Risk Assessment Based Access Control with Text and Behavior Analysis for Document Management

Zhuo Lu
Department of Electrical Engineering
University of South Florida, Tampa FL 33620
Email: z.l.lu@ieee.org

Yalin Sagduyu
Intelligent Automation Inc.
Rockville MD, 20855
Email: ysagduyu@i-a-i.com

***Abstract*—In computerized systems, documents with sensitive information are generated, stored and accessed every day in large volumes. These documents are classified and disseminated only to appropriate personnel. Unintentional disclosure of sensitive information should be ultimately avoided. How to effectively provide access control of document disclosure is a key for secure business, government and military operations. Traditional access control is based on a *simple* rule, i.e., to test whether a user account that requests the access to information has been granted such an access. However, this design has been shown to provide no security guarantee due to emerging incidents including insider threats, account hacking, and human classification errors. In this paper, we propose a new access control mechanism based on a *flexible* decision design, which will not simply guarantee access to a document when a user account has been granted such an access, but comprehensively use text analysis and behavior analysis in a complementary way to quantify the risk of information disclosure and grant the access only if the risk is assessed low. Our evaluation based on *notional* documents demonstrates the effectiveness of this new access control design against erroneous document classification and malicious user behavior. The proposed access control mechanism shows potential to enhance the overall security in today's access control systems for document management.**

I. Introduction

Documents with sensitive information for business, government, and military operations must be classified and disseminated only to appropriate personnel. Unintentional disclosure of or malicious access to such sensitive information can lead to significant unwanted effects. This is particularly important for data management and access control in the government and military information systems to protect data privacy, mission confidentiality and national security.

Government operations routinely maintain classified or sensitive documents in computer systems. Access to these documents must follow strict rules and be assigned to the personnel with proper authorization. For example, paragraphs or documents can be classified into four security classification levels: Unclassified (U), Confidential (C), Secret (S) and Top Secret (TS) [1]. A staff member is granted a security clearance level [2] such that he/she can have access to documents with the same (or lower) security classification level. In this way, the classified information will be properly disseminated to the personnel with appropriate access.

DISTRIBUTION A. Approved for public release; distribution unlimited. (AFRL/RIE; 88ABW-2016-2042).

Traditional access control is to make a binary "hard" decision based on the user's role and defined policies. For example, when a user account requests the access to information, the access will be granted only if the user is granted such an access by a policy [3]–[5] or the user's role has the access [6], [7]. However, in recent years, there have been many occasions that such an access control design failed to protect sensitive information due to various threats, such as insider threats, account hacking, and human classification errors. For example, a malicious insider user with some granted access may attempt to download a large amount of secret information then send the data to the adversary.

In this paper, we are motivated to design a new access control mechanism to protect sensitive information from unintentional or malicious access and disclosure. The idea behind our design is that instead of simply granting document access when a user has been granted the access, we should analyze the user's behavior based on the underlying textual content of the documents that the user requests, then assess the risk of document disclosure to such a user. Access will be denied or at least logged with notification if the risk is assessed high. Such a new mechanism can be integrated into an existing access control system to enhance the security and robustness for sensitive document management.

In our evaluation, we generated *notional* documents with different security classification levels to test the effectiveness of such a new access control design and showed that our design outperforms conventional access control design in the presence of insider threats and classification errors. The proposed access control mechanism has a wide range of applications in protecting sensitive information in business, government and military operations.

The reminder of this paper is organized as follows. In Section II, we introduce the models and state our research problem. In Section III, we present the design of the proposed access control, including the risk assessment and user analysis methods. In Section IV, we describe our security evaluation results. Finally, we conclude this paper in Section V.

II. Models and Problem Statement

We consider a document management system in which access control is enforced. The system stores a set of N documents $\{D_k\}_{k\in[1,N]}$. Each document must be classified

into a security classification level, which is in an ordered set $\mathbf{C} = \{C_k\}_{k\in[1,|\mathbf{C}|]}$. For example, $\mathbf{C}$ = {U, C, S, TS} with U $<$ C $<$ S $<$ TS for information control in many Department of Defense (DoD) operations.

