Improved Multilevel Security with Latent Semantic Indexing

Dirk Thorleuchter[1] and Dirk Van den Poel[2]

[1] Fraunhofer INT, Appelsgarten 2, 53879 Euskirchen, Germany
[2] Ghent University, Faculty of Economics and Business Administration, Tweekerkenstraat 2, 9000 Gent, Belgium

Abstract
Multilevel security (MLS) is specifically created to protect information from unauthorized access. In MLS, documents are assigned to a security label by a trusted subject e.g. an authorized user and based on this assignment; the access to documents is allowed or denied. Using a large number of security labels lead to a complex administration in MLS based operating systems. This is because the manual assignment of documents to a large number of security labels by an authorized user is time-consuming and error-prone. Thus in practice, most MLS based operating systems use a small number of security labels. However, information that is normally processed in an organization consists of different sensitivities and belongs to different compartments. To depict this information in MLS, a large number of security labels is necessary.
The aim of this paper is to show that the use of latent semantic indexing is successful in assigning textual information to security labels. This supports the authorized user by his manual assignment. It reduces complexity by the administration of a MLS based operating system and it enables the use of a large number of security labels. In future, the findings probably will lead to an increased usage of these MLS based operating systems in organizations.

Keywords: Information Security, Text Mining, Data protection, Decision support, Multilevel Security

1 Introduction

A well known problem in information security (Wright, 1998) is the unauthorized access, use, disclosure, disruption, modification, or destruction of information. The new challenge in protecting information is that current operating systems become more and more heterogenic and that these systems are connected in a complex way to further operating systems e.g. via

the internet (Moskovitch, Elovici, & Rokach, 2008). During the last decade, an increasing number of successfully processed security attacks per year can be seen (Bompard, Napoli, & Xue, 2009). It shows that today, operating systems cannot be protected against security exploits because of their inherent bugs and their vulnerabilities. This includes external attacks as well as internal attacks occurred by e.g. unwitting user behavior (Mazzariello, Lutiis, & Lombardo, 2011; Martinez-Moyano, Conrad, & Andersen, 2011; Chen, Chiu, & Chang, 2005).

The information flow of an organization is characterized by receiving information from different sources and by distributing information to different drains (Akella, Tang, & McMillin, 2010; Chou & Chen, 2006; Zhang, 2007). This information normally consists of different sensitivities concerning data protection and information security aspects (Ford, Millstein, Halpern-Felsher, & Irwin, 1996). As an example, strategic documents containing a company's strategy should not be distributed to customers or competitors. This specific information is more sensitive than information about a standard product that is published e.g. on the company's website. Some companies consider this aspect by assigning information to a self-defined security grading e.g. 'company confidential' (Dubash, 2011). Further, some governmental organizations assign information to an official security grading based on national or international agreements for the protection of data and classified information e.g. 'NATO secret' (Gericke et. al, 2009).

Besides assigning information to different security gradings, different compartments also have to be considered (Caroll, 1988). An example for this is that a company gets personal information from customers (e.g. name, birth date, earning, and credit card number). This personal information consists of different sensitivities e.g. a credit card number is more sensitive than a surname. However, this personal information also belongs to a different compartment than e.g. product or technological information. In general, people should not be able to access information in compartments they do not belong to regardless of whether the information is assigned to a security grading or not (Bell & LaPadula, 1976).

To ensure data protection and information security in environments with several security gradings and several compartments, a certain organization-wide access control policy is necessary (Ward, 2002; Pavlich-Mariscal, Demurjian, & Michel, 2010). It describes how to handle with different security gradings and with different compartments. Further, it describes how to determine access rights, e.g. to prevent people from obtaining access to sensitive information for which they lack authorization (Bell & LaPadula, 1973).

Operating systems that are based on multilevel security (MLS) use such a (mandatory) access control policy (Lunt, 1989; Shaikh, Adi, & Logrippo, 2012). They enable the processing of information without causing a security compromise. Thus, MLS protects an operating system from external security attacks as well as from internal security attacks. This includes unwitting user behavior by preventing users from obtaining access to information for which they are not authorized (Bell & LaPadula, 1976). Example for these specific operating systems are Secure VMS, BAE Systems XTS-400, Secure Versions of Windows Vista and Linux, Trusted Solaris, Compartmented Mode Workstation, etc. (Gericke et al., 2009).

A disadvantage of these operating systems is that the mandatory access control policy enforces a manual assignment of documents to a security grading and to one or several compartments by an authorized user (McLean, 1985; Obiedkov, Kourie, & Erloff, 2009). This results in a high complexity by the administration of these systems and in a low usability (Pfleeger and Pfleeger, 1990) especially by using many security gradings and many compartments. Therefore, these systems are rarely applied in practice (Holeman et al., 2002). They can be found in specific governmental environments e.g. in military (Von Solms & Geldenhuys, 1999) or in financial environments e.g. stock market transactions (Kjaerland, 2006) where security is much more important than administrational complexity.

As described above, a large number of security gradings and of compartments is necessary to depict the information flow of an organization in MLS. Thus in this paper, a high granular MLS approach as well as latent semantic indexing (LSI) as text classification methodology is introduced (see Sect. 2). A methodology is contributed in Sect. 3 and based on this, a case study (see Sect. 4) shows how LSI can be used to support the assigning of textual information to a security grading and to compartments in MLS.

As a result, it can be shown that text classification helps the authorized user to administrate an operating system based on the high granular MLS approach. This reduces administrative complexity by using a large number of security gradings and compartments. It also ensures data protection and information security for the information flow of an organization. Thus, the distinctive feature and main contribution of this paper is to propose a LSI approach in the field of MLS to make an increased usage of these MLS based operating systems possible for commercial and governmental organizations.

2 Background

In this chapter, the access control policies used in MLS are described and the predominant model for mandatory access control is introduced (see Sect. 2.1). Further, a high granular

MLS approach is formalized. It is used to separate information of different security gradings and different compartments as introduced through an example in Sect 2.2. Additionally, LSI as text classification methodology is introduced that enables the assignment of textual information to a security grading and to compartments (see Sect. 2.3).

2.1 Multilevel security

MLS describes the capability of a computer system or network to process information with different security gradings as well as with different compartments. It also prevents users from obtaining access to information for which they lack authorization (Bell & LaPadula, 1976). In MLS, information is stored in objects and each object is assigned to a security label. Additionally, a subject represents a user or an active entity - such as a process – in MLS. Each subject is assigned to a clearance that is also represented by a security label.

To implement MLS in an operating system, a mandatory access control policy (Lunt, 1989) normally is used. Access to objects is only allowed by a strict policy that is enforced by the system. Subjects cannot change neither this policy nor access rights for own objects. This mandatory model is in contrast to discretionary models where a subject - who is the owner of an object - is responsible for the allowing or denying other subjects access to this object. However in practice, the use of a discretionary model often leads to ad hoc decisions by individual users concerning access rights. Therefore, mandatory models are more reliable than discretionary models and the access to information is easier to control (Li, Du, & Wong, 2007).

A predominant model for mandatory access control is the Bell-LaPadula model (BLPM) (Bell and LaPadula, 1973) that is the formal security policy of the Trusted Computer System Evaluation Criteria (Orange Book) (Chokhani, 1992). The BLPM as well as its dual Biba model (Biba, 1977) uses security labels from a partially ordered universe that is named a lattice. An information flow is only allowed within the lattice. The partially order of the security labels determines the degree of object's security or subject's clearance. A security label consists of two categories: The security category (also known as security level or security grading) consists of a hierarchical structure e.g. top secret > secret > confidential > restricted etc. The needs-to-know category (also known as compartmented information) consists of specific restrictions e.g. US eyes only, personal compartment only, atomic, crypto etc.

