

Content-based Management of Document Access

Control

Edgar Weippl, Ismail Khalil Ibrahim

Software Competence Center Hagenberg

Hauptstr. 99, A-4232 Hagenberg, Austria

{edgar.weippl, ismail.khalil-ibrahim}@scch.at

Werner Winiwarter

Electronic Commerce Competence Center

Siebensterngasse 21/3, A-1070 Vienna, Austria

werner.winiwarter@ec3.at

Abstract

Three different security models have been advocated to determine the access rights to documents in a network of computers; DAC, MLS, and RBAC. Each of these models has its strength and weakness. No one can urge that any of these models can offer alone by itself an automated support for either the security or the access rights. In this paper, we propose a content-based management model for the access control to documents in a large enterprise. The model determines the access rights to documents based on their content and automatically classifies the access levels or detects possible incorrect settings. The content-based document access control can be used in advanced business applications to allow developers to provide a high level integration of security models within business applications.

1 Introduction

During the last decade the continuing spread of PCs lead to an ever-increasing number of digital documents. At first, most documents were stored locally and

information security depended heavily on the physical security of individual computers. However, when documents are stored on networked computers, access control becomes essential for security.

Various security models have been proposed; they can roughly be categorized into discretionary access control (DAC) and multi-level security (MLS). For MLS a user must specify the security level of new data. Changing this categorization to lower the level is commonly prohibited (i.e. no-write-down property).

On the other hand, DAC implies that the access rights are specified for every single object. Obviously, this approach makes it difficult to implement a consistent security policy for documents across a large company. Today, role-based access control (RBAC) is used to address this problem by introducing hierarchies of roles to facilitate administration. However, most current systems still do not support clear RBAC allowing the specification of DAC on a user-object basis; this is mainly due to requirements for backward compatibility.

The approach described in this paper offers automated support for either setting the security level (MLS) or the access rights (DAC). The document's access rights will be assigned according to its content. For instance, if a CEO writes a memo about an invitation to all employees, there will hardly be any access restrictions whereas if she composes a memo about the company's future business strategy access would be strictly limited.

The paper is organized as follows: Section 2 gives an indicative overview of security models and highlights properties of these models that are relevant in the context of this paper. In Section 3, we elaborate on the processing steps required to automatically classify text in documents for which we are trying to determine access control. In Section 4, we propose the use of content-based analysis to either automatically classify access levels or to detect possibly incorrect settings. We conclude in Section 5 with discussion and ongoing work.

2 Access Control

Computer security deals with the prevention and detection of unauthorized actions by users of a computer system. Therefore computer systems control access to data and shared resources, like memory, printers, etc., both for reasons of integrity and for confidentiality. Access control is at the core of computer security. In more general terms access control tries to enforce security by controlling and limiting subjects' access to objects. Knowledge management is basically concerned with providing access to required information. At the same time, however, content not appropriate to certain people should not be disclosed. Hence effective access control can be deemed to be very relevant for managing a large corpus of information.

Within the last decades three different forms of access control evolved: multi-level security (MLS, also referred to as mandatory access control (MAC)), discretionary access control (DAC) and role-based access control (RBAC) [Gollmann 99].

Discretionary access control (DAC) controls access to an object on the basis of an individual user's permissions and/or prohibitions. This means that it is at the discretion of an object's owner to decide who has which kind of access. For instance, in the UNIX file system the owner of a file can decide whether to grant read, write or execute rights to herself, her group or to everyone. In the context of knowledge management this would require every user to decide for himself who should be allowed to access a document, rendering it very difficult to enforce a consistent security policy across an organization.

A role is a collection of operations (on specific objects) needed for an application. Assigning access rights to subjects based on their role is called RBAC. RBAC is commonly used in database systems as it offers two main advantages. First and foremost, RBAC clearly separates between "what has to be done" and "who has to

do it” in that it assigns users to roles and defines permissions on roles, which in turn takes effect when users activate the corresponding role.

Second, RBAC clearly distinguishes an application’s object model from the subject and authorization model. The object model provides a view on objects to protect (e.g., tables, columns, entities), the subject model highlights which entities are active within a system (e.g., users, processes), and the authorization model describes rules regulating access between subjects and objects and the administration thereof.

MLS systems follow a fundamentally different approach. In MLS systems content is classified into different levels (e.g. ‘secret’, ‘internal use only’ and ‘public’) and ensure that the content can never be moved to a level of lower classification. For example, information about planned investments, classified at the ‘secret’ level cannot be copied to ‘public’ level.

These forms can be compared qualitatively and quantitatively in terms of their (1) adequacy to produce a high level, software independent conceptual model that gives a clear abstraction of what is needed to be protected in the system and describes the functional and structural properties of the security policy and (2) their efficiency to allow developers to give a high level definition of the protection requirements and system policies as well as to produce a concise description of the desired system behavior.