We denote the classification process as a function, denoted as $\Phi(\cdot)$, mapping from a document to a security classification level. The classification can be done either manually or using a classifier, such as support vector machine (SVM) [8], [9], and Bayesian classifiers [10]–[12].

A user of the system is associated with an assigned security clearance level $C_u \in \mathbf{C}$. In traditional access control design, the user can access any document d as long as $\Phi(d) \leq C_u$, i.e., document d's classification level is lower than or equal to the user's security clearance level. However, as aforementioned, there are two major issues missing in the traditional design.

- $\Phi(\cdot)$ is a classification function by either human being or machine classifiers that can potentially make errors. There always exists a risk that a document with a high level (e.g., S) is classified to a low level (e.g., U), leading to erroneous *information disclosure*. Similarly, a document with a low level being classified to a high level may also happen, which leads to erroneous *information blocking*.
- Due to emergence of insider threats and account hacking, it may be suspicious for a user to request a large number of documents that the user does not regularly access.

To solve these two issues, we state our research problem in this paper as follows:

Definition 1: Give all N documents $\{D_k\}_{k\in[1,N]}$ stored in the system and a user's security clearance level $C_u \in \mathbf{C}$, determine whether to grant the user's current request to access documents $\{P_k\}_{k\in[1,L]}$, where L is the number of currently requested documents.

III. Risk Analysis based Access Control

A. General Architecture

The architecture of our access control design is shown in Fig. 1. When a user with a particular clearance level $C_u \in \mathbf{C}$ requests access to documents $\{P_k\}_{k\in[1,L]}$, the request will go to the risk analysis based access control, which consists of the following two major modules.

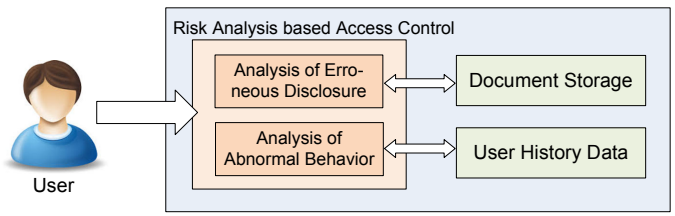

Fig. 1. Architecture of the proposed access control mechanism.

- *Analysis of Erroneous Disclosure (AED).* The AED module will interact with document storage to measure the similarity in textual features between the requested documents and other documents with the same and higher security levels to determine the risk of the request documents being disclosed to the user with clearance level C_u.
- *Analysis of Abnormal Behavior (AAB).* The AAB module will interact with the user history data storage to measure the risk of the user exhibiting abnormal behavior, which may indicate insider threats or account hacking.

If the risk from either AED or AAB modules is greater than a threshold, access will be denied. In the following, we will present the design of both AED and AAB modules in the proposed access control mechanism.

B. The Analysis of Erroneous Disclosure (AED) Module

When a user requests documents $\{P_k\}_{k\in[1,L]}$, which may have already been classified into some levels, denoted by $\{C_{P_k}\}_{k\in[1,L]}$, by security experts. The role of the AED module is to determine the risk of these documents being erroneously classified. Such a risk in fact contains two major factors:

1) *Risk due to classification errors/mismatch, called Type I risk.* Such a risk happens when there is a potential error in the document classification process by security experts. A statistical classification algorithm is needed to re-scan the document to ensure the correct access. However, machine classification may also lead to errors, which can lead to either *information leakage* (document with a high level is classified to a low level) or *information blocking* (document with a low level is classified to a high level). In practice, classification error is unavoidable in both human and machine classifications. Therefore, the AED module must access such a risk.
2) *Risk due to information similarity, called Type II risk.* Documents may share information. The risk increases if there is some shared information between two documents with different security classification levels; i.e., there may be a *"hidden" leakage* from a document with a low security level to the one with a high security level. As a result, the AED module will also evaluate Type II risk.