Normally, a practical implementation and administration of MLS based operating systems with a medium to large number of security categories and of needs-to-know categories causes performance problems because of the linearly grows in the number of security

categories and the exponentially grows in the number of needs-to-know categories (Obiedkov, Kourie, & Erloff, 2009). For example, if a lattice contains five security categories and five needs-to-know categories then the number of security labels equals 160. This is calculated by the cardinality of the power set of five needs-to-know categories (32) multiplied by 5 security categories. Using a number of ten security categories and ten needs-to-know categories already leads to 10.240 security labels. Additionally, the number of objects in a system (e.g. data, files) could be very high. The (manual) assignment by an authorized user (a trusted subject) of a large number of objects to such a large number of security labels is time-consuming and error-prone. This explains why lattice-based access control models in practical use are restricted to a very small number of security labels (Pfleeger & Pfleeger, 1990; Holeman, Manimaron, Davis, & Chakrabarti, 2002).

In lattice-based access control policies that based on BLPM two fundamental precepts can be found. Firstly, it is not allowed to read up (Lindgreen & Herschberg, 1994). This means a subject cannot read information that is of a higher security category than the subject's clearance. For example, a subject (e.g. a program) with secret clearance must not read information classified as top secret. Additionally, subjects must not be able to read information in compartments to which they do not have access. Secondly, it is not allowed to write down (Anderson, Stajano, & Lee, 2002). Subjects on the system must not be able to write information in objects that are labeled by a lower security category then the security clearance level of the subject. For example, a subject with secret clearance must not be able to write secret information into a confidential object. To allow such an action would cause a security compromise that means an unauthorized access to information (Bell & LaPadula, 1973). Using these two precepts, a confidential subject must not be able to read from a secret object and a secret subject must not be able to write (secret) information in a confidential object. However, with these restrictions both subjects are not able to exchange information or to transfer a secret object e.g. from the secret level to the confidential level. Therefore, with BLPM it is not possible to depict the information flow in an organization where information is very often exchanged or transferred to persons with different security clearances as described in Sect. 1.

To bypass these restrictions, Gericke et al. (2009) propose the use of a more granular view on the information. This is because a secret object (e.g. a text) probably contains secret information as well as unclassified, restricted, or confidential information. If text phrases (e.g. a paragraph / a section) from a secret text are identified, which can be classified e.g. as confidential then these text phrases can be transferred to the confidential level without causing a security compromise. In Gericke et al. (2009) this requirement is satisfied by a

security gateway where the transferred information is monitored at the interfaces. The information is displayed on a viewer and manually analyzed by an authorized user. As a result, the authorized user has to identify text phrases in a text that can be assigned to a different security label than the text itself and if necessary, he has to change the labels of these text phrases manually.

This cannot be realized with a MLS approach that is based on BLPM where the classification of the object equals the highest security category and the union of all needs-to-know categories of the information stored in this object. The granularity of an object is not defined in BLPM (Lindgreen & Herschberg, 1994). In literature, examples can be found that identifies an object with a segment that may be a file or a multiple variable (Bell & LaPadula, 1976, Saltzer & Schroeder, 1975). However, most of the approaches identify an object as a file (Landwehr, 1981; McLean, 1985; Feiertag, Levitt, & Robinson, 1977).

A high granular MLS approach has been introduced by the authors (Thorleuchter & Van den Poel, 2011c; Thorleuchter, Weck, & Van den Poel 2012a; Thorleuchter, Weck, & Van den Poel 2012b) that focuses specifically on textual objects and uses an increased granularity (sections, paragraphs, text phrases, words, syllables, or signs). This approach contributes frame objects as a list of objects with different security categories and different sets of needs-to-know categories. In contrast to objects from BLPM, frame objects are not assigned to a security label that means each subject is allowed to access a frame object but subjects do not obtain access to a single object inside a frame object if they lack authorization. Therefore, a frame object creates texts or files that contain textual information from different security categories and needs-to-know categories.

The increased granularity of this manual approach enables workflows and the exchange and transfer of information. Further, it separates text patterns stored in objects according to their security labels. These text patterns can be used as training and test examples for text classification to enable an automatic assignment of texts to security labels.

2.2 Formal description of a high granular MLS architecture

The authors have introduced a high granular MLS approach (Thorleuchter & Van den Poel, 2011c; Thorleuchter, Weck, & Van den Poel 2012a; Thorleuchter, Weck, & Van den Poel 2012b). Below, the formulization is summarized in five aspects: frame object, object's categories, reading, writing, and deleting.

Definition 1 (Frame object): Let O{i,j} be an object (data, files, and programs of the MLS based operating system that are not in execution). Let O^{sup}_i be a frame object. $n \in N$ is defined as the number of frame objects in a MLS based operating system and $i \in \{1,..,n\}$. $m_i \in N$ is defined as the number of objects in O^{sup}_i and $j \in \{1,.., m_i\}$. Then, a frame object is defined as a list of objects by

$$O^{sup}_i \equiv [O\{i,1\}, .., O\{i,m_i\}] \quad (1)$$

Definition 2 (Categories): C is defined as a classification category (security level) and $C^{O\{i,j\}}$ is defined as the classification category of an object O{i,j}. K is defined as the needs-to-know categories and P is the power set. Then, $PK^{O\{i,j\}}$ represents the power set of all needs-to-know categories of an object O{i,j}. $Del^{O\{i,j\}} \in \{true, false\}$ is defined as the deleting category of an object O{i,j}. The categories of an object O{i,j} can be defined by

$$(C^{O\{i,j\}}, PK^{O\{i,j\}}, Del^{O\{i,j\}}) \quad (2)$$

Definition 3 (Reading): A subject S{k} is defined as process or program in execution. It consists of a security category (subject's security clearance) and of needs-to-know categories (the compartments to which a subject is authorized to access): $(C^{S\{k\}}, PK^{S\{k\}})$. $p \in N$ is defined as the number of subjects in the MLS based operating system and $k \in \{1, ..., p\}$. Subject S{k} is allowed to read in object O{i,j} if and only if

$$(C^{S\{k\}} \geq C^{O\{i,j\}}) \wedge (PK^{O\{i,j\}} \subseteq PK^{S\{k\}}) \wedge (Del^{O\{i,j\}} = false) \quad (3)$$

Definition 4 (Writing): Let an object $O\{i,j\} \equiv [data\{i,j,1\}, .., data\{i,j,q_{i,j}\}]$ be defined as a list of data units. Data units can be images, lines, sentences, text phrases, words, syllables, signs etc. $q_{i,j} \in N$ is the number of data units in an object O{i,j}. Let $l \in \{1, .., q_{i,j}\}$ be the position where a subject S{k} intents to insert content. Ow{i,j,l} is defined as a writing split on position l of an object O{i,j}. It consists of a list of three objects: $Ow\{i,j,l\} \equiv [Ow1\{i,j\}, Ow2\{i,j\}, Ow3\{i,j\}]$ with $Ow1\{i,j\} \equiv [data\{i,j,1\}, ..., data\{i,j,l-1\}]$ and $Ow3\{i,j\} \equiv [data\{i,j,l\}, \ldots, data\{i,j,q_{i,j}\}]$. Further, $Ow2\{i,j\} \in \varnothing$ is an empty object. Let the classification category $(C^{Ow1\{i,j\}} = C^{Ow3\{i,j\}} \equiv C^{O\{i,j\}})$ of object Ow1{i,j} and of object Ow3{i,j} be equal to the classification category of object O{i,j}. Let the needs-to-know categories $(PK^{Ow1\{i,j\}} = PK^{Ow3\{i,j\}} \equiv PK^{O\{i,j\}})$ of object Ow1{i,j} and of object Ow3{i,j} be equal to the needs-to-know categories of object O{i,j}. Let the deleting category $(Del^{Ow1\{i,j\}} = Del^{Ow3\{i,j\}} \equiv Del^{O\{i,j\}})$ of object Ow1{i,j} and of object Ow3{i,j} be equal to the deleting categories of object O{i,j}. Let the classification category $(C^{Ow2\{i,j\}} \equiv C^{S\{k\}})$ of object Ow2{i,j} be equal to the classification category of subject S{k}. Let the needs-to-know categories $(PK^{Ow2\{i,j\}} \equiv PK^{S\{k\}})$ of object Ow2{i,j} be equal to the needs-to-know categories of

subject S{k}. Then, subject S{k} is allowed to write in object Ow2{i,j} and thus, in O{i,j} if and only if

$Del^{O\{i,j\}} = false$ (4)

After writing, the corresponding frame object is defined as a list of objects where the object O{i,j} is replaced by the writing split object Ow{i,j,l}.