3 Classification of Context

In this section we elaborate on how text can be categorized based on its content. The described approach works even if no meta information is available as it can be solely based on the texts’ words. However, if meta information is available, some straightforward changes in the algorithm should be made that dramatically improve the quality of the clustering.

3.1 Key Term Vectors

We start by building a matrix in which each row represents a text and each column a keyword. For each element of the matrix we count how often a keyword appears in each text. Thus the matrix consists of a key-term vector for each text. The term 'key-term vector' refers to 'term vectors' mentioned by Chalmers [Chalmers 96]. By using stemming and only considering nouns we focus on key-terms only and therefore speak of key-term vectors.

As not all keywords are equally important we apply a well-known transformation thereby multiplying each item of the matrix by a weight, the so called 'Term Frequency - Inverse Document Frequency' (tf-idf) [Baezo-Yates 99], [Salton 89].

3.2 Clustering

The key-term matrix represents the content of all texts in a high dimensional vector space. Within this space texts form clusters according to the topics they deal with. The task is to find a clustering that also preserves the topology of the space. To facilitate the visualization of relations between text documents we propose to find a 2D projection of the high dimensional vector space. The WebSOM project [Honkela 97] illustrates how well a Kohonen net [Kohonen 97] can perform this task. Figure 1 shows how approximately 300 texts have been categorized into 49 clusters.

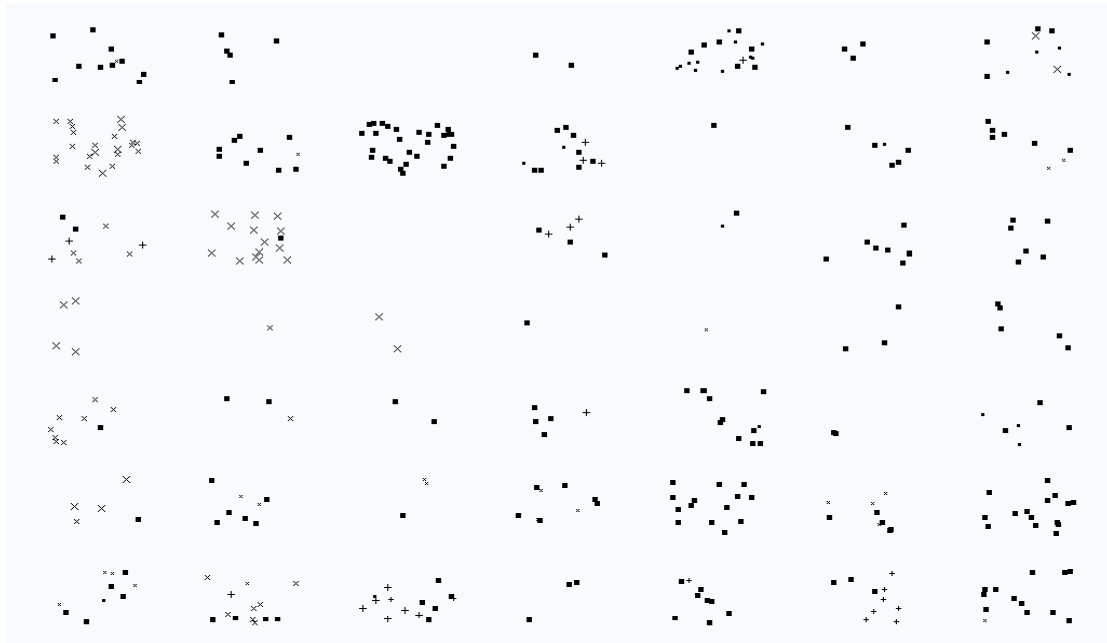


Figure 1: Using a 7 by 7 SOM to visualize content-based relationships between approximately 300 texts.

4 Automatic Assignment

The main contribution of this paper is to propose the use of content-based analysis as previously described to either classify access levels (MLS) automatically or to detect possibly incorrect settings (DAC).

4.1 Assigning MLS levels

Every organization stores documents dealing with different topics such as files about employees and customers, technical documents, sales and marketing information, accounting data, etc. Using the techniques described above all these documents are placed in different regions in a 2D SOM. Figure 2 shows how topics could be spread across the display.

