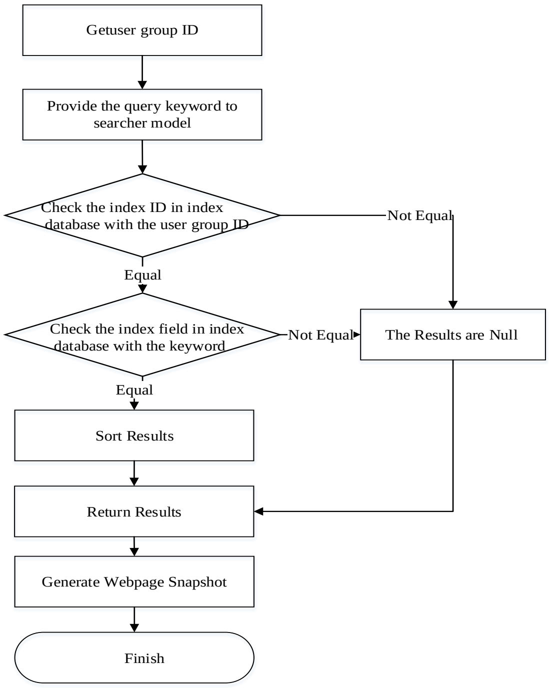
Getuser group ID
Provide the query keyword to searcher model
Check the index ID in index database with the user group ID
Not Equal
Equal
Check the index field in index database with the keyword
Not Equal
The Results are Null
Equal
Sort Results
Return Results
Generate Webpage Snapshot
Finish

Figure 3. Steps of the Searcher Model

4) Access Control Model: The access control model is an interface for the administrator representing the access control information for the controlled resource sites, users, user groups respectively. As shown in Table 1, the administrator can operate the information table on database server to register the description information for the resource sites, such as the default user login access website name, the website home page, login URL, the access control strategy and the default login parameters. The description information of users includes user ID, password, level (for hierarchical access control), the user sites such as ID attributes, which a user site ID is founder of the user belongs to the website ID, as shown in Table 2. For a site, the number of users may be a lot of, the access control policies can be different, but there are always many users have the same permissions. Thus the system will have the same permissions to users on the same site for the same user group, the description information is shown in Table 3.

5) Access Database: When the administrator describes the different access control information for the resource sites and users, the enterprise security search engine system needs store the information into the access control database. The access control database will provide the user's name and password and role information for the crawler in order to get the resources from the different controlled resource sites. When the access control database receives the user's name and password from the single sing on model, it will check the details of the user according to information which already were stored into the access control database and send the results to the single sing on model.

6) Single Sign On model: Single sign on (SSO) model is an independent software system that also a property that has the multiple related with the access control system.

SSO is mechanism whereby a single action of user authentication and authorization can permit a user to access all computers and systems where he has access permission, without the need to enter multiple passwords. As different applications and resources support different authentication mechanisms, single sign on has to internally translate to and store different credentials compared to what is used for initial authentication. In this architecture the single sign on model will receive the information from the requestor or crawler and then provide them to the access control database, if the user's name and password are certified, the single sign on model will send the authentication information about the user to all the resources site, the authorization information also will send to the searcher model in order to get the different results according to the user's identities and permissions.

4. Conclusions

With the rapid development of computer and information technology, people want to search valuable information from the vast contents fast and efficiently. However, it should protect the integrity and safety of the information stored within the system at the same time when the search engine system searches the valuable information according to the users request query where security is a critical requirement. Hence, in this paper, we have introduced a role-based access control architecture for enterprise security search engine system to solve this problem, the RBAC architecture integrates single sing on and RBAC policies not only satisfy the access control requirements about the controlled resources site, but also have the merit of searching different resources according to the users different identities by employing the single sing on technology. For the future work, we will evaluate our technique with others. Moreover, we will utilize the access control information to further protect the information safety and efficient.

5. Acknowledgements

This work is supported by the Joint Fund for National Natural Science Foundation of China and Henan Province for Fostering Talents under grants U1304619, National Natural Science Foundation of China under grants 61173170, 61300222, 61433006 and Innovation Fund of Nanyang Normal University under grants ZX2013013. We thank the anonymous reviewers for their helpful comments. This paper is a revised and expanded version of a paper entitled "A Search Engine System Based on Role-based Access Control" presented at The 6th International Conference on Computational and Information Science, May 30-June 1, China West Normal University.