To this end, for a document P_{in} in $\{P_k\}_{k\in[1,L]}$, the AED module uses four processes to assess the overall risk, as shown in Fig. 2: 1) information processing and feature extraction process, in which texts and phrases are extracted and form a feature set $\mathcal{F}(P_{\text{in}})$, 2) Type I risk assessment, 3) Type II risk assessment, and 4) the overall risk assessment. In the following, we will discuss each process of AED module in detail.

1) Type I Risk Assessment: As shown in Fig. 2, after extracting $\mathcal{F}(P_{\text{in}})$ for document P_{in}, the Type I assessment component uses the Bayes classifier, denoted as $\Phi_B(\cdot)$, to compute the probability of the appearance of such features $\mathcal{F}(P_{\text{in}})$ in each classification level $s \in \mathbf{C}$:

$$\pi(s) = \sum_{f\in\mathcal{F}(P_{\text{in}})} \mathbb{P}(f|s). \tag{1}$$

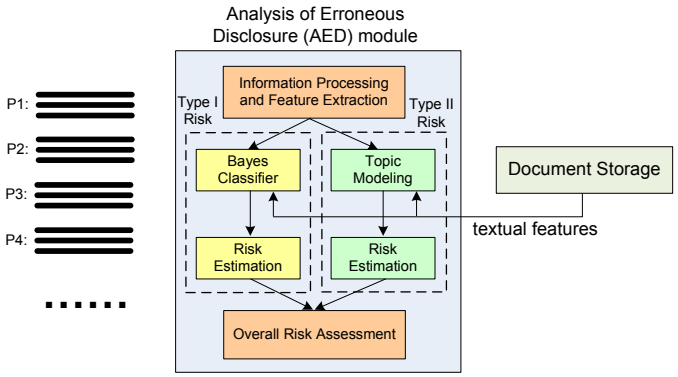

Fig. 2. AED module: measuring the risk the request documents being disclosed to the user.

The Bayes classifier can also give the best classification level $\hat{c} = \Phi_B(P_{\text{in}})$ that document P_{in} belongs to with the highest chance under its criterion. Note that the document $\mathcal{F}(P_{\text{in}})$ may have already been classified into a level c by security experts. So we may have two classified levels c and $\hat{c}$. If $c \neq \hat{c}$, what should happen? Our observation is that we should always trust human reasoning more than machine reasoning. Thus, when $c \neq \hat{c}$, the document still more likely belongs to classification level c. Therefore, we define Type I risk as the probability that P_{in} is not of level c. Ignoring $\hat{c}$ does not mean that the Bayes classifier is completely useless. We still use (1) in the Bayes classifier to compute Type I risk.

We define the two sub-types in Type I risk: (i) Information leakage in the risk. This happens when a document with a high security level is classified to a low level (e.g., TS level becomes U level). In this case, information is unintentionally disclosed. (ii) Information blocking in the risk. This happens when a document with a low security level is classified to a high level (e.g., S level becomes TS level). In this case, the content of a document is prevented from proper dissemination.

Given the extracted text features $\mathcal{F}(P_{\text{in}})$ from a document, we define the risk of information leakage as the probability that the document has already been classified into a level c should be classified to a higher level $s > c$, i.e.,

$$R_d(P_{\text{in}}) = \sum_{s>c} \pi(s) = \sum_{s>c} \mathbb{P}(f|s), \tag{2}$$

where $\pi(s)$ is given in (1).