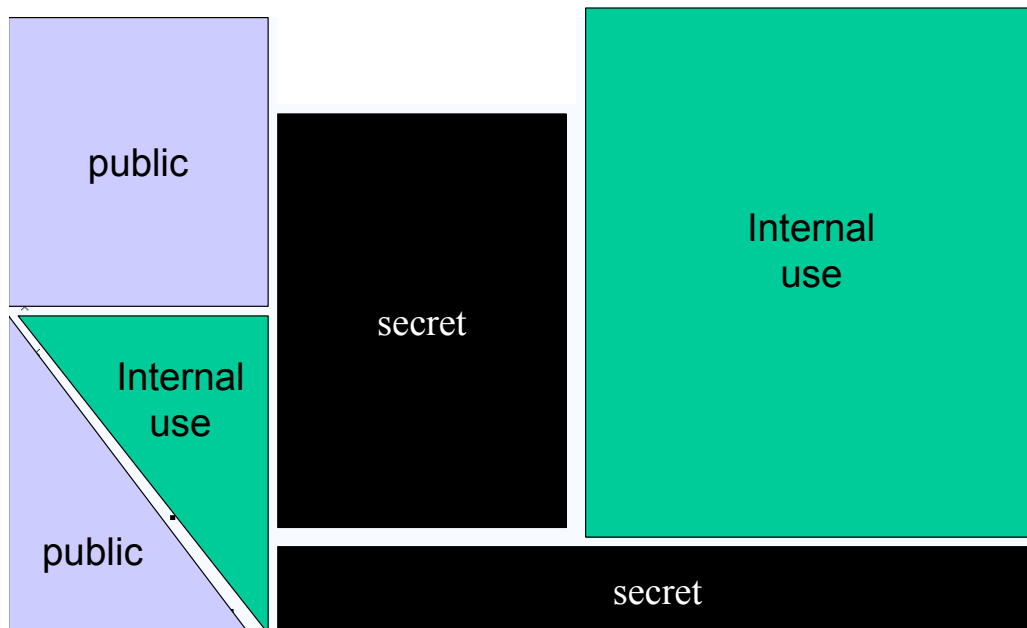


Figure 2: By marking regions and assigning access control levels to them, all existing and future documents that fall within the cluster will be classified accordingly.

The only human input required to set the MAC levels is the assignment of security levels to regions previously identified. If new documents are added that fit within the existing categorization, they will automatically be assigned the corresponding access control level. By measuring how well a document fits into a cluster or more precisely by observing the magnitude of the error, it is straightforward to detect when a new cluster has to be created.

4.2 Checking DAC settings

As MLS is not commonly used outside the military, we will elaborate on how a similar content-based approach can be used to check DAC settings. In DAC access rights can be defined quite simply in the form of an access control matrix; each row corresponds to one subject (e.g. user), each column to one object (e.g. file). Every element of this matrix contains information which (if any) access rights a certain subject has for a specific object.

If access rights are consistently specified for all documents according to their content, rows in the access control matrix of neighboring (in the 2D SOM), i.e. related texts, should not differ greatly. It is therefore feasible to detect texts that have been assigned unusual and possibly incorrect access rights.

5 Discussion and Ongoing Work

In this paper, we have proposed a content-based management of document access rights in large networked enterprises. This approach is based on building a key-term matrix that shows how often a keyword appears in a text and by using a well known transformation technique the texts are classified into clusters. The main contribution of this paper is to propose a content-based analysis of the clusters to automatically classify access levels or to detect possibly incorrect settings. This approach has its advantages over previous forms by relieving the application developer from the burden of having to design, implement, and test access control mechanisms. However, two major shortcomings were identified in our approach that need to be addressed in future efforts in this direction. Firstly, as most organizations do not use MLS and find DAC's organizational overhead too large, they adopt role-based access control (RBAC) as the best strategy for the future. The extension of our approach to RBAC is obvious only if access rights on a subject-object basis should be checked. As every hierarchical specification of role models can easily be transformed into an access control matrix, we can directly apply the previously mentioned approach. However, another essential feature in the context of RBAC is to check whether roles and not only the implied access rights are assigned in a consistent way.

Secondly, in the current state our concept is only usable in a very limited environment as the assigning of access rights is based on content only and not on how detailed information is. The level of detail, however, significantly influences who should be allowed access. For example, consolidated financial statements of listed

corporations are freely available whereas details on financial plans should be kept secret.

6 References

- 1 Ricardo Baezo-Yates and Berthier Ribeiro-Neto: Modern Information Retrieval. Addison-Wesley, 1999
- 2 Matthew Chalmers: A Linear Iteration Time Layout Algorithm for Visualising High-Dimensional Data. Visualization 96 Proceedings. IEEE Society Press, 1996
- 3 Dieter Gollmann: Computer Security. John Wiley & Sons, 1999
- 4 Timo Honkela et al.: WEBSOM - Self-Organizing Maps of Document Collections. Proceedings of WSOM'97, Workshop on Self-Organizing Maps, Espoo, Finland, June 4-6. Helsinki University of Technology, Neural Networks Research Centre, 1997
- 5 Teuvo Kohonen: Self-Organizing Maps. Springer-Verlag, 1997
- 6 Gerard Salton: Automatic Text Processing: The Transformation, Analysis and Retrieval of Information by Computer. Addison-Wesley, 1989