A Multi-Agent System for Enforcing "Need-To-Know" Security Policies

Young-Woo Seo, Joseph Giampapa, and Katia Sycara

The Robotics Institute, Carnegie Mellon University, Pittsburgh PA 15213, USA

Abstract. We propose a multi-agent system architecture for the adaptive authorization of access to confidential information. The proposed multi-agent system provides "need-to-know" content-based authorization of requests to access confidential information. "Need-to-know" authorization is that which grants access to confidential information only if that information is necessary for the requester's task or project. In our system, we treat the authorization task as a text classification problem in which the classifier must learn a human supervisor's decision criteria with small amounts of labeled information, e.g. 20 to 30 "documents", and to be capable of generalizing to other documents with a zero, or near-zero, false alarm rate. Since "need-to-know" authorizations must be determined for multiple tasks, multiple users, and multiple collections of confidential information, with quick turn-around from definition to use, the authorization agent must be adaptive and capable of learning new profiles quickly and with little impact on the productivity of the human supervisor and the human end-user. To this end, we examined five different text classification methods for solving this problem, "agentified" the best performer, and inserted it in a secure document management system context.

1 Introduction

Most information systems for securing confidential information have relied on a manually compiled access control list (ACL). Each item of confidential information is associated with an ACL, which ensures a corresponding level of security and can be accessed by anyone who has been authorized. For the purposes of indexing and security, some information is grouped into containers based on similarity of their contents or similar levels of confidentiality. A secure repository holds all these containers encompassed by a limited access system. A request to access the confidential information may occur, for example, when an employee is assigned to a new project and needs to access background knowledge.

This approach, however, has a crucial security weakness in that an authorized user to a segment of confidential information in a container is actually able to access the entire container. For example, an employee who is authorized to look at a progress report on the development of new technology is able to access the information about a financial plan for that project; the two pieces of information are rated at the same level of confidentiality and hence are held in the same container. An engineer may not need not to know – indeed perhaps should not know – a financial plan. Therefore the supervisor of the collection of confidential information will either hand select only those documents that he will let the user see, or completely bar access to the entire collection rather than risk exposing documents that should not be exposed.

Furthermore, this approach is inflexible. It does not allow easy adjustment to frequent changes of a user's task assignment. Project assignments for an employee may be changed quite often and hence the employee needs to access confidential information related to the newly assigned project. In addition, access to a previously assigned project may need to be revoked. In order to ensure that authorized access is granted and unauthorized access is denied, the ACLs for all information associated with the project must be updated according to the rights and the permissions of employees assigned to the project.

In order to provide a solution for these problems, we propose a multi-agent system that takes advantage of classification algorithms for the authorization of requests for confidential information. Instead of relying on coarse-grained ACLs and hand- selected information, our system reads the content of requested confidential information and compares it with the content of the requester's project. The request is accepted if the two contents are similar, otherwise it is rejected. In the case of either acceptance or rejection, the event can be logged for security audits and alarms.

By doing this, our method allows the supervisor a means of specifying subsets of per-user and per-task access control policies, and a way to automatically enforce them. Since the proposed system learns, or adapts to, the supervisor's decision criteria based on a small number of supervisor-provided examples, and generalizes from such examples, the supervisor need not identify all relevant information. Through our proposed system, it then becomes possible for the supervisor to define, assign, and enforce a security policy for a particular subset of confidential information.

The paper is organized as follows. Section 2 describes the motivations for this work and observations on the characteristics of the domain of confidential information management. Section 3 outlines related research. In the following section 4 we describe a conceptual architecture of our multi-agent system. Section 5 details the adaptive authorization process in terms of text classification. Experiments are described in section 6 and we concluded with discussions in section 7.

2 Background

Most organizations protect their confidential information via hierarchical access control policies, ensuring the principle that confidential information should be accessed only by those who have been authorized. Here, confidential information refers to property containing knowledge that is sensitive to an individual or organizations; hence its careless release may have a damaging effect. Consider the case, for example, of a celebrity or politician admitted to a hospital. There might be strong incentives for physicians or medical staff not directly tasked