$O^{sup}_i \equiv [O\{i,1\}, \ldots, O\{i,j-1\}, Ow\{i,j,l\}, O\{i,j+1\}, \ldots, O\{i,m_i\}]$ (5)

Definition 5 (Deleting):

Subject S{k} intents to delete a list of data units [data{i,j,x}, ..., data{i,j,y}] $\subseteq$ O{i,j} from object O{i,j}. x $\in$ {1,…,$q_{i,j}$} is the start position, y $\in$ {1,…,$q_{i,j}$} is the end position, and y ≥ x. Od{i,j,x,y} is defined as a deleting split from position x to position y of an object O{i,j}. It consists of a list of three objects: Od{i,j,x,y} ≡ [Od1{i,j,x,y}, Od2{i,j,x,y}, Od3{i,j,x,y}] with Od1{i,j,x,y} ≡ [data{i,j,1}, ..., data{i,j,x-1}], Od2{i,j,x,y} ≡ [data{i,j,x}, ..., data{i,j,y}], and Od3{i,j,x,y} ≡ [data{i,j,y+1}, …, data{i,j,$q_{i,j}$}]. Let the security category ($C^{Od1\{i,j,x,y\}} = C^{Od2\{i,j,x,y\}} = C^{Od3\{i,j,x,y\}} \equiv C^{O\{i,j\}}$) of object Od1{i,j,x,y}, object Od2{i,j,x,y}, and of object Od3{i,j,x,y} be equal to the security category of object O{i,j}. Let the needs-to-know category ($PK^{Od1\{i,j,x,y\}} = PK^{Od2\{i,j,x,y\}} = PK^{Od3\{i,j,x,y\}} \equiv PK^{O\{i,j\}}$) of object Od1{i,j,x,y}, object Od2{i,j,x,y}, and of object Od3{i,j,x,y} be equal to the needs-to-know categories of object O{i,j}. Let the deleting category ($Del^{Od1\{i,j,x,y\}} = Del^{Od2\{i,j,x,y\}} = Del^{Od3\{i,j,x,y\}} \equiv Del^{O\{i,j\}}$) of object Od1{i,j,x,y}, object Od1{i,j,x,y}, and of object Od3{i,j,x,y} be equal to the deleting categories of object O{i,j}. Then, subject S{k} is allowed to delete content in object Od2{i,j,x,y} and thus, in O{i,j} if and only if

$(C^{O\{i,j\}} \leq C^{S\{k\}}) \wedge (PK^{O\{i,j\}} \subseteq PK^{S\{k\}}) \wedge (Del^{O\{i,j\}} = false)$ (6)

After deleting, the deleting category of object O{i,j} is set to true ($Del^{O\{i,j\}} \equiv true$) and the corresponding frame object is defined as a list of objects where the object O{i,j} is replaced by the deleting split object O_{del}\{i,j,x,y}}.

$O^{sup}_i \equiv [O\{i,1\}, \ldots, O\{i,j-1\}, O_{del}\{i,j,x,y\}, O\{i,j+1\}, \ldots, O\{i,m_i\}]$ (7)

The high granular MLS approach enables editing of documents by persons with different security clearances and it enables the exchange of information or the transfer to persons with different security clearances. To explain this, a simple example for the contribution of different sensitive information based on the granularity ‘sentence’ is presented below.

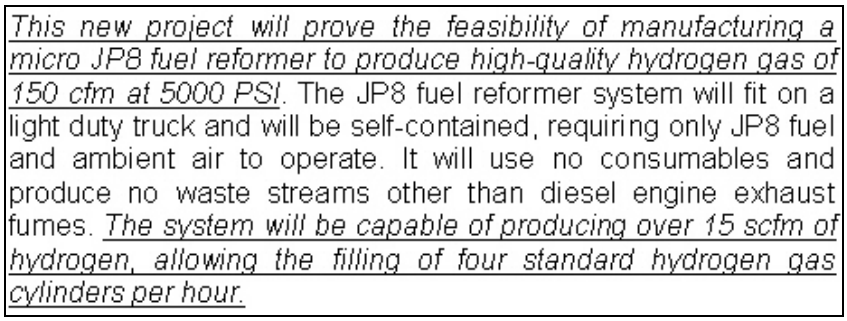

Figure 1: This text contains text phrases classified as company confidential (in underlined print) as well as unclassified text phrases. A person who has authorization to access on this classified information can see the text as presented.

People from research and development department of a company add information that is classified as company confidential (in underlined print) to a text that is unclassified (see Fig. 1). A person who has authorization for company confidential information can see the classified information as well as the unclassified information in the text. Using the standard BLPM, this text consists of one object that is classified as company confidential in total and only persons that have the corresponding clearance can access to the document. Using the high granular MLS approach, all persons can access this text but the text is presented to persons in different ways. The text is represented by a frame object O^{sup}_{i} that consists of three objects. O{i,1} contains the first sentence, O{i,2} contains the two sentence in the middle of the text, and O{i,3} contains the last sentence. The security category of O{i,1} and of O{i,3} is company confidential. The security category of O{i,2} is unclassified. A person with security clearance for company confidential is allowed to view all three objects (see Fig. 1). Fig. 2 shows how a person who lacks authorization to company confidential views the document because the person is only allowed to view the unclassified object O{i,2}.

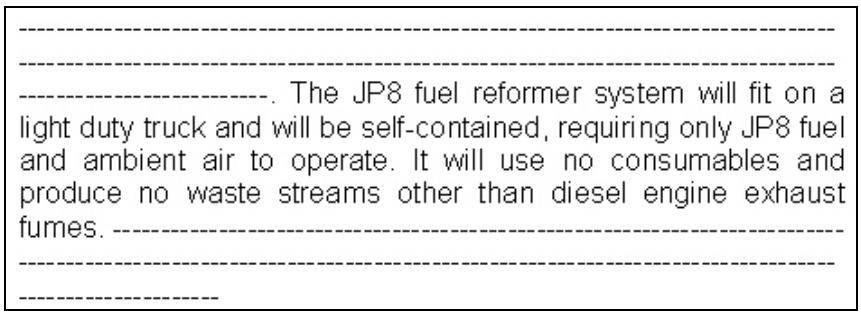

Figure 2: This is the same text with company confidential text phrases as shown in Fig. 1. It is presented in a way how a person without authorization (e.g. a marketing professional) sees the classified document. In this form, the information is released to publish e.g. in the internet. Then, information security compromises (e.g. by publishing company confidential information on company's website) do not occur because the high granular MLS approach prevents the marketing professional from obtaining access to company confidential information.

As an example, a marketing professional receives the order to publish the unclassified information from this text in the internet. If he lacks authorization for company confidential information, then the high granular MLS approach prevents him from obtaining access to company confidential information and therefore, he cannot cause a security compromise by publishing these sentences on company's website.

2.3 Text classification in MLS

Text classification aims at assigning pre-defined classes to text documents (Thorleuchter, Van den Poel, & Prinzie, 2010d; Ko & Seo, 2009). By applying text classification in MLS; a class can be defined as a security label. However, the use of a large number of security labels in MLS probably causes performance problems in text classification methodologies because of the large number of classes. Further, each class needs a minimum size of training examples to represent the class properly. If the number of classes is too large then a MLS based operating system probably does not contain enough objects associated with each specific security label.

Alternatively, each security category can be represented by a class and additionally, each needs-to-know category can be represented by a class, too. This reduces the number of classes in total. Additionally, the number of objects associated with a specific security category or with a specific needs-to-know category is much larger then the number of objects associated with both (a specific security label). Thus, classes can be represented more properly by using security categories and needs-to-know categories as classes.

Assigning an object to a security category or a needs-to-know category depends on aspects of meaning but not on aspects of words (Thorleuchter & Van den Poel, 2011b). A single word will never be classified e.g. as secret but the semantic meaning of several words that occur together in a text pattern probably will be (Thorleuchter, 2008). Thus for text classification in MLS, it is important to recognize that two objects share the aspect of meaning even if they do not share a single word. This can be done by identifying and comparing underlying dimensions of meaning from the objects (Park, Kim, Choi, & Kim, 2012). Then, an object can be assigned to the same class as a second object if their dimensions of meaning are similar.

