UNIVERSITEIT GENT

FACULTEIT ECONOMIE EN BEDRIJFSKUNDE

TWEEKERKENSTRAAT 2
B-9000 GENT
Tel. : 32 - (0)9 – 264.34.61
Fax. : 32 - (0)9 – 264.35.92

WORKING PAPER

Technology Classification with Latent Semantic Indexing

Dirk Thorleuchter[1]

Dirk Van den Poel[2]

September 2012

2012/814

[1] Fraunhofer INT, Appelsgarten 2, 53879 Euskirchen, Germany
[2] Ghent University, Faculty of Economics and Business Administration, Tweekerkenstraat 2, 9000 Gent, Belgium, http://www.crm.UGent.be

D/2012/7012/47

Technology Classification with Latent Semantic Indexing

Dirk Thorleuchter[a,*], Dirk Van den Poel[b]

[a] Fraunhofer INT, D-53879 Euskirchen, Appelsgarten 2, Germany, dirk.thorleuchter@int.fraunhofer.de
[b] Ghent University, Faculty of Economics and Business Administration, B-9000 Gent, Tweekerkenstraat 2, Belgium, dirk.vandenpoel@ugent.be URL: http://www.crm.UGent.be

* *Corresponding author at: Fraunhofer INT, Appelsgarten 2, 53879 Euskirchen, Germany. Tel.: +49 2251 18305; fax: +49 2251 18 38 305*
E-mail address: Dirk.Thorleuchter@int.fraunhofer.de (D. Thorleuchter).

Abstract
Many national and international governments establish organizations for applied science research funding. For this, several organizations have defined procedures for identifying relevant projects that based on prioritized technologies. Even for applied science research projects, which combine several technologies it is difficult to identify all corresponding technologies of all research-funding organizations. In this paper, we present an approach to support researchers and to support research-funding planners by classifying applied science research projects according to corresponding technologies of research-funding organizations. In contrast to related work, this problem is solved by considering results from literature concerning the application based technological relationships and by creating a new approach that is based on latent semantic indexing (LSI) as semantic text classification algorithm. Technologies that occur together in the process of creating an application are grouped in classes, semantic textual patterns are identified as representative for each class, and projects are assigned to one of these classes. This enables the assignment of each project to all technologies semantically grouped by use of LSI. This approach is evaluated using the example of defense and security based technological research. This is because the growing importance of this application field leads to an increasing number of research projects and to the appearance of many new technologies.

Key Words: Latent semantic indexing, SVD, Classification, Research Funding.

1 Introduction

Research funding for applied science research projects is done by many national and international organizations (Beaudry & Allaoui, 2012; Lepori, 2011). They evaluate proposals for new research projects and based on self-defined procedures, they identify the relevant projects, which are accepted for funding (Hicks, 2012; Mobjörk & Linnér, 2006). An important criterion for technological research is that the technologies standing behind the proposed research project are also mentioned in a specific list or taxonomy of prioritizes technologies (Choi, Lee, & Sohn, 2009; Bradshaw et al., 2008). In general, these technology lists or taxonomies consist of a manually created label for each technology and of a description. The descriptions contain terms from the technology as well as from potential application fields (Thorleuchter, Van den Poel, & Prinzie, 2010c). For example, the European Union establishes a Framework Research Programme (FP7) theme for security that has the objective to develop technologies needed to ensure the security of citizens from threats. It uses a list of prioritized technologies (ESRAB technology list) for research funding decisions (Remuss, 2010). That means proposals of research projects that do not fit with these prioritized technologies and the corresponding application field e.g. ‘security’ normally are not accepted (McLeish & Nightingale, 2007; Jiricka & Pröbstl, 2012).

For a researcher, it is often difficult to identify the corresponding prioritized technologies and corresponding application fields concerning each research-funding organization (Grimpe, 2012). Additionally, it is also difficult for research planners to assign applied science research projects to prioritized technologies of their research-funding organization manually (Ludwig, Roson, Zografos, & Kallis, 2011). Therefore, in this paper, we present an automated approach based on text classification that supports researchers as well as research-funding planners by the identification of relationships between applied science research projects and technologies extracted from lists or taxonomies.