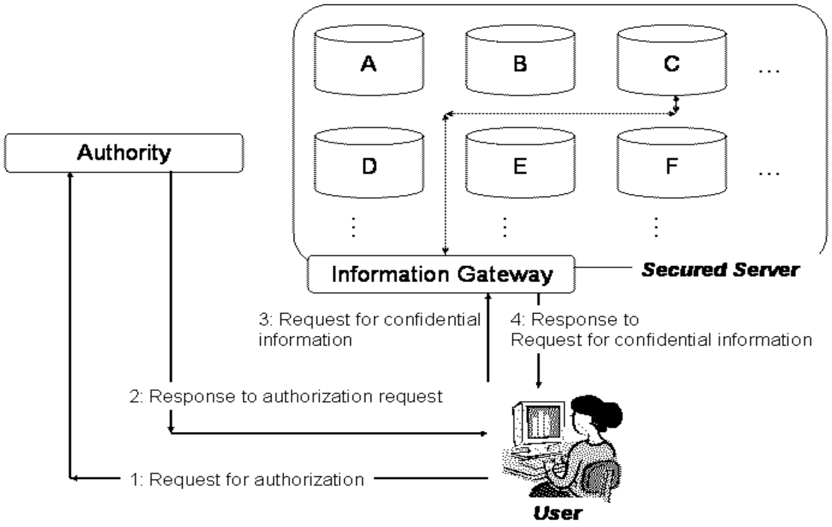
A
B
C
...
D
E
F
...
Authority
Information Gateway
Secured Server
3: Request for confidential information
4: Response to Request for confidential information
2: Response to authorization request
1: Request for authorization
User

Fig. 1. A typical scenario of access to confidential information.

with caring for the patient to anonymously leak the medical condition of the patient to a third party. The confidential information would be more secure if it was only released to primary care physician and the assigned nursing staff on shift, than to all physicians and nursing staff for that unit. Therefore access should be granted according to an employee's role and task.

Figure 1 shows a typical scenario of access to confidential information. The underlying assumption we make in this scenario is that all confidential information is stored in a secured repository (e.g., secured server). A piece of confidential information can be archived in many forms such as a text file, images, voice-recordings, etc. The secured repository is a strong-hold and the only way to access information in that repository is to present a request over an information gateway. Requests for confidential information are successful only when they are granted by the appropriate authority. Another assumption we have is that an employee or user needs to acquire background knowledge related to the project to which he has been assigned. He must get authorization to access the background knowledge if it is rated as confidential.

In Figure 1, the confidential information that an employee depicted as “user” needs to know is part of the container labeled “C”, not the entire contents of “C.” In order to access it, the employee must get authorization, which is usually done by requesting that his supervisor – an example of a controlling authority – allows him access. With the authorization to a part of the container “C,” the employee is now able to access the entire part of the container, which includes other information that the employee “need not know.”

For this typical scenario, we would like to enumerate the properties that an information system should have in order to ensure that confidential information is secured.

- **Coverage** An information system must be able to protect all confidential information according to given security policies. In other words, an item of confidential information should only be accessed by employees who need to know it, in order to perform their duties.

- **Adaptiveness** The work-flow of an information system must be flexible enough to adapt to frequent changes of project assignments to personnel. Regardless of current position, an employee of an organization is not always expected to participate in a particular project during the entire period of employment; there are frequent changes in project assignments, which last only for limited periods. Likewise, authorizations for access to discrete units of information often need to change.
- **Inexpensive Maintenance** The maintenance of an information system must be relatively inexpensive in terms of human time and effort, and computational resources, and be easily updated.

The simplest and most intuitive way of applying the "need-to-know" principle is to rely on the conscientiousness of employees. This approach may work in cases where the constituents all have the same objective and will not pursue individually-motivated objectives. For most cases, however, this method is far from reality.

Another approach is to have a completely sealed space. Access requests to confidential information are then processed only within that space. The space is built in a high security facility and only allows an authorized employee to use the consoles wherein the confidential information is stored. An authorized employee must verify his identity using a retinal or finger-prints scanning system. This approach may assure the high-level security for the confidential information. However, it is difficult for an ordinary organization to achieve, due to the high monetary maintenance costs. In this age of ubiquitous information access, it may also be very inconvenient for employees to use this facility because they must physically present themselves there. A medical doctor in a time-critical situation, for example, will not always be able to go to this facility to look at his or her patient's records given a time-critical medical emergency.

The most frequently used technique for protecting access to confidential information is the use of an Access Control List (ACL). There have been three major approaches for this type of security protection [Gollmann, 1999]. The first is a multi-level security (MLS) approach that imposes mandatory security policies on all confidential information hierarchically. The second approach is a discretionary access control (DAC) that works on the basis of the creator's permissions: the creator of confidential information is responsible for the decision of who has which kinds of access. Finally, a role-based access control (RBAC) approach maintains a list of roles that encapsulates access rights to a particular set of confidential information. The users under this type of system are assigned corresponding roles according to their responsibilities. Basically, these approaches compile relatively static and complex ACLs and impose them in response to requests for confidential information. However they are slightly different from one another in what they emphasize: DAC and MLS focus on the item to protect and RBAC focus on the role of user. However, these approaches have three problems:

- **Incomplete Coverage** It is difficult for these approaches to avoid all possible unauthorized attempts, even though a complete list of ACLs is complied with. In particular, an employee who is allowed to access a certain level of security might be able to access all information at that level, even if they should not. He is not supposed to access all of those items, but at the same time he is officially allowed to access them.