Similarly, we define the risk of information blocking as the probability that the document has already been classified into a level c should be classified to a lower level $s < c$, i.e.,

$$R_b(P_{\text{in}}) = \sum_{s<c} \pi(s) = \sum_{s<c} \mathbb{P}(f|s). \tag{3}$$

Finally, Type I risk can be quantified as

$$R_I(P_{\text{in}}) = R_d + R_b = \sum_{s\neq c} \sum_{f\in\mathcal{F}(P_{\text{in}})} \mathbb{P}(f|s). \tag{4}$$

Note that up to this point, Type I risk only means the risk of a document being classified with a potential error. How to prevent a user from accessing such a document is assessed at the last process, the overall risk assessment, in the AED module, as shown in Fig. 2.

It also worth mentioning that inside the AED module, the two sub-types of risks $R_d(P_{\text{in}})$ and $R_b(P_{\text{in}})$ can be computed independently and provide interfaces for other modules for fine-grained and extended analysis.

2) Type II Risk Assessment: Type II risk assessment runs concurrently with Type I risk assessment. Type II risk may happen when there is some similar content between two documents with different classification levels; therefore, it is not due to classification error. In the following, we design a method based on topic modeling[1] [13]–[16] to measure the similarity of the textual content between documents.

By applying the topic modeling method in [15], namely Latent Dirichlet Allocation (LDA), we can obtain the probability distribution that a document is associated with a certain topic. In this formulation, topics act as an abstract later that connects word distributions with documents. Intuitively, if the topic distributions between two documents are similar (indicating they have similar content), the document may belong to the same level or close levels. On the other hand, different topic distributions between two documents carry no direct meaning in terms of the similarity between their security levels.

Therefore, a direct way to reveal the similarity between two documents is to use the cosine similarity, which has been used in data mining and information retrieval, as a measure of similarity. The cosine similarity, having value between 0 and 1, measures the cosine of the angle between two topic distribution vectors of two documents in an inner product space. For two documents P_1 and P_2, the cosine similarity is defined as

$$C(P_1, P_2) = \frac{\sum_{i=1}^{T} |X_1(i)X_2(i)|}{\sqrt{\sum_{i=1}^{T} X_1(i)^2}\sqrt{\sum_{i=1}^{T} X_2(i)^2}}, \tag{5}$$

where $X_1(i)$ and $X_2(i)$ are the i-th topic distributions (over T topics) of P_1 and P_2, respectively.

For the given document P_{in}, we define Type II risk as the average cosine similarity between the document and other documents with higher security classification levels as

$$R_{II}(P_{\text{in}}) = \frac{1}{|\mathbf{P}^*|} \sum_{p\in\mathbf{P}^*} C(P_{\text{in}}, p), \tag{6}$$

where $\mathbf{P}^*$ is the set of documents with higher security levels. To reduce the complexity, we use a random sample to form $\mathbf{P}^*$, instead of exhaustively using all other documents with higher security levels.

3) Overall Risk Assessment and Access Control: When Type I and Type II risks are computed, the AED module will decide whether to grant access of document P_{in} to the

[1]Topic modeling is a well-developed technique to analyze the statistical relations between textual structures. As we directly use an existing topic modeling method instead of developing new topic modeling techniques, we omit the detailed procedure of topic modeling, which is widely available in the literature

user with clearance level C_u. The following two rules apply towards information disclosure prevention.

1) *Backward Compatibility*. If the given security classification level c of document P_{in} is higher than C_u, immediately reject the access. This is because the system should never disclose a document to a user with lower clearance level (even if the document may be mistakenly classified as a high-level one). This is compatible with conventional access control design. In the case, the system immediately rejects the access, but logs all detailed AED risk evaluation for administrative analysis.
2) *Cautious Grant of Access*. If a user has enough clearance level to access P_{in}, we compute the overall risk as a weighted version of $R_I(P_{\text{in}})$ and $R_{II}(P_{\text{in}})$:

$$R_o(P_{\text{in}}) = w_1 R_I(P_{\text{in}}) + w_2 R_{II}(P_{\text{in}}), \tag{7}$$

where $w_1 \geq 0$, $w_2 \geq 0$ and $w_1 + w_2 = 1$. The access is granted only if $R_o(P_{\text{in}}) \leq T_{\text{AED}}$, where T_{AED} is a given threshold that represents the tolerant risk level. In this way, the AED module comprehensively takes into account the risks of a document being wrongly classified to prevent an erroneous document disclosure.