A classification that considers aspects of meaning cannot be done by use of knowledge structure approaches e.g. Decision trees, support vector machine (SVM), naive Bayes classifier, k nearest neighbour (k-NN) classification (Shi & Setchi, 2012; Lee & Wang, 2012).

These approaches are frequently used in literature (Palmieri & Fiore, 2010; Herranza, Matwin, Nind, & Torra, 2010) however they do not identify the underlying dimensions.

The identification of underlying dimensions from the patterns of word usage in objects can be done by computational techniques that base on statistical procedures using some variation of eigenanalysis (eigenvectors) (Jiang, Berry, Donato, Ostrouchov, & Grady, 1999; Luo, Chen, & Xiong, 2011). It is important to know that the dimensions do not represent the words that are in an object but they represent the words that might be in the object (Thorleuchter, Van den Poel, & Prinzie, 2011; Thorleuchter & Van den Poel, 2012c). Thus, an object that is similar to a secret object can be classified as secret even if it consists of different words that are not equal to words in the secret object (Tsai, 2012; Christidis, Mentzas, & Apostolou, 2012). A well known technique of these eigensystems is LSI. After extracting a large number of underlying dimensions of meaning, LSI reduces the number of dimensions to get a manageable form (Thorleuchter, Van den Poel, & Prinzie, 2012).

3 Methodology

3.1 Overview

Sect 2.1 describes the problem of a MLS based operating system where the authorized user has to assign a large number of objects to a security category and to needs-to-know categories and thus, to a security label. The proposed methodology based on the high granular MLS approach as described in Sect. 2.2. It uses LSI as described in Sect. 2.3 to support the authorized user by his manual assignment.

This methodology uses textual information from objects stored in a high granular MLS based operating system. Since these operating systems do no exist; we use a self developed converter based on the formulization in Sect. 2.2 to emulate a high granular MLS based operating system. The converter uses documents in ‘edit mode’ as input information edited by different persons with different security categories and needs-to-know categories. Each document is split in several objects and each object consists of textual information that occur together in the corresponding document and that is assigned to the same security categories and to the same needs-to-know categories. As an example, the text in Fig. 1 is split in three objects. The first object contains the first sentence and the third object contains the last sentence where the security category is company confidential and a needs-to-know category is not given. The second object contains the two sentences in the middle of the text where

the security category is unclassified and a needs-to-know category is not given, too. Thus, each object is assigned to a specific security category and to a needs-to-know category (if available). They are divided in a training set and a test set and they are also pre-processed by use of text mining methods (see Sect. 3.2). A term-object matrix based on the training set is created that is used to identify the latent semantic patterns of the training objects. The test objects are projected into the same latent semantic concept-space (see Sect. 3.3). A logistic regression model is built on this concept-space matrix. It shows that test objects can be assigned successfully to a specific security category and to a needs-to-know category (see Sect. 3.4). The results are evaluated by comparing them to the frequent baseline and to results from SVM (Support Vector Machine) classification (see Sect. 3.5). Fig. 3 shows the methodology of this approach.

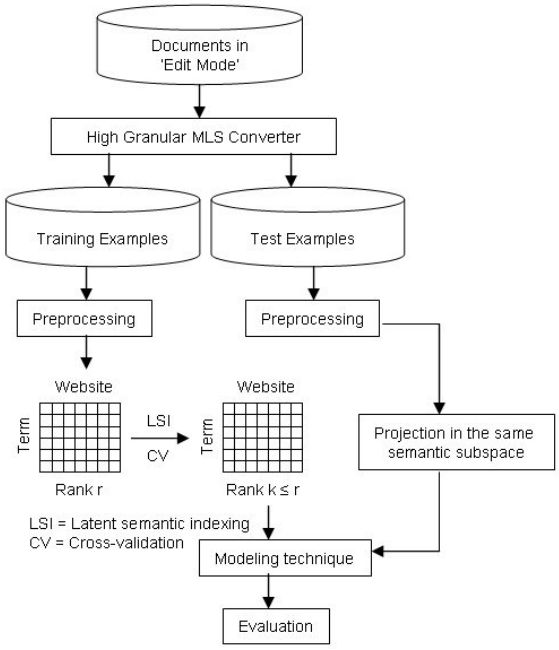

Figure 3: Methodology of the approach

3.2 Pre-processing

Objects created from the high granular MLS converter consist of textual information. Text preparation and term filtering is used to create a term vector in vector space model for each object. The size of the vectors is determined by the number of distinct terms in the collection of all objects. Weighted frequencies are used instead of raw term frequencies for the vectors' components. The term vectors are used to build a term-by-object matrix.

3.2.1 Text preparation

The raw text is cleaned in a first step by removing scripting code, tags, images etc. Further, specific characters are deleted. We use a dictionary to correct typographical errors (Thorleuchter, Van den Poel, & Prinzie, 2010a). In a second step, the text is tokenized that means it is split in terms where the term unit is word. All terms are converted in lower case and in a capitalized first sign.

3.2.2 Term filtering

We use term filtering to reduce the number of different terms (Thorleuchter, Van den Poel, & Prinzie, 2010b). Stop word filtering is done to identify terms that are non-informative (Thorleuchter & Van den Poel, 2011a). Further, part-of-speech tagging assigns terms to a syntactic category and terms that belong to a specific category also are non-informative (Thorleuchter, Van den Poel, & Prinzie, 2010c). Half of all terms in the collection of all objects appear only once or twice (Zipf, 1949; Zeng, Duan, Cao, & Wu, 2012). They are also non-informative terms. These non-informative terms are discarded. The number of different terms can be reduced further by term summarizing. Terms are converted to their stem by use of a dictionary or (if not in the dictionary) by use of a set of production rules as described by Porter (1980).

3.2.3 Term vector weighting

After text preparation and term filtering, a term vector in vector space model can be build. However, the use of weighted frequencies for the vectors' components leads to significant improvement (Sparck Jones, 1973). A component of a vector has a large weight if the corresponding term occurs frequently in a small number of objects but rarely in the collection of all objects (Salton & Buckley, 1988). We define $w_{i,j}$ as the weight that is assigned to term i in object j. Further, n is the number of objects and m is the number of terms in the vectors (m-dimensional term vectors). Then, df_i is the number of objects that contain term i (Chen, Chiu, & Chang, 2005). We calculate the weight by term frequency $tf_{i,j}$ times inverse object frequency idf_i divided by a length normalization factor (Salton, Allan, & Buckley, 1994).

$$w_{i,j} = \frac{tf_{i,j} \cdot \log(n / df_i)}{\sqrt{\sum_{p=1}^{m} tf_{i,j_p}{}^2 \cdot (\log(n / df_{i_p}))^2}} \quad (1)$$

3.3 Concept identification with LSI and singular value decomposition

The term-by-object matrix is of high dimensionality because it is based on the size of the term vectors as calculated by the number of distinct terms in the collection of all objects. Most values of the matrix components are zero thus, the high dimensionality can be reduced. This leads to a semantic generalization and helps to identify the underlying semantic textual patterns in the objects. We use LSI combined with singular value decomposition (SVD) as method for this reduction.

A is defined as the term-by-object (m x n) matrix and r is its rank ($r \leq \min(m,n)$). The SVD of A is a product of three matrices. The first matrix is the term-concept similarity (m x r) matrix U. The second matrix is the diagonal (r x r) matrix Σ containing positive singular values of matrix A. The third matrix is the concept-website similarity (n x r) matrix V.

$$A = U \Sigma V^t \quad (2)$$

We reduce the rank r of A to k by considering the first k ($k \leq r$) singular values in Σ. Further positive singular values are discarded. As the selection of k is critical for the predictive performance, we create several rank k-models on the training examples. The most favourable model is selected. Then, a prediction model is built (see Sect. 3.4) that integrates the test examples into the same semantic subspace (Deerwester, 1990).

3.4 Prediction Modelling

Logistic regression is used for prediction modelling based on the maximization of a maximum likelihood function (Allison, 1999). We use logistic regression because of its conceptual simplicity (DeLong, DeLong, & Clarke-Pearson, 1988) and its robustness concerning the predictive results (Greiff, 1998).