- **Non-Adaptiveness** Since ACL assignments to confidential information always meets an organization's "need-to-know" principle, the assignments must be done before an employee is assigned to a new project. If the time period between task assignments is relatively short, the maintenance of static ACLs may require a lot of effort. Given these constraints, it is quite difficult for these approaches to adapt to the frequent changes of project assignments. ACL and role-based approaches become inconvenient if the content to be protected cannot be easily or uniformly specified.
- **Expensive maintenance** The maintenance and assignment of ACLs to all possible combinations is not easy to accomplish. Since the number of ACLs that must be updated increases cubically in proportion to the units of confidential information, the groups of constituents, and the number of operations, it is computationally and monetarily expensive to maintain.

3 Related Works

There has been extensive research on multi-agent systems (MAS) in a wide spectrum of applications from financial portfolio management [Sycara, 1998] to air traffic control.

Sycara [Sycara, 1998] specifies the desiderata for multi-agent systems and introduces a multi-agent system called RETSINA. She also presents applications of this system, such as financial portfolio management, decision support, and crisis management. Some work in the multi-agent community associated with the issue of security deal with guaranteeing secure interactions between agents, including secure communications between agents and the estimation of agent reliability in a multi-agent community [Mouratidis et al., 2003], [Golbeck et al., 2003].

Wagner proposed a multi-level security policy for multi-agent communications, in order to ensure that only authorized agents could share information [Wagner, 1997]. Tan and his colleagues [Tan et al., 2003] suggested a dynamic security policy model for heterogeneous multi-agent systems that provide open services, where a policy regulates ACLs and communication services.

[Ferraiolo et al., 1995] outlined the characteristics of a role-based access control. [Chandramouli and Sandhu, 1998] had a comparative study of RBAC implementations in a couple of commercial databases, given that role-based access control has primarily been applied to database management systems.

[Weippl and Ibrahim, 2001] proposed a content-based management of text document access control. They applied a self-organized map (SOM) to cluster a given collection of text documents into groups which have similar contents. It also allowed humans to impose DAC to identified text document groups. However they did not address a potential problem that occurs when the security policy for individual documents of a cluster does not match with the security policy

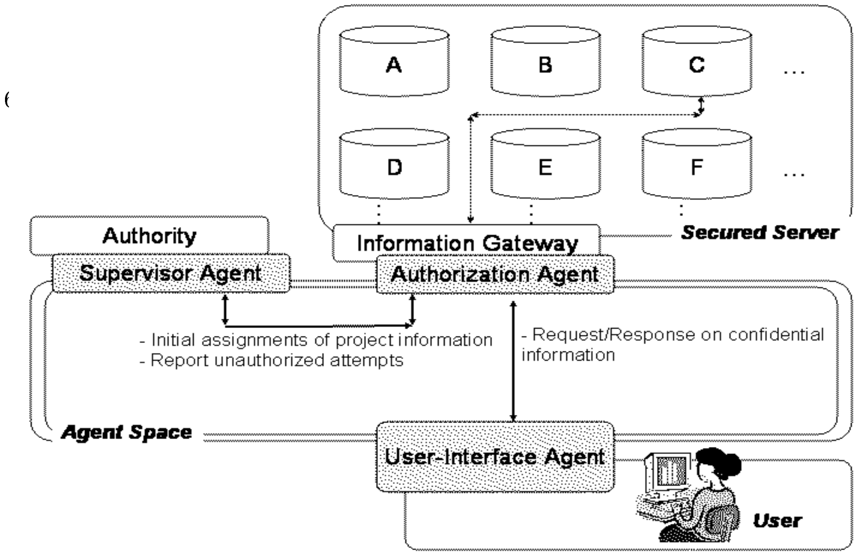

Fig. 2. A conceptual architecture of a multi-agent system for managing confidential information. An agent-space replaces interactions between humans with communications between software agents.

for that cluster. [Giuri and Iglio, 1997] proposed an approach that determines a user's access to confidential information, which is based on the content of the information and the role of the user. For example, they consider subdividing medical records into several different categories (e.g., pediatrics), and allow that only relevant physicians (e.g., pediatrician) can access them. Since they do not mention automatic techniques in their paper, one is left with the suspicion that they manually categorize content and roles.