Literature proposes application based technological relationships (Yu, Hurley, Kliebenstein, & Orazem, 2012). Here, it is shown that during the process of creating an application, technologies are related to their substitutive, integrative, predecessor, and successor technologies (Geschka, 1983). An example for substitutive technologies is electrical fuel cells, electrical batteries, and solar cells in the context of creating an energy supply application. A research project that has the aim to create a new approach for an energy supply application can combine all three substitutive technologies to build this new approach. Alternatively, it can focus on one technology e.g. fuel cells. However then, it has to consider research results from the further substitutive technologies. This is because the newly created fuel cell approach for energy supply has to be compared to existing potential energy supply applications to indicate its advances. This full cell project processes knowledge from electrical battery and solar cells and thus, is related to the electrical battery technology and to the solar cell technology, too; even if key words from electrical battery technology or from solar cell technology do not occur in the project description (Geschka, Lenk, & Vietor, 2002).

Applied science research projects have to combine or to consider these related technologies to create an application (Thorleuchter, Van den Poel, & Prinzie, 2010b). This describes a binary classification problem because the test examples (research projects) are associated with a specific class (a set of related technologies) (Kim, Toh, Teoh, Eng, & Yau, 2012). To identify related technologies, LSI is used. This is because semantically, all related technologies consist of the same terms describing the technology or the application field. LSI identifies the semantic textual patterns in the descriptions of the technologies and it also identifies the impact of each technology description on each semantic textual pattern (Thorleuchter & Van den Poel, 2012b). Then, each semantic pattern represents a set of related technologies where the corresponding impact is larger than a specific threshold. The descriptions of the projects are projected in the same semantic subspace. An assignment of each project on a set of technologies can be done based on the calculated impact of each project on each semantic textual pattern (Thorleuchter, Van den Poel, & Prinzie, 2012).

Previous work calculates the similarity between each project and each technology separately assuming that all technologies are independent (Thorleuchter & Van den Poel, 2011). It uses machine-learning techniques as supervised learning methods and a knowledge structure text classification approach that uses a similarity measure (Jaccard's coefficient) as well as a specific threshold to enable a multi-label classification. This knowledge structure approach often fails because prevalent features that are characteristic for a technology are not simultaneously present in all projects that belong to one technology.

In contrast to previous and related work, this work considers research results from the application based technological relationships as mentioned above. Aspects that are relevant for this task are extracted and used for this approach. Related technologies are grouped in several sets as represented by semantic textual patterns and each project has to be assigned to one set of related technologies. This can be done by using a binary textual classification instead of using a multi-label classification and this enables the use of LSI as a binary semantic classification algorithm.

This approach is evaluated using the example of defense and security (D&S) based technological research projects. This is because the growing importance of this application field leads to an increasing number of research projects and the appearance of many new technologies as indicated by the occurrence of several technology lists or taxonomies (e.g. EDA, WEAG, STACCATO, ESRAB, MCTL, and DSTL) during the last years (Gericke et al., 2009; Te Kulve & Smit, 2003).

The results are compared to a standard text classification algorithm that applies a multi-label classification on the same data set. A centroid vector is created that represents the term vectors from the training examples (projects) of each class (technology) (Takci & Güngör, 2012). This vector is the average vector of all vectors that are assigned to this class in the training phase. Term vectors from further research projects (test examples) are compared to all centroid vectors for identifying similar centroid vectors. We use a well-known similarity measure (Jaccard's coefficient) and a specific threshold to assign test examples to classes that means to identify none, one, or several technologies for each project (Madjarov, Kocev, Gjorgjevikj, & Džeroski, 2012).

The evaluation shows, that the new LSI based approach outperforms the centroid based text classification algorithm concerning the calculated performance measures precision and recall.

2 Background

In this approach, we consider findings of literature that focus on the application based technological relationships. Some important aspects are adapted to this approach and mentioned in Sect. 2.1. Further, text classification approaches that are used in this study are described in Sect. 2.2 and it is explained; why LSI is a good mean to identify the technological relationships from Sect. 2.1. Further, a knowledge structure based classification approach is selected for evaluation purposes. It outperforms further knowledge structure approaches considering the aspects in Sect. 2.1.