Let $T = \{(x_i, y_i)\}$ be a training set and let $i = \{1, ..., n\}$ be an index. Then, $x \in R^n$ is an n-dimensional input vector (a concept-object vector) as representative for the impact of objects on the concepts. Further, w is the parameter vector and w_0 is the intercept. $x_i \in R^n$ represents the input data and $y_i \in \{0,1\}$ represents the corresponding binary target labels (textual information from an object is assigned to a specific security label or not). Then, logistic regression estimates the probability $P(y = 1 \mid x)$ by

$$P(y = 1 \mid x) = \frac{1}{1 + exp(-(w_0 + wx))} \; . \tag{3}$$

3.5 Evaluation criteria

The aim of the prediction model is to show that latent semantic concepts from the objects can be used for an assignment of objects to a specific security label. We evaluate the prediction model to show its success. For this, well-known criteria are used: the cumulative lift, the precision and recall, the sensitivity and specificity, and the misclassification rate.

The most commonly used performance measure is the lift that measures the increase in density of objects that are successfully assigned to a security label relative to the density of objects that belong to the security label in total. Based on the objects that belong to the security label, TP (true positive) is defined as the number of correctly assigned objects and FN (false negative) is defined as the number of incorrectly assigned objects. Based on the objects that do not belong to the security label, TN (true negative) is defined as the number of correctly assigned objects and FP (false positive) is defined as the number of incorrectly assigned objects. The sensitivity (TP/(TP+FN)) is defined as the proportion of positive cases that are predicted to be positive, the specificity (TN/(TN + FP)) is defined as the proportion of negative cases that are predicted to be negative, the precision (TP/(TP+FP)) is defined as a measure of exactness or fidelity, and the recall (TP/(TP+TN)) is defined as a measure of completeness (Jones, 1972).

The receiver operation characteristics curve (ROC) (Van Erkel & Pattynama, 1998; Halpern et al., 1996) as a two dimensional plot of the sensitivity versus (1-specificity) is used to calculate the AUC (area under the ROC). The AUC is a well-known measure to compare the performance of binary classification models (Hanley and McNeil, 1982). For the calculation of the optimal number of concepts a cross-validated misclassification rate is used.

The frequent baseline as calculated from Table 1 is used as baseline for the evaluation. To compare the results to existing text classification approaches, a support vector machine (SVM) classifier is used, too. For each security label, a SVM (Palmieri & Fiore, 2010) separates objects that are assigned to this security label from objects that are not assigned to this security label in a training phase. Then, a hyperplane, which is located between the positive and negative training examples, is determined by a small number of training examples (support vectors) and the test examples are assigned based on this hyperplane.

4 Case study

4.1 Overview

The aim of the case study is to show that LSI can be used to support the authorized user by the administration of a high granular MLS based operating system that means by assigning objects to a security category or to needs-to-know categories. To select an application field for a case study, three conditions have to be considered: the information has to be available, subjects and objects have to be determined concerning their categories / clearances, and the number of test and training examples has to be sufficiently high for a statistical evaluation. All three conditions are fulfilled by selecting the application field ‘defense based research and technology’. German Ministry of Defence (GE MoD) provides the authors about 800 textual documents from 2000 to 2006 dealing about the planning of research and technology for GE MoD. The documents are available for the evaluation because they are not subjected to an official security grading based on national or international agreements for the protection of data and classified information.

Some textual parts of the documents are sensitive because they deal about technological areas that are of German national security interests. In these areas, the complete know-how for processing research and technology projects has to be nationally available. This excludes the processing of collaboration projects with further nations where the technological knowledge is split among the nations. National security interest is in contrast to the European or to international security interest where projects normally are processed in collaboration with further nations.

The documents are written by use of Microsoft Word and they are available in ‘Edit Mode’. Several persons from different departments of GE MoD as well as from several subsidiary departments have edited the documents. The structure of these departments is based on technological areas. Thus, departments can be assigned to technological areas related to national security interest or to European or to international security interest. Based on their affiliation, some authors of the documents are assigned to a security clearance for national security interest. Thus, in this case study we consider two security categories ‘national security interest’ and ‘unclassified’ and we do not consider needs-to-know categories. This reduces complexity and it ensures that the number of test and training examples is sufficiently high for an evaluation. Further, a successful binary LSI classification (a training object can be assigned to national security interest or not) shows the feasibility of the proposed approach.

Most documents contain several technologies and are edited by authors with the security clearance for national security interest as well as by authors without the security clearance.

The documents are not processed in a high granular MLS thus, a self-developed converter is used to emulate this considering the formulation of the high granular MSA (see Sect. 2.2). As a result, each document is split in many objects and the objects are assigned to a security category based on author's clearance.

After converting, the 800 documents are split into 4126 objects. These objects are randomly split into training and test set. The training set is used to obtain the optimal SVD dimension and the model estimates, while the test set is used to validate and compare the different models. The data characteristics are shown in Table 1.

	Number of objects	*Relative percentage*
Training set:		
National security interest	392	19
Unclassified	1671	81
Total	4126	
Test (incl. Validation) set:		
National security interest	392	19
Unclassified	1671	81
Total	4126	

Table 1: Characteristics of national security interest documents and unclassified objects

A LSI based binary classification model is created. For evaluation purpose, a SVM based binary classification model is created, too. Both models assign the objects to a security category 'national security interest' or otherwise to unclassified.

4.2 Optimal dimension selection and interpretation

The rank of the high dimensional term-by-object matrix is reduced to obtain the optimal number of SVD dimension (concepts). Thus, a cross-validation procedure on the training data was applied (Thorleuchter, Herberz, & Van den Poel, 2012; Thorleuchter & Van den Poel, 2012a). The x-axis in Fig. 4 represents the number of concepts and the y-axis represents the cross-validated misclassification rate. It can be seen that in the range of 1–20 concepts, the cross-validated misclassification rate was decreasing rapidly. From 20 concepts on, it was decreasing less rapidly, while in the region around 60 concepts, the cross-validated performance was stabilizing. Including more than 60 concepts resulted in a

more complex prediction model, while the misclassification rate hardly decreased. Thus, 60 concepts were chosen as the optimal number of SVD dimension in our study. At this point, a good balance was achieved between the number of concepts and the predictive performance.