4 A Multi-Agent System for Adaptive Authorization of Confidential Information

A multi-agent system (MAS) is an appropriate approach for managing confidential information in that the multiple threads of control are a good match for the distributed resources to be secured and the ever-changing nature of task assignments in typical organizations. A single-agent system could still take advantage of an intelligent agent's properties, such as adaptiveness and proactiveness. However, it would be vulnerable to a "single point of failure" and could not manage the large amount of requests for information access on time [Sycara, 1998]. Furthermore, an MAS can more easily manage the work-flow and response to changes of environment and can provide the spontaneous mapping of multiple requests to confidential items while ensuring that all confidential information is secured.

Figure 2 shows the conceptual architecture of a multi-agent system for managing confidential information.

- **Supervisor Agent** A supervisor agent helps a human supervisor monitor attempts to access confidential information, assign initial information to a user's project, and control the work-flow of authorization of confidential information.

- **User-Interface Agent** An employee is only allowed to interact with a user-interface agent. This agent places a person's requests for confidential information with the authorization agent and allows a user to browse authorized information.
- **Authorization Agent** An authorization agent is responsible for deciding if requests for the confidential information are consistent with the employee's task. Section 5 details the process of adaptive authorization.

An agent space in this architecture replaces interactions between humans with communications between software agents. Specifically, a supervisor agent assigns initial information of an employee's projects to an authorization agent. An authorization agent learns profiles of projects for that employee. A user-interface agent requests pieces of confidential information on behalf of the employee. The employee will be able to browse the requested information if his request is authorized.

5 Adaptive Authorization

Needless to say, the most important goal of our system is to provide a highly reliable authorization for all requests for confidential information. We tackle this problem as a classification problem. Specifically, the request for access to a unit of confidential information is accepted only if the content of the requested item is relevant to the requester's task. It is reasonable to verify whether or not the content of a requested confidential item is associated with the content of requester's project because the requester only needs to know information related to his project in order to conduct the given task.

The authorization agent first learns the content of what each registered user needs to know from a small amount of confidential information, which has been assigned by a human supervisor. Note that we assume that all confidential information has a textual description and/or has textual content. When a request for confidential information occurs, the authorization agent compares the content of the requested information to the description of the requester's project. The request is approved for access if the requester's project description is determined to be "relevant" to the requested item. Rather than considering the match between a requester's access level and the requested item's ACL, our authorization is processed based on the comparison of similarity between the requester's project description and the content of the requested information.

An unauthorized attempt to access a unit of confidential information that the requester does not "need to know" is undoubtedly rejected because the requester's project description is not at all similar to that information. Moreover, this approach is quite flexible and adaptive to changes of project assignment in that only an updated description of newly assigned projects is necessary, instead of re-compiling the ACL on all changing relevant information. Therefore, it is much less expensive, in terms of computation and human time and effort, than an ACL-based or role-based approach.

However, there is an underlying assumption that there must be some amount of initial information which allows the authorization agent to learn the content of the project. In other words, a supervisor – or an employee who rightfully has full access to all confidential items – should obtain a certain amount of information and assign it to our agent. Since the anticipated context of use for the employee will involve from tens to a few hundreds of documents, it would be highly undesirable to ask a supervisor to assign 100 or so documents to the authorization agent for a given task. However, the approach could be realistic if the number of documents is small enough (e.g., 20 or 30 documents) such that a supervisor would not feel it to be overly time-consuming and intractable.

5.1 Classification Methods for Adaptive Authorization

In this section, we describe text classification methods for an authorization process. Since there are a large number of classification methods available, it is desirable to list guidelines to help us handpick an appropriate classification method. First, we should keep in mind that only a small amount of initial data for training classifiers will be available, because no supervisor is willing to spend time reading and providing large amounts of data to the authorization agent. Given this criterion, we need to choose a classification method that is able to provide reasonable classification performance given only a small amount of training data. Second, we should choose a classification method that can be easily trained computationally while requiring a relatively short training period, due to the fact that changes of project assignment happen often. No one would be willing to wait to access confidential information until the authorization agent finishes learning the newly assigned project. Finally, there should be a straightforward way of improving the performance of a selected classification method when it makes an intolerable number of mistakes.

To this end, we initially chose 5 different classification methods. Largely, they are comprised of three representative classification schemes: discriminative, generative, and transductive.