C. The Analysis of Abnormal Behavior (AAB) Module

The AED module can be effective in finding classification errors in some documents, and preventing erroneous access to such documents. However, it cannot prevent some other scenarios in which a user (e.g., an inside attacker or an account hacker) that has obtained a high security clearance level maliciously attempts to access a large number of documents that are indeed correctly classified.

The AAB module is used to detect and prevent such a malicious access. The process of the AAB module is shown in Fig. 3.

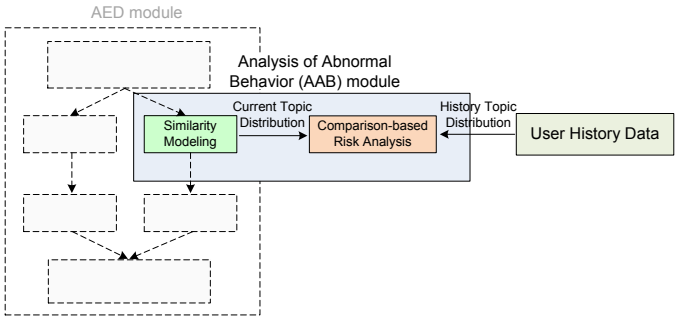

Fig. 3. AAB module: measuring the risk of the user exhibiting abnormal behavior.

The AAB module leverages the topic modeling component in the AED module, which computes the probability distribution that a document is associated with a certain topic as discussed in Section III-B2. This means that each time a user requests a document and is granted, the topic modeling probability distribution for that document can be stored as the user's history data, which gives the behavior pattern of what types (i.e., topics) of documents the user usually requests.

The intuition behind our design is as follows: every user has a routine job that needs access to various documents in an access controlled document management system. Due to the nature of a user's job, the documents that the user requests should have similar, if not the same, topics. When an account hacking happens, the hacker can use a user's account to download all possible documents that the account has the access to, which should exhibit quite distinct topic distributions analyzed by the topic modeling component in the AAB module as shown in Fig. 3.

As a result, when a user requests access of documents $\{P_k\}_{k\in[1,L]}$, we define the risk of abnormal behavior of such a user's action as the deviation between the averaged topic distributions of documents $\{P_k\}_{k\in[1,L]}$ between the user's historic topic distribution, i.e.,

$$R_a(\{P_k\}_{k\in[1,L]}) = \frac{1}{T}\sum_{i=1}^{T} \left| X_k(i) - U(i) \right|, \tag{8}$$

where $X_k(i)$ is the i-th topic distribution (over T topics) for document P_k, $U(i)$ is the i-th averaged topic distribution in the user's history data, and T is a given parameter in topic modeling.

Given (8), the AAB module currently uses one of the following rules for user access control.

1) *Strict rule* (strict rejection during abnormal behavior). A user's access requests will be rejected by the AAB module if $R_a(\{P_k\}_{k\in[1,L]}) > T_{\text{AAB}}$, where T_{AAB} is a given threshold in the AAB module.
2) *Loose rule* (notification during abnormal behavior). A user's access requests will always be granted by the AAB module, which will notify the administrator of a user's request and AAB analysis data if $R_a(\{P_k\}_{k\in[1,L]}) > T_{\text{AAB}}$.

Choosing the strict or loose rule depends on security requirements of a document management system.

IV. Evaluations

We generate documents to evaluate the effectiveness of our proposed access control system. In particular, we downloaded 100 documents from the Internet, treated them as *notional* documents, and manually classified them into four *notional* categories: A (lowest level), B, C, and D (highest level). The made-up rules of our manual classification are as follows (serving as the ground truth).