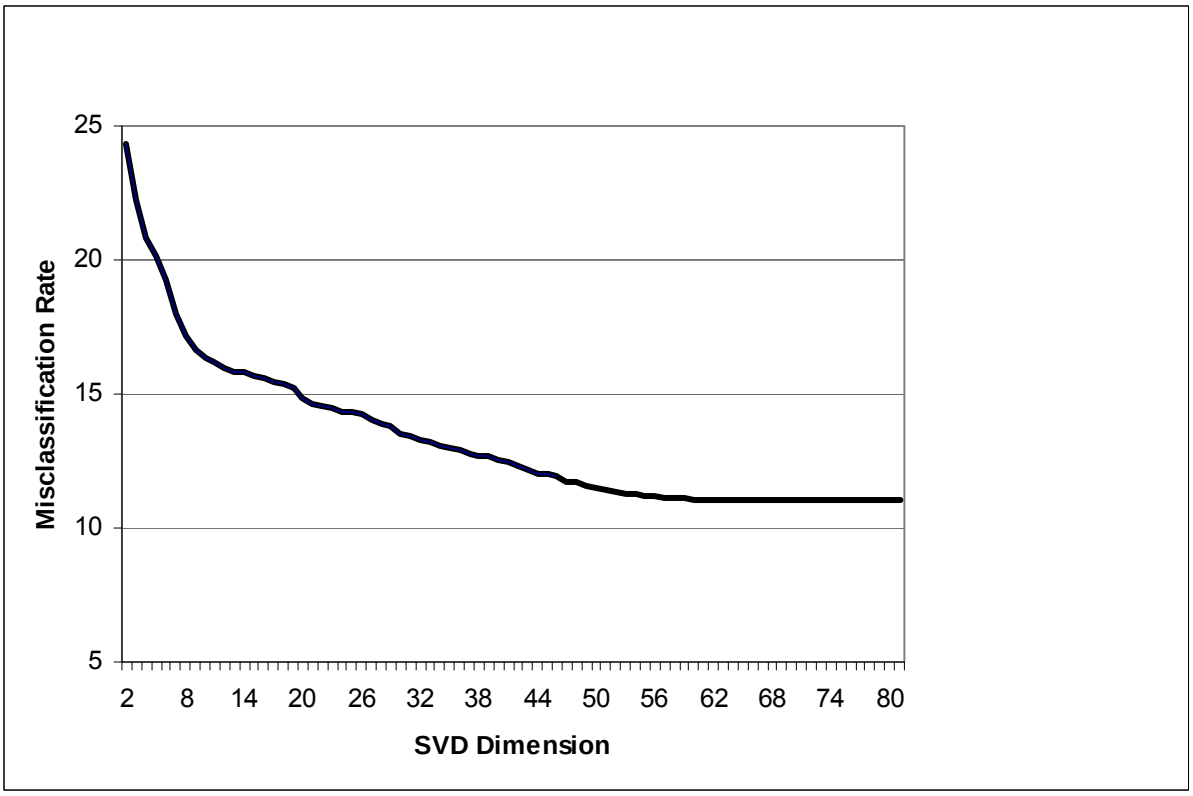

Figure 4: SVD Dimension

Each calculated SVD dimension represent the above-chance frequent occurrence of several terms in an object that can be used to assign this object to the security category national security interests. The terms represent words in German language and in stemmed form because a German stemmer is used as described in Sect. 3.2.2. We translate the terms to the English language to present examples for the interpretation of some single SVD dimensions.

A. To show the important results in detail, three groups of terms that are representative for the ‘national security interests’ objects are presented below:

A1. Electronic (including electronics, electronically etc.) and warfare are two terms that occur above-chance frequently in objects assigned to national security interests together with the following terms in stemmed form: optronic, intelligent, digital, receiver, millimeter, sub-millimeter, frequency, band, radar, infrared, combat, attack, protection, spectrum, energy,

etc. The terms describe the research area of electronic warfare where research projects normally are processed nationally.

A2. Intelligent and ammunition are two terms that occur above-chance frequently in objects assigned to national security interests together with the following terms in stemmed form: energetic, materials, weapon, effect, precision, guided, etc.. The terms describe the area of area of weapon and ammunition systems. Most of the research projects in this area are not processed in collaboration projects.

A3. Protection and decontaminate are two terms that occur above-chance frequently in objects assigned to national security interests together with the following terms in stemmed form: automatic, portable, real-time, diagnosis, nuclear, biological, chemical, biotechnology, etc. The terms represent the area of defense against nuclear, biological, and chemical threats. Based on a strategic decision of MoD, the know-how in this area has to be nationally available.

B. Furthermore, three groups of terms that are representative for the unclassified sections are presented below:

B1. Architectures and modeling are two terms that occur above-chance frequently in text patterns of the unclassified objects together with the following terms in stemmed form: data, information, fusion, simulation software, communication, environments, radio, encryption, etc. The terms describe the area of communications and simulation where most of the research projects are collaborative that means they are processed together with other nations and thus, they are unclassified.

B2. Unmanned and system are two terms that occur above-chance frequently in text patterns of the unclassified objects together with the following terms in stemmed form: armours, visibility, reduction, sense-and-avoid, unmanned, intelligent, multifunctional, materials, temperature, propulsion, fuel, cell, etc. The terms represent the area of platforms where mainly collaborative research is done and thus, the projects normally are unclassified.

B3. Personal and protection are two terms that occur above-chance frequently in text patterns of the unclassified sections together with the following terms in stemmed form: textiles, sensors, integrated, computing, communications, soldier, electric, energy, individual,

passive, management, human, factor, etc. The terms represent the area of soldier technologies where most of the research projects are unclassified, too.

4.3 Comparing predictive performance

Fig. 5 and 6 show that the predictive performance of the regression model (test set) significantly outperforms the baselines.

The cumulative lift curve of the test set lay above the SVM baseline and above the frequent baseline, respectively. Thus, LSI is able to identify more classified sections than the baselines within a specific percentile, e.g. the lift value in the top 30 percentile increases from 1 (frequent baseline) and 1.66 (SVM baseline) to 1.79 (test set). The ROC curve of the test set lay above the baselines and the ROC curve of the test set is located further from the frequent baseline than that of the SVM baseline. Thus, the AUC of test set (81.36) is larger than that of SVM baseline (72.42). This improvement is significant. This shows that the LSI model is able to better distinguish 'national security interests' objects from unclassified objects than the SVM model.