A discriminative classification assumes that there is a specific parameterized functional form (e.g., $\mathbf{w}$) to be optimized, and then the values of the parameters (e.g., components in $\mathbf{w}$) are determined from a given training data set by means of a suitable learning algorithm (e.g., mean square error) [Bishop, 1995]. For this category, we made use of three classifiers: Widrow-Hoff (WH), Exponentiated Gradient (EG) [Lewis et al., 1996], and a multi-layer neural network [Bishop, 1995]. In this scheme, a request for a confidential item, $\mathbf{x}$, on project, j, $(req(\mathbf{x}, j))$ is approved if $req(\mathbf{x}, j) = \mathbf{w}^T_{accept}\mathbf{x} > \mathbf{w}^T_{rejct}\mathbf{x}$, where $\mathbf{x} =< x_1, x_2, ..., x_{|\mathbf{x}|} >$ is a requested confidential text document in a multi-dimensional vector and $\mathbf{w}^T_{accept}$ is a weight vector obtained from training a discriminative classifier with the positive example of project j. Otherwise, it will be rejected.

A generative classification method assumes that there are a number of probability distributions (usually, the same number of classes) and that the task of the classification method is to assign the most probable class to an instance. Naive Bayes classification [Mitchell, 1997] is chosen for this category. A request for a confidential item, $\mathbf{x}$, on the jth project, $(req(\mathbf{x}, j))$ is accepted if $\frac{p(\mathbf{x}|C_{accept})P(C_{accept})}{p(\mathbf{x}|C_{reject})P(C_{reject})} > 0$, where $p(\mathbf{x}|C_{accept})$ is a likelihood of probability density function given class "accept" and $P(C_{accept})$ is a prior probability of class "accept."

Finally, a transductive classification method does not make any classification decisions until an instance is given. Given an instance to be classified, a transductive classification investigates all training examples to identify the k most similar examples and makes a decision based on their class labels. We chose k-nearest neighbors [Mitchell, 1997] for this category. A request for a confidential item, $\mathbf{x}$, on the project, j, $(req(\mathbf{x}, j))$ is accepted if $vote(C_{accept}) > vote(C_{reject})$, where $vote(C_{accept})$ is the number of training examples that belong to class "accept," otherwise, it is rejected.

The reason that we chose a multi-layer neural network and a naive Bayes classifiers is that we would like to see how two performance-proven and computationally expensive classifiers work in comparison with the other classifiers.

6 Experiments

A number of text classification experiments were performed, in order to find an appropriate classification method for adaptive authorization. Concisely, our task is a binary classification – to dichotomize a given text document into either positive or negative classes, where each text document is regarded as a piece of confidential information. If the document is labeled as "positive" for the user and their tasks, then a a request to access it will be granted.

6.1 Experimental Settings

Since there are no datasets available that are comprised of confidential information, especially text documents, we compiled two textual datasets from publicly available sources. The first textual dataset is a set of web documents downloaded from the Google directory [1]. The other one is comprised of three selected categories from the Reuters-21578 dataset [2]. Table 1 shows the distribution numbers of text documents in the two text datasets.

In order to make experiments more realistic, a relatively small number of data is used for training 5 classifiers. In particular, we separate two datasets into 20% training data and 80% testing data. In particular, for the Google dataset, 40 out of 200 web documents are used as training data. Specifically,

[1] http://directory.google.com

[2] It is publicly available at http://www.daviddlewis.com/resources/testcollections/-reuters21578/

Categories	Arts/Classical Studies	Arts/Literature	Society/Government	Total
# of web documents	68	68	64	200
Categories	Cocoa	Corn	Dlr	Total
# of documents	76	254	224	554

Table 1. The distribution of two text datasets. The top two rows are about the Google dataset and the bottom two rows are about the Reuters-21578 dataset. The "/" symbol in a category name of the Google dataset represents the hierarchical structure of the Google directory. For instance, "Arts" is a parent category that includes "Classical Studies" as one of sub categories.

the training data is 34 web documents that are comprised of 14 web documents from "Arts/Classical Studies", 14 from "Arts/Literature," and 12 from "Society/Government." For the Reuters dataset, 111 out of 554 Reuters documents are used for training. It is comprised of 30 text documents from the "cocoa" category, 40 from the "corn" category, and 41 from the "dlr" category. Note, we did not follow any recommended ways of separating the Reuters-21578 dataset, such as "ModApte" or "ModLewis." For the authorization scenario, "Arts/Classical studies" in the Google data is labeled as the "positive" class and the two remaining classes, "Arts/Literature" and "Society/Government", are labeled as the "negative" class. For the Reuters dataset, the "cocoa" category is used as the positive class and the remaining two categories are negative.

Text documents used for training are represented in three different methods: binary, raw-frequency, and TF·IDF. Since we are looking for a training-time bounded text classification for adaptive authorization, it is desirable to measure all possible costs of a text classification. By estimating the ratio of time to performance, we will be able to find a cost-effective text representation.