- If a document contains some National Football League topics (e.g., Superbow, Denver Broncos), the document will be classified into B.
- If a document discusses food or fruits topics, the document will be classified into C.
- if a document with the fruits topic contains words "apple" or "orange", the document will be classified into D.
- if a document with the food topic contains words "rice" or "beef", the document will be classified into D.
- All other documents will be classified into A.

During our classification, we intentionally left some errors, such as classifying a document containing word "apple" into C or B, to test the effectiveness of the proposed access control mechanism. As our access control mechanism consists of the AED and AAB modules that are relatively independent, we evaluate them individually.

A. The AED Evaluation

The AED module is to provide access control in presence of document classification errors. We simulate the behavior of a user with clearance level B and C to randomly access a number of documents in the system, some of which are manually classified into a wrong category. For example, the following text should belong to D (as it contains both "apple" and "orange"), but it was wrongly classified to be A.

- *Apple Granita: Simmer 4 cups apple juice with 1 cinnamon stick, 2 cloves and 1 strip orange zest, 10 minutes; strain and cool. Freeze in an 8-inch-square pan. To serve, scrape with a fork; top with minced green apple and candied ginger tossed with lemon juice.* (classified to A)

Each time, the simulated user randomly requests 20 documents within his clearance level. In other words, each document he requests is manually marked as a level equal to or below the user's clearance level. Fig. 4 shows an example of the overall risks computed in the AED module according to (7) for 7 documents that the user requests. In the 7 documents, documents 2 and 5 are D-level but intentionally misclassified to level A to test the effectiveness of the AED module. From Fig. 4, we can see that the overall risk assessment of documents 2 and 5 are higher than the threshold 0.5, and thus the AED module immediately rejects the requests of the two documents.

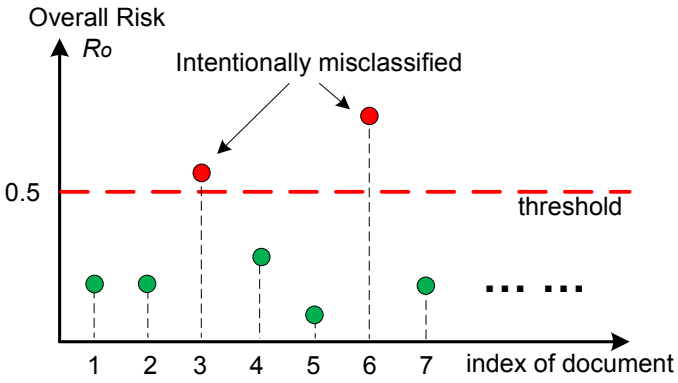

Fig. 4. Example of risk assessments of document requests.

The example in Fig. 4 shows the initial effectiveness of the AED module against erroneous disclosure due to classification errors. We then move on to comprehensively evaluate the performance of the AED module with a large number of document requests. We say the access control system makes a correct decision if the system correctly grants or rejects the access of a document to the user based on the ground truth (not based on the manual classification containing intentional errors). We define the correct decision rate as the rate of the number of correct decisions over the total number of document requests.

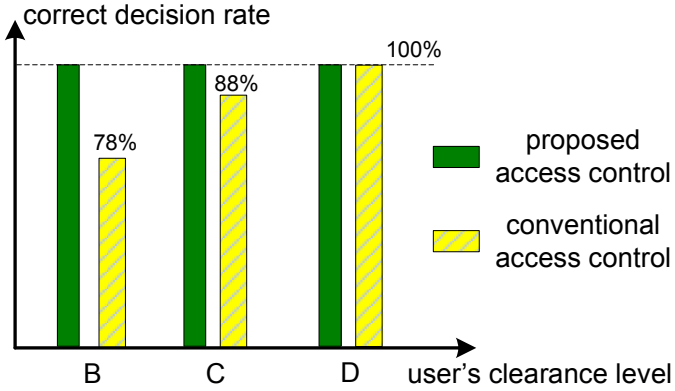

Fig. 5. Correct decision rate of AED versus conventional design.