References

[1] R.A. Baeza-Yates and B. Ribeiro-Neto, "Modern Information Retrieval", Addison Wesley Longman Publishing Co., Inc., Boston, MA, USA **(1999)**

[2] M. Xiaopu, L. Ruixuan and L. Zhengding, "Role Mining Based on Weights", Proceedings of the 15th ACM Symposium on Access Control Models and Technologies (SACMAT), **(2010)** June 9-11; Pittsburgh, USA

[3] D.F. Ferraiolo, J.A. Cugini and D.R. Kuhn. "Role-based access control (RBAC): features and motivations", Proceedings of the 11th Annual Computer Security Applications Conference, **(1995)** December 11-15; New Orleans, Louisiana

[4] R. Sandhu, E.J. Coyne, F. Hal and C.E. Youman. "Role-based access control models. IEEE Computer", vol. 2, no. 29, (**1996).**

[5] Z. Zhengli, Q. Yan and S. Mingwen. "Analysis on Grid Security Patterns Based on PKI", AISS: Advances in Information Sciences and Service Sciences, vol. 4, no. 23, **(2012).**

[6] T. Joachims, D. Freitag and T. Mitchell, "Web Watcher: A tour guide for the World Wide Web", Proceedings of the15th International Joint Conference on Artificial Intelligence, **(1997)** August 23-29; Nagoya, Aichi, Japan.

[7] S. Lawrence and C.L. Giles, "Searching the World Wide Web", Science, vol. 3 **(1998)**

[8] R. Singh, "Performance of World Wide Web" search engines: A comparative study. Library Herald, **(2006).**

[9] B.S. Biradar and S.B.T. Kumar "Internet search engines", A comparative study and evaluation methodology", SRELS journal of information management, vol. 3, no. 43, **(2006)**

[10] M. E. Manoj and Jacob. "Information retrieval on Internet using meta-search engines", A review, JSIR (CSIR), vol. 67, no. 10, (**2008).**

[11] L. Ruixuan, W. Wei and M. Xiaopu, "Mining roles using attributes of permissions", International Journal of Innovative Computing, Information and Control. vol. 8, no. 11, **(2012)**

[12] M. Xiaopu, L. Ruixuan, Z. Lu, J. Lu and Meng Dong. "Specifying and enforcing the principle of least privilege in role-based access control", Concurrency and Computation: Practice and Experience. vol. 23, no. 12, **(2011)**

[13] L. Chen and J. Crampton, "Inter-domain role mapping and least privilege", Proceedings of the 12th ACM Symposiums on Access Control Models and Technologies (SACMAT), **(2007)** June 20-22; Antipolis, France

[14] F.B. Schneider. "Least privilege and more", IEEE Security and Privacy, vol. 1, no. 5, **(2003)**

[15] N. Li, M.V. Tripunitara and Z. Bizri, "On mutually-exclusive roles and separation of duty", Proceedings of the 11th ACM conference on computer and communications security. **(2004)** November 3-7; Scottsdale, Arizona, USA

[16] D. Ferraiolo, R. Sandhu, S Gavrila, D.R. Kuhn and R.Chandramouli. Proposed nist standard for role-based access control. ACM Transactions on Information and System Security, vol. 4, no. 3, **(2001)**

Authors

Xiaopu Ma, He received the M.S. degree in School of Computer Science and Engineering at University of Electronic Science and Technology of China in 2004, and the Ph.D. degree in School of Computer Science and Technology at Huazhong University of Science and Technology in 2011. He is an associate professor in the School of Computer and Information Technology at Nanyang Normal University. His research interests include distributed system security and access control.

Xing Chen, He received the M.S. degree in School of Computer Science and Technology at Wuhan University of Technology. He is a lecturer in the School of Computer and Information Technology at Nanyang Normal University. His research interests include parallel computing and grid computing.

Ning Cheng, He received the M.S. degree in School of Computer Science and Technology at Jilin University. He is a lecturer in the School of Computer and Information Technology at Nanyang Normal University. His research interests include distributed system security and access control.

Ruixuan Li, He received the B.S. M.S., and Ph.D. degrees from School of Computer Science and Technology at Huazhong University of Science and Technology in 1997, 2000, and 2004, respectively. He is now a full Professor of School of Computer Science and Technology at Huazhong University of Science and Technology. His research interests include distributed system security, information retrieval, peer-to-peer computing, and social network.