To this end, we first identified a vocabulary set (V) from the text documents assigned to training data after removing stop-words[3] and a set of words infrequently and highly frequently observed in the training data according to Zipf's Law. Each training text document is then projected onto a "word-by-document" matrix [Salton, 1989], where each row represents a word (i.e., unigram) in the identified vocabulary and each column represents a text document belonging to training data.

For the raw-frequency representation, the value of a cell (e.g., ith row and jth column) in the matrix is the frequency of the ith word from the identified vocabulary in the jth text document in the training data. We normalized each cell value by using a row's standard deviation and mean value. This avoids a scaling problem, created when the frequency of a word is much more variable than the others causing it to dominate any computation, e.g., similarity compu-

[3] A list of stop words is defined as common functional words such as "and", "of", "or", "the", etc., which are irrelevant for the representation of text content. [Salton, 1989]

output \ target	Accept	Reject
Accept	a	b
Reject	c	d

Table 2. A contingency table for evaluating a binary classification. In the table, the column a is the number of text documents that match a method's outputs and target values, b and c are the number of items that mismatch a method's outputs and target values respectively [Manning and Schutze, 1999].

tation. For the binary representation, if a word occurs one or more times, it is represented by "1," otherwise "0." In the TF·IDF text representation, the value of each cell is computed as follows:

$$w_{i,j} = \frac{1 + log_2 tf_{i,j} \times log_2(N/n_j)}{||\mathbf{d}||}$$

where $w_{i,j}$ is the weight value of the ith word in the jth document from the identified vocabulary set (V), $tf_{i,j}$ is the frequency of the ith word in the document j, N is the total number of training data, n_i is the number of training documents where word i occurs, and $||\mathbf{d}|| = \sqrt{\sum_w w_{i,j}}$.

We evaluate the performance of a text classification in five different measures that are defined using the contingency table in the Table 2.

- Precision, $p = \frac{a}{a+b}$
- Recall, $r = \frac{a}{a+c}$
- F1, $f1 = \frac{2pr}{p+r} = \frac{2a}{2a+b+c}$
- Miss, $m = \frac{c}{a+c}$
- False Alarm, $f = \frac{b}{b+d}$

Note all denominators should be greater than 0, otherwise they are not defined.

All performance metrics are important in terms of various evaluation perspectives. For example, precision can tell us how accurate a method's classification decisions are, relative to the total number of classification decisions it has made. For our scenario, the error rate is a more important measure than the others because we are more interested in having a classification method that makes fewer authorization errors. To this end, we elaborate the error in terms of "miss" and "false alarm." The miss measures the error rate of the number of text documents that are determined not to be authorized, which should be allowed to access, whereas the false alarm measures the error rate of the number of text documents that are allowed to access that should not be allowed access. In the case of a high miss rate, an employee might sense that the authorization process is not working properly. However, not approving valid requests does not cause a serious problem from the security perspective. Conversely, a high false alarm rate is a serious problem because confidential information, which should not be revealed, can be accessed. Therefore, it is more tolerable to have an authorization process with a

	A	P	R	F1	F	M	P	R	F1	F	M	T
WH	.743	.933	.259	.405	.009	.47	.724	.99	.836	.74	.009	37
EG	.668	.507	.629	.561	.311	.37	.784	.688	.733	.37	.311	37
*k*NN (5)	.737	.666	.444	.533	.113	.555	.758	.886	.817	.555	.113	11
NN	.656	0	0	0	.009	1	.66	.99	.792	1	.009	437
WH	.743	1	.24	.388	0	.759	.721	1	.837	.759	0	35
EG	.687	.547	.425	.479	.179	.574	.737	.82	.776	.574	.179	35
*k*NN (15)	.662	0	0	0	0	1	0.662	1	.796	1	0	10
NN	.656	0	0	0	.009	1	.66	.99	.792	1	.009	451
WH	.75	.937	.277	.428	.009	.722	.729	.99	.84	.722	.009	24
EG	.6	.446	.777	.567	.49	.222	.818	.509	.627	.222	.49	29
*k*NN (25)	.668	1	.018	.036	0	.981	.666	1	.8	.981	0	9
NN	.656	0	0	0	.009	1	.66	.99	.792	1	.009	440
Naive Bayes	.481	.389	.944	.551	.75	.055	.896	.245	.385	.05	.754	12

Table 3. Classification experiments with the Google dataset were evaluated by six different performance metrics. "A" is "accuracy," "P" is "precision," "R" is "recall," "F" is "false alarm," "M" is "miss," and "T" is the elapsed time for training a classifier in seconds. The number in parentheses next to *k*NN is the size of *k*. The five columns from third to seventh are the results for the positive class and the other five consecutive columns are the results for the negative class. Since we represented text documents in the given training data in three different text representation methods, three different results are presented from the top to to bottom at intervals of three rows: raw, binary, and TF·IDF, respectively. There is only one result from the naive Bayes classifier because it is a probabilistic classifier and hence used a different representation. Specifically, a text document is modeled as a collection of word probabilities and each word probability is used to generate a multinomial probability distribution.

high miss rate than one with high false alarm rate. Therefore, the method with a higher precision rate and a lower false alarm score is the preferable classification method.