Fig. 5 shows a comparison between the AED module and conventional design in which access will be granted as long as the user has enough clearance. It is noted from Fig. 5 that both AED and conventional design have 100% correct decision rate if the user has the D clearance (i.e., the highest clearance level), which means that the user can access everything. But if the user only has B or C clearance, we see that the conventional design will make some errors to disclose wrongly classified documents to users. However, AED does not make such errors since it does not absolutely trust the manual classification, and tries to analyze all probabilities and compute the risk of the document being wrongly classified and disclosed to reach a final decision.

During our simulation, we did not observe a case in which the overall risk of the access to a correctly classified document is assess higher than the threshold 0.5 in Fig. 5. In other words, there is no false alarm in the simulations with the AED module.

B. The AAB Evaluation

The AAB module is to detect abnormal user activities. In our AAB evaluation, we simulate the behavior of a user with C clearance level routinely accessing a number of food related documents for his jobs duty. We consider a scenario that the account of the user has been hacked, and the hacker requests a large amount of A, B, and C documents.

As shown in Fig. 6, the first 5 requests are made by the user for job duty and the 6th request is made by the hacker after account hacking to request all possible documents that can be accessed under the user's C clearance. The AAB module analyzes such behavior and obtains a very high risk of abnormal behavior according to (8), then immediately rejects such a request.

Fig. 6 illustrated that the AAB module shows its good potential to prevent unauthorized document disclosure due to insider threats and account hacking. We also simulate a comprehensive scenario where the hacker randomly requests

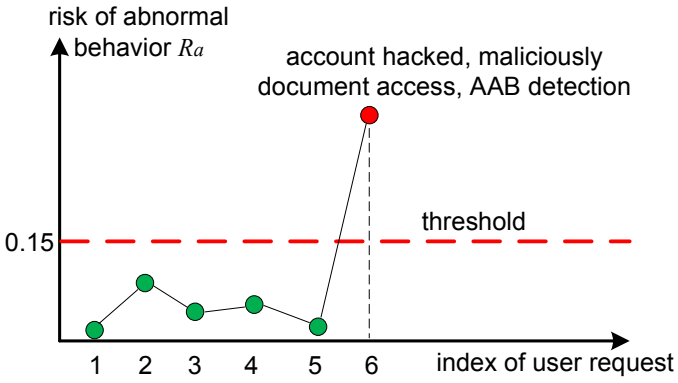

Fig. 6. Computations of the risk of abnormal behavior over time.

a large number of (but not all) documents in the system to evaluate the performance of the AAB module. Our performance metrics are 1) detection rate that represents the probability that the hacker's activity can be correctly detected, and 2) false alarm that represents the probability that a user's normal activity is mistakenly identified as malicious document request.

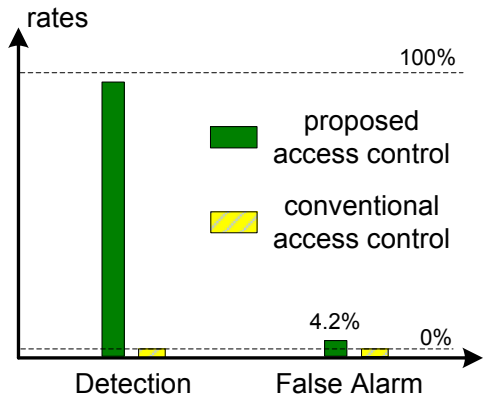

Fig. 7. Decision rate and false alarm of AED versus conventional design.

Fig. 7 shows the performance of the AAB module in comparison with conventional access control design. It is obvious that the conventional access control design always grants a user's access as long as the user has enough clearance, therefore does not provide any security against insider threats and account hacking. As a result, we observe from Fig. 7 that the conventional design has 0% detection rate and 0% false alarm in our account hacking scenario. In contrast, we can see from Fig. 7 that the AAB module has a detection rate of 96.5% and a false alarm of 4.2%. We find that the AAB module cannot achieve 100% detection ratio because the hacker requests a number of random documents, and there always exists a small probability that the requested documents are textually similar to the user's history data. There is also a positive false alarm shown in Fig. 7 because a user may occasionally request some documents, which show different textual structures and can be considered as outlier events.