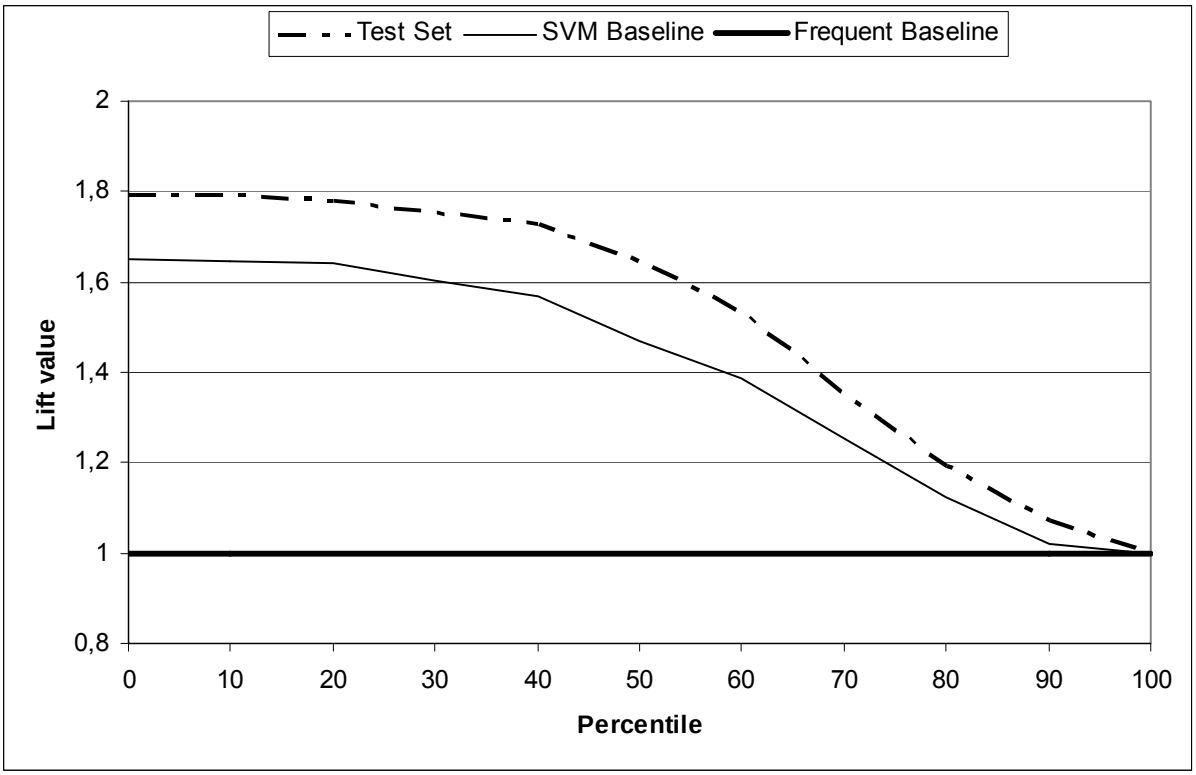

Figure 5: Lift for the logistic regression model

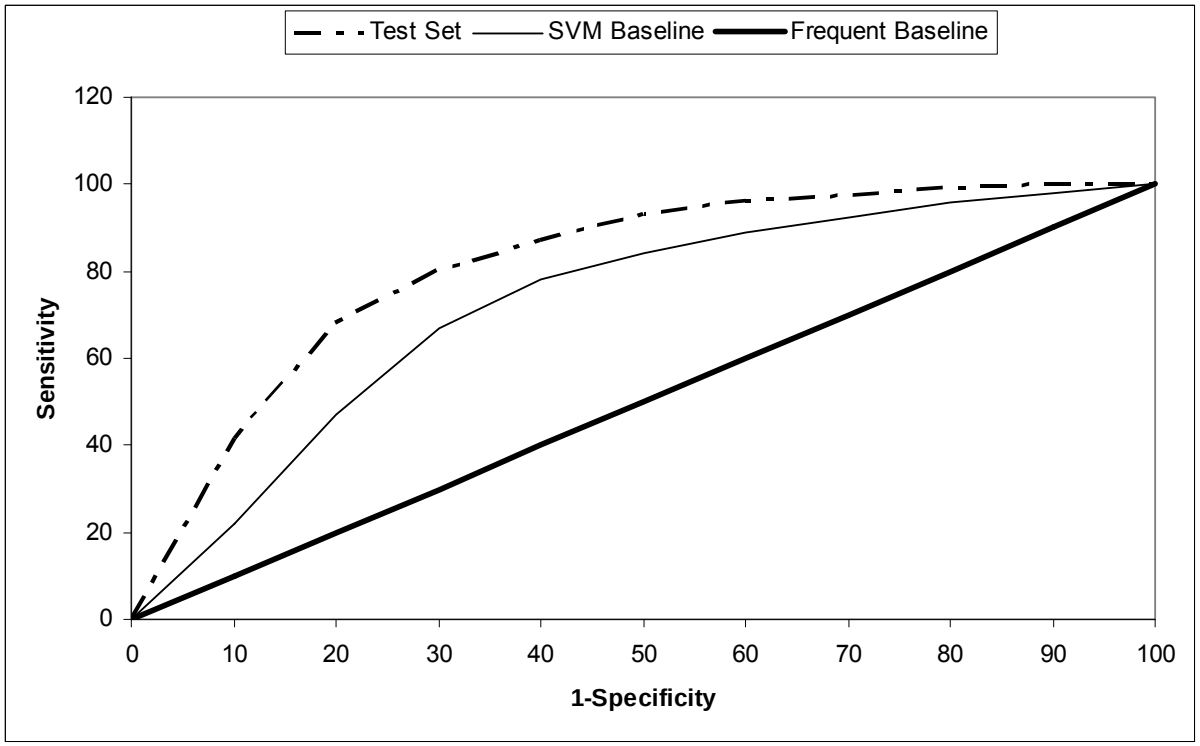

Figure 6: Sensitivity / Specificity Diagram

5 Summary and conclusions

In an organization, information is received from several sources and distributed to several drains. Concerning data protection and information security, this information consists of several different sensitivities (security categories) and compartments (needs-to-know categories). To depict this information flow in MLS operating systems, a large number of security labels is necessary where an authorized user has to assign information from objects to a large number of security labels. This leads to a high complex administration of MLS based operating system. Additionally, to realize an information exchange between subjects of different security clearances, a high granularity MLS approach is necessary. With this approach, text pattern that represent classified information can be separately stored in objects labeled with the corresponding security labels.

As text classification methodology, LSI is presented that classifies information concerning their aspects of meaning. LSI is used to recommend the assigning of information to a specific security label as decision support for the authorized user. In a case study, it is shown that a logistic regression model based on LSI is successful in this assigning.

As a result, LSI can be used to support the authorized user by assigning classified information to security labels. This support reduces the complexity to administrate a MLS based operating system. Further, a high granular MLS approach permits the information transfer to objects of different security labels. This is useful to depict the information flow of an organization in MLS where information is often exchanged. Therefore, these results will probably lead to an increased usage of MLS based operating systems in organizations.

The access of subjects to objects is controlled by the mandatory access control policy and each request is stored in log files in the MLS based operating system. Therefore, future research should focus on classifying the information from these log files to identify behavior issues of subjects. This could be a further example for the use of text classification in MLS.

6 Acknowledgments

This project was realized using SAS v9.1.3, SAS Text Miner v5.2, Web Genesis SVM Classifier, and Matlab v7.0.4.

References

Allison, P. D. (1999). *Logistic Regression using the SAS System: Theory and Application.* Cary: SAS Institute Inc.

Akella, R., Tang, H., & McMillin, B. M. (2010). Analysis of information flow security in cyber–physical systems. International Journal of Critical Infrastructure Protection, 3(3-4), 157-173.

Anderson, R., Stajano, F., & Lee, J. H. (2002). Security policies. Advances in Computers, 55, 185-235.

Bell, D. E., & LaPadula, L. J. (1973). Secure Computer Systems: Mathematical Foundations. Bedford: Mitre Corp.

Bell, D. E., & LaPadula, L. J. (1976). Secure Computer System: Unified Exposition and Multics Interpretation. Bedford: MITRE Corp.

Biba, K. J. (1977). Integrity Considerations for Secure Computer Systems. Bedford: Mitre Corp.

Bompard, E., Napoli, R., & Xue, F. (2009). Assessment of information impacts in power system security against malicious attacks in a general framework. Reliability Engineering & System Safety, 94(6), 1087-1094.

Carroll, J. M. (1988). Implementing multilevel security by violation privilege. Computers & Security, 7(6), 563-573.

Chen, M. C., Chiu, A. L., & Chang, H. H. (2005). Mining changes in customer behavior in retail marketing. *Expert System with Applications*, 28(4), 773-781.

Chokhani, S. (1992). Trusted products evaluation. *Communication of the ACM*, 35(7), 64-76.

Chou, S. C., & Chen, Y. C. (2006). Managing role relationships in an information flow control model. Journal of Systems and Software, 79(4), 507-522.

Christidis, K., Mentzas, G., & Apostolou, D. (2012). Using latent topics to enhance search and recommendation in Enterprise Social Software. Expert Systems with Applications, 39(10), 9297-9307.
Deerwester, S., Dumais, S., Furnas, G., Landauer, T., & Harshman, R. (1990). Indexing by latent semantic analysis. *Journal of the American Society for Information Science,* 41(6), 391-407.
DeLong, E. R., DeLong, D. M., & Clarke-Pearson, D. L. (1988). Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. *Biometrics,* 44(3), 837-845.
Dubash, M. (2011). Do you have the right security? Computer Fraud & Security, 2011(3), 14-16.
Ford, C. A., Millstein, S. G., Halpern-Felsher, B., & Irwin, C. E. (1996). Confidentiality and adolescents' disclosure of sensitive information. Journal of Adolescent Health, 18(2), 111.
Feiertag, R. J., Levitt, K. N., & Robinson, L. (1977). Providing multilevel security of a system design. Proceedings of the 6th ACM Symp. Operating System Principles, 1977.
Gericke, W., Thorleuchter, D., Weck, G., Reiländer F., & Loß, D. (2009). Vertrauliche Verarbeitung staatlich eingestufter Information - die Informationstechnologie im Geheimschutz. *Informatik Spektrum,* 32(2), 102-109.
Greiff, W. R. (1998). A theory of term weighting based on exploratory data analysis. In W. B. Croft, A. Moffat, C. J. van Rijsbergen, R. Wilkinson, & J. Zobel (Eds.), *Proceedings of the 21st SIGIR Conference* (pp. 11-19). New York: ACM.
Halpern, E. J., Albert, M., Krieger, A. M., Metz, C. E., & Maidment, A. D. (1996). Comparison of receiver operating characteristic curves on the basis of optimal operating points. Academic Radiology, 3(3), 245-253.
Hanley, J.A., & McNeil, B.J. (1982). The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology,* 143(1), 29-36.
Herranza, J., Matwin, S., Nind, J., & Torra, V. (2010). Classifying data from protected statistical datasets. Computer and Security, 29(8), 875-890.
Holeman, S., Manimaron, G., Davis, J., & Chakrabarti, A. (2002). Differentally secure multicasting and its implementation methods. Computer Security, 21(8), 736-749.
Jiang, J., Berry, M. W., Donato, J. M., Ostrouchov, G., & Grady, N. W. (1999). Mining consumer product data via latent semantic indexing. Intelligent Data Analysis, 3(5), 377-398.
Jones, K. S. (1972). A statistical interpretation of term specificity and its application in retrieval. *Journal of Documentation*, 28(1), 11-21.
Kjaerland, M. (2006). A taxonomy and comparison of computer security incidents from the commercial and government sectors. Computers & Security, 25(7), 522-538.
Ko, Y., & Seo, J. (2009). Text classification from unlabeled documents with bootstrapping and feature projection techniques. Information Processing & Management, 45(1), 70-83.
Landwehr, C.E. (1981). Formal models for computer security. Computing Surveys, 13(3).
Lee, C.H., & Wang S.H. (2012). An information fusion approach to integrate image annotation and text mining methods for geographic knowledge discovery. Expert Systems with Applications, 39(10), 8954-8967.
Li, E. Y., Du, T. C., & Wong, J. W. (2007). Access control in collaborative commerce. Decision Support Systems, 43, 675-685.