6.2 Experimental Results

In order to provide a general sense of performance, we made use of the "accuracy" metric, which is a ratio of the number of documents correctly classified to the number of testing documents. For example, in Table 3, WH has .743 accuracy when the training documents are represented in the raw-frequency and this value comes from the fact that WH classified 119 out of 160 web-documents correctly (i.e., classified positive examples as positive and negative examples as negative). For *k*NN classification, three different sizes of *k* are chosen as the best performers in each of the text representation methods from the different trials.

Considering that relatively good performance by linear classifiers such as WH and EG was observed, the Google dataset seems linearly separable. In other

	A	P	R	F1	F	M	P	R	F1	F	M	T
WH	.954	.933	.608	.736	.005	0.391	.956	.994	.975	.391	.005	22
EG	.724	.086	.173	.115	.211	.826	.891	.788	.836	.826	.211	13
*k*NN (5)	.963	.941	.695	.8	.005	.304	.965	.994	.98	.304	.005	5
NN	.896	0	0	0	0	1	.896	1	.945	1	0	1111
WH	.948	.925	.543	.684	.005	.456	.949	.994	.971	.456	.005	19
EG	.74	.08	.152	.108	.191	.847	.891	.808	.848	.847	.191	22
*k*NN (15)	.902	1	.065	.122	0	.934	.902	1	.948	.934	0	5
NN	.896	0	0	0	0	1	.896	1	.945	1	0	1082
WH	.952	.931	.586	.72	.005	.413	.954	.994	.974	.413	.005	11
EG	.534	.152	.76	.253	.491	.239	.948	.508	.662	.239	.491	15
*k*NN (25)	.966	1	.673	.805	0	.326	.963	1	.981	.326	0	4
NN	.896	0	0	0	0	1	.896	1	.945	1	0	946
Naive Bayes	.582	.199	1	.332	.465	0	1	.534	.696	0	.465	4

Table 4. The results from text classification with three selected categories from the Reuters-21578 dataset.

words, the boundary between categories on Google is relatively clear and accordingly linear classifications worked well. The reason we collected web documents from two categories (i.e., "Classical Studies" and "Literature") under the same top category ("Arts") is to make the boundary between those two selected categories vague and accordingly to make the binary classification difficult. With the same reason, we chose two Reuters categories (i.e., "cocoa" and "corn") under the same conceptual category, "commodities."

A multi-layered neural network trained by back propagation was used [Mitchell, 1997]. It is a non-linear classifier that should be good at identifying the boundary between two classes – at least, better than the linear classifiers. The performance of the multi-layer neural network, however, was unusually poor. It was only capable of classifying negative examples.

WH showed the best classification performance for the Google dataset in terms of three criteria: a higher precision, a lower error rate (i.e., measured in miss and false alarms), and a small elapsed time. For example, its .009 false alarm rate can tell us that 9 out of 1000 requests for confidential information could be incorrectly authorized. This false alarm rate is verifiable if a user attempts indiscriminate access to a large number of documents. It is possible to further reduce the risk of inappropriate access to "false alarm material" by defining and enforcing a policy that would alarm or penalize the end user for attempting to access such information.

We hypothesized that the fastest way to represent text documents would be a raw-frequency representation because it leaves the word-by-document matrix intact. However, it turned out that there was a negligible difference in elapsed time between representation methods. Representing word-by-document matrix in raw-frequency is highly undesirable because it is not able to capture the characteristics of vocabulary distribution. TF·IDF representation, however, is desirable because it enables all selected words to have roughly the same impact on classification.

It is not surprising to see that the performance of the naive Bayes classifier was the worst because it usually suffers from the data variance problem. In other words, its performance monotonically increases to converge at a level of good performance only if it is given a sufficient amount of data.

kNN is the best performer for the Reuters dataset in that it showed the highest precision and the lowest false alarm. The reason that kNN was quite competitive is that it is very robust to noisy training data. However, its miss rate is quite high, meaning that personnel might feel that the automatic authorization is not working well, thus making it more likely for them to resort to a grievance process. Moreover, there is one clear drawback of using kNN: the cost of classifying new instances could be high because it is in proportion to the number of documents used for comparison.