Our simulations show that the AAB module is effective in detecting abnormal use activities, such as account hacking. On the other hand, we also need to be aware of the small probability of false alarm.

V. Conclusions

In this paper, we proposed an access control mechanism based on text analysis and behavior analysis to quantify the risk of information disclosure and grant the user access only if the risk is assessed low with respect to the user's credentials. Our proposed access control system consists of two relatively independent AED and AAB modules We used simulations to assess the performance in terms of correct decision rate of the AED module, detection rate and false alarm of abnormal user behavior analysis in the AAB module. Simulation results showed the effectiveness of our design against erroneous document classification and malicious user behavior.

Our future work includes comprehensive evaluation and parallelizations of the AED and AAB modules that will be developed for large-volume data processing in cloud computing based platforms.

References

[1] "Classificaiton Levels," *http://www.fas.org/sgp/library/quist2/chap_7.html*.
[2] "Security clearance frequently asked questions," *https://www.clearancejobs.com/security_clearance_faq.pdf*.
[3] P. W. Fong, "Relationship-based access control: protection model and policy language," in *Proc. of ACM Conference on Data and Application Security and Privacy*, 2011, pp. 191–202.
[4] P. Samarati and S. D. C. Di Vimercati, "Access control: Policies, models, and mechanisms," *Lecture Notes in Computer Science*, pp. 137–196, 2001.
[5] S. Jajodia, P. Samarati, M. L. Sapino, and V. Subrahmanian, "Flexible support for multiple access control policies," *ACM Trans. Database Systems*, vol. 26, no. 2, pp. 214–260, 2001.
[6] D. F. Ferraiolo, R. Sandhu, S. Gavrila, D. R. Kuhn, and R. Chandramouli, "Proposed nist standard for role-based access control," *ACM Trans. Information and System Security*, vol. 4, no. 3, pp. 224–274, 2001.
[7] J. B. Joshi, E. Bertino, U. Latif, and A. Ghafoor, "A generalized temporal role-based access control model," *IEEE Trans. Knowledge and Data Engineering*, vol. 17, no. 1, pp. 4–23, 2005.
[8] T. Joachims, "Transductive inference for text classification using support vector machines," in *Proc. of International Conference on Machine Learning*, vol. 99, 1999, pp. 200–209.
[9] C.-W. Hsu, C.-C. Chang, and C.-J. Lin, "A practical guide to support vector classification," 2003.
[10] H. Zhang, "The optimality of naive bayes," *Artificial Intelligence*, vol. 1, no. 2, p. 3, 2004.
[11] P. Domingos and M. Pazzani, "On the optimality of the simple bayesian classifier under zero-one loss," *Machine Learning*, vol. 29, pp. 103–137, 1997.
[12] A. Y. Ng and M. I. Jordan, "On discriminative vs. generative classifiers: a comparison of logistic regression and naive bayes," in *Proc. of Annual Conference on Neural Information Processing Systems*, 2001.
[13] A. Asuncion, P. Smyth, and M. Welling, "Asynchronous distributed learning of topic models," in *Proc. of Annual Conference on Neural Information Processing Systems*, 2008.
[14] M. D. Hoffman, D. M. Blei, and F. Bach, "Online learning for latent dirichlet allocation," in *Proc. of Annual Conference on Neural Information Processing Systems*, 2010.
[15] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," *Journal of Machine Learning Research*, vol. 3, pp. 993–1022, 2003.
[16] F. Yan, N. Xu, and Y. Qi, "Parallel inference for latent dirichlet allocation on graphics processing units," in *Proc. of Annual Conference on Neural Information Processing Systems*, 2009.