Lindgreen, R., Herschberg, I. S. (1994). On the validity of the Bell-LaPadula model. Journal of Computer Security, 13, 317-333.
Lunt, T. F. (1989). Access control policies: Some unanswered questions. Computers & Security, 8(1), 43-54.
Luo, Q., Chen, E., & Xiong, H. (2011). A semantic term weighting scheme for text categorization. Expert Systems with Applications, 38(10), 12708-12716.
Mazzariello, C., Lutiis, P. D., & Lombardo, D. (2011). Clustering NGN user behavior for anomaly detection. Information Security Technical Report, 16(1), 20-28.
Martinez-Moyano, I. J., Conrad, S. H., & Andersen, D. F. (2011). Modeling behavioral considerations related to information security. Computers & Security, 30(6–7), 397-409.
McLean J. (1985). A comment on the Basic Security Theorem of Bell and LaPadula. Information Process Letter, 20.
Moskovitch, R., Elovici, Y., & Rokach, L. (2008). Detection of unknown computer worms based on behavioral classification of the host. Computational Statistics & Data Analysis, 52(9), 4544-4566.
Obiedkov, S., Kourie, D. G., & Erloff, J. H. P. (2009). Buildung access control models with attribute exploration. Journal of Computer Security, 28, 2-7.
Palmieri F., & Fiore, U. (2010). Network anomaly detection through nonlinear analysis. Computer Security, 29, 737-55.
Park, D, H., Kim, H. K., Choi, I. Y., & Kim, J. K. (2012). A literature review and classification of recommender systems research. Expert Systems with Applications, 39(11), 10059-10072.
Pavlich-Mariscal, J. A., Demurjian, S. A., & Michel L. D. (2010). A framework of composable access control features: Preserving separation of access control concerns from models to code. Computers & Security, 29(3), 350-379.
Pfleeger, P., & Pfleeger, S. L. (2003). Security in computing. Old Tappan: Prentice Hall.
Porter, M.F. (1980). An algorithm for suffix stripping. Program, 14(3), 130-137.
Saltzer, J. H., & Schroeder, M. D., The protection of information in computer systems. Proceedings of the IEEE, 63(9), 1278-1308.
Salton, G., Allan, J., & Buckley, C. (1994). Automatic structuring and retrieval of large text files. *Communications of the ACM,* 37(2), 97-108.
Salton, G., & Buckley, C. (1988). Term-weighting approaches in automatic text retrieval. *Information Processing & Management,* 24(5), 513–523.
Shaikh, R. A., Adi, K., & Logrippo, L. (2012). Dynamic risk-based decision methods for access control systems. Computers & Security, In Press
Shi, L., & Setchi, R. (2012). User-oriented ontology-based clustering of stored memories. Expert Systems with Applications, 39(10), 9730-9742.
Sparck Jones, K. (1973). Index term weighting. Information Storage and Retrieval, 9(11), 619-633.
Thorleuchter, D. (2008). Finding Technological Ideas and Inventions with Text Mining and Technique Philosophy. In C. Preisach, H. Burkhardt, L. Schmidt-Thieme, & R. Decker (Eds.), *Data Analysis, Machine Learning, and Applications* (pp. 413-420). Berlin: Springer.
Thorleuchter, D., Van den Poel, D., & Prinzie, A. (2010a). Mining Ideas from Textual Information. *Expert Systems with Applications,* 37(10), 7182-7188.

Thorleuchter, D., Van den Poel, D., & Prinzie, A. (2010b). A compared R&D-based and patent-based cross impact analysis for identifying relationships between technologies. *Technological Forecasting and Social Change,* 77(7), 1037-1050.
Thorleuchter, D., Van den Poel, D., & Prinzie, A. (2010c). Extracting consumers needs for new products - A web mining approach. In *Proceedings WKDD 2010* (p. 441.). Los Alamitor: IEEE Computer Society.
Thorleuchter, D., Van den Poel, D., & Prinzie, A. (2010d). Mining innovative ideas to support new product research and development. In H. Locarek-Junge, & C. Weihs (Eds.),,*Classification as a Tool for Research,* Berlin: Springer.
Thorleuchter, D., Van den Poel, D., & Prinzie, A. (2011). LSI based Profitability Prediction of new Customers," In Proc. SIAM International Workshop on Data Mining for Marketing (pp. 62-67). New York: SIAM.
Thorleuchter, D., & Van den Poel, D. (2011a). Companies Website Optimising concerning Consumer's searching for new Products. In Proc. Uncertainty Reasoning and Knowledge Engineering (pp. 40-43). New York: IEEE.
Thorleuchter, D., & Van den Poel, D. (2011b). Semantic Technology Classification - A Defence and Security Case Study. In Proc. Uncertainty Reasoning and Knowledge Engineering (pp. 36-39), New York: IEEE.
Thorleuchter, D., & Van den Poel, D. (2011c) High Granular Multi-Level-Security Model for Improved Usability. In: System Science, Engineering Design and Manufacturing Informatization 1 (pp. 191-194), New York: IEEE.
Thorleuchter, D., Van den Poel, D., & Prinzie, A. (2012). Analyzing existing customers' websites to improve the customer acquisition process as well as the profitability prediction in B-to-B marketing. Expert Systems with Applications, 39(3), 2597-2605.
Thorleuchter, D., Herberz, S., & Van den Poel, D. (2012). Mining Social Behavior Ideas of Przewalski Horses. Lecture Notes in Electrical Engineering, 121, 649-656.
Thorleuchter, D., & Van den Poel, D. (2012a). Extraction of Ideas from Microsystems Technology. In: David Jin, Sally Lin (Eds.) Computer Science and Information Engineering. Advances in Intelligent and Soft Computing 168, Berlin: Springer, in press.
Thorleuchter, D., Weck, G., & Van den Poel, D. (2012a). Granular Deleting in Multi Level Security Models. In: Mechanical and Electronic Engineering. Lecture Notes in Electrical Engineering, Berlin: Springer, in press.
Thorleuchter, D., Weck, G., & Van den Poel, D. (2012b). Usability based Modeling for Advanced IT-Security. In: Mechanical and Electronic Engineering. Lecture Notes in Electrical Engineering, Berlin: Springer, in press.
Thorleuchter, D., & Van den Poel, D. (2012c). Using Webcrawling of Publicly-Available Websites to Assess E-Commerce Relationships, Service Research and Innovation Institute SRII Global Conference, New York: IEEE, in press.
Tsai, H.H. (2012). Global data mining: An empirical study of current trends, future forecasts and technology diffusions. Expert Systems with Applications, 39(9), 8172-8181.

Van Erkel, A. R., & Pattynama, P. M. T. (1998). Receiver operating characteristic (ROC) analysis: Basic principles and applications in radiology. European Journal of Radiology, 27(2), 88-94.
Von Solms, S.H., & Geldenhuys, J.H.S. (1999). Managing multi level security in a military intranet environment. Computers & Security, 18(3), 257-270.
Ward, P., & Smith, C. L. (2002). The Development of Access Control Policies for Information Technology Systems. Computers & Security, 21(4), 356-371.
Wright, M. A. (1998). The need for information security education. Computer Fraud & Security, 1998(8), 14-17.
Zeng, J., Duan, J., Cao, W., & Wu C. (2012). Topics modeling based on selective Zipf distribution. Expert Systems with Applications, 39(7), 6541-6546.
Zhang, G. Q. (2007). Mediating secure information flow policies. Information and Computation, 205(9), 1413-1425.
Zipf, G. K. (1949). *Human Behaviour and the Principle of Least Effort.* Cambridge: Addison-Wesley.