7 Discussions

In this paper, we presented the architecture of a multi-agent system and text classification algorithms for the adaptive authorization of confidential information. In contrast to conventional systems based on coarse grained ACLs and hand-selected documents, our system provides content-based authorization. In particular, an unauthorized attempt to access confidential information that the requester does not "need to know" is rejected because the requester's project description is not similar to the information. In addition, this method is quite flexible and adaptive to changes of project assignment. Instead of re-compiling the ACL on all changing relevant information, only an updated description of newly assigned projects is necessary. Therefore, it is much less expensive, computationally, and in terms of human time and effort, than an ACL-based approach.

This work is significant in that it enables a human supervisor to conveniently and cost-effectively identify arbitrary subsets of confidential information and to associate security policies to it. The multi-agent system, by integrating with a secure document management system, enables the automatic enforcement of such security policies, as well as tracks authorized and unauthorized attempts to access the confidential information.

As for the adaptive authorization process, we tested 5 different text classification methods. We made use of two textual datasets and set up a configuration of experiments in order to reflect the most plausible scenario in which such classifiers could be used.

From several different experimental results, we found that WH and kNN are the best candidates for an adaptive authorization process. However, as cumulative learning is necessary for the continued training of a selected classification method, WH is preferred over kNN, because kNN is not a trainable classifier.

As future work, we would like to investigate the usefulness of relevance feedback for cumulative learning. In addition, we would like to investigate how an approach to cost-sensitive learning would be beneficial to minimize the authorization errors such as “miss” and “false alarm.”

8 Acknowledgement

We would like to thank Robin Glinton and Michael Rectenwald for their friendly and responsive help and comments. This research has been supported by ARDA under the CTA, Incorporated, sub-contract, BAA 03-02-CMU.

References

[Bishop, 1995] Bishop, C.M.: *Neural Networks for Pattern Recognition*, Oxford University Press (1995)

[Chandramouli and Sandhu, 1998] Chandramouli, R. and Sandhu, R.: Role based access control features in commercial database management systems, In *Proceedings of 21st National Information Systems Security*, (1998)

[Ferraiolo et al., 1995] Ferraiolo, D.F., Cugini, J., and Kuhn, D.R.: Role Based Access Control: Features and Motivations, In *Proceedings of Computer Security Applications Conference*, (1995)

[Giuri and Iglio, 1997] Giuri, L. and Iglio, P.: Role templates for content-based access control, In *Proceedings of ACM Workshop on Role Based Access Control*, pp. 153-159, (1997)

[Golbeck et al., 2003] Golbeck, J., Parsia, B., and Hendler, J.: Trust Networks on the Semantic Web, In *Proceeding of Cooperative Information Agents* (CIA-2003), pp. 238–249, (2003)

[Gollmann, 1999] Gollmann, D.: *Computer Security*, John Wiley and Sons, Inc., (1999)

[Lewis et al., 1996] Lewis, D.D., Schapire, R.E., Callan, J.P., and Papka, R.: Training algorithms for linear text classifiers, In *Proceedings of International ACM Conference on Research and Development in Information Retrieval* (SIGIR-96), pp. 298–306, (1996)

[Manning and Schutze, 1999] Manning, C.D. and Schutze, H.: *Foundations of Statistical Natural Language Processing*, MIT Press (1999)

[Mitchell, 1997] Mitchell, T.: *Machine Learning.* Prentice Hall (1997)

[Mouratidis et al., 2003] Mouratidis, H., Giorgini, P., and Manson, G.: Modelling secure multiagent systems, In *Proceedings of the Second International Joint Conference on Autonomous Agents and Multiagent Systems* (AAMAS-2003), pp. 859–866, (2003)

[Salton, 1989] Salton, G.: *Automatic Text Processing: The Transformation, Analysis, and Retrieval of Information by Computer.* Addison-Wesley (1989)

[Sycara, 1998] Sycara, K.: Multiagent systems, *AI Magazine*, 10:2, pp. 79–83, (1998)

[Tan et al., 2003] Tan, J.J., Poslad, S., and Titkov, L.: Agent driven policy management for securing open services, In *Proceedings of the Second International Joint Conference on Autonomous Agents and Multiagent Systems* (AAMAS-2003), pp. 1134–1135, (2003)

[Wagner, 1997] Wagner, G.: Multi-Level Security in Multiagent Systems, In *Proceedings of Cooperative Information Agents* (CIA-97), pp.272–285, (1997)

[Weippl and Ibrahim, 2001] Weippl, E. and Ibrahim, K.: Content-based management of document access control, In *Proceedings of the 14th International Conference on Applications of Prolog*, (2001)