2.1 Application based technological relationships

A large number of literature studies the relationships between technologies (Choi et al., 2012; Subramanian & Soh, 2010; Radder, 2009; Jiménez, Garrido-Vega, Díez de los Ríos, & González, 2011; Herstatt & Geschka, 2002; Rubenstein et al., 1977; Fleck & Howells, 2001). Below, most important findings are adapted specifically for this study.

a) An applied science research project can be classified according to a technology only if there is a relation between the project and the technology. The simplest relation is that a project contains research activities concerning the core area of a technology. Then both, the project description and the technology description consist of the same technology specific terms that describe the technological field. Therefore, the project can be directly assigned to one technology by computing the similarity of both descriptions.

b) Technologies are not single data points but they describe a technological field that consists of many different research topics. Inside this field multiple research projects occur. Two research projects, which focus on different topics in a technological field, consist of project descriptions with different terms although they belong to the same technology. Therefore, prevalent features that are characteristic for a technology are not simultaneously present in all projects that belong to one technology.

c) Technological project descriptions consist of a high percentage of term co-occurrence. This is because to describe a technical topic, several technical terms are used that normally occur together in a text phrase. Therefore, conditioned on each technology and on each project, different terms do not occur independently.

d) Applied science research projects focus on an application field and use many different technologies. Literature indicates that these projects consist of up to ten technologies. Therefore, these research project descriptions consist of features from several different technologies.

e) If a research project is assigned to a technology and this technology is related to further technologies then the project can be assigned to these further technologies, too. One kind of relationship is that technologies can be similar to other technologies. They deal with the same technology field but have a different focus e.g. passive radar technology and active radar technology. Technologies are not completely delimited from their similar technologies, which means in some research areas similar technologies overlap. Descriptions of similar technologies also consist of technology specific terms that describe the technological field. Then, a research project can be assigned to a similar technology by comparing the project description to the technology description.

f) A further relationship is seen between a technology and its substitutive technology. These technologies substitute each other e.g. electrical fuel cells, electrical batteries and solar cells in the context of energy supply. An applied science research project normally examines several substitutive technologies to create an application. Then its description consists of terms from different technology fields. By comparing this description to a technology description, we do not get a large similarity because terms from the further technology fields do not appear in the technology description. If the research project examines fuel cell, electrical battery, and solar cell technology in an equally distributed way then the similarity by comparing the project description to the fuel cell technology description is about one third. Therefore, it is necessary to get project and technology descriptions that also contain terms, which describe the application field. Then, one gets a higher similarity by comparing and a better success by assigning a project to a substitutive technology.

g) Integrative technologies sometimes are named complementary technologies and occur together by realizing an application. Examples for two integrative technologies are fuel and lubricants technology. This is because both technologies are used e.g. to create a new power plant prototype. Additionally, predecessor or successor technologies are technologies that precede or succeed another in the process of creating an application. Thus, it is important to use project and technology descriptions that contain terms, which describe the application field, too.

2.2 Text Classification

In general, the aim of text classification is the assignment of pre-defined classes to text documents (Ko & Seo, 2009; Sudhamathy & Jothi Venkateswaran, 2012; Lin & Hong, 2011; Finzen, Kintz, & Kaufmann, 2012). For the identification of technologies standing behind projects, a class can be defined in two different ways. First, each technology can be represented by one class. Using this definition leads to the use of a multi-label classification because a project consists of several technologies and thus, it should be assigned to several classes. Second, a set of related technologies can be represented by one class. As shown in Sect. 2.1, the descriptions of related technologies consist of similar terms that describe application fields or technology areas. Based on these characteristic textual patterns, related technologies can be identified. Using this definition leads to the use of a binary classification where a project is assigned to one class or not.

Extracting technologies from lists or taxonomies normally leads to a large number of technologies. E.g. in the case study (see Sect. 4) 2.850 technologies are extracted from the application field security and defense. Defining a class as a technologies leads to a large number of classes that probably causes performance problems in text classification. Semantic generalizations by grouping related technologies are a good mean to reduce the number of classes.

The assignment of a project to a technology or to a set of related technologies depends on semantic aspects (aspects of meaning) and not on knowledge structure aspects (aspects of words) as described in Sect. 2.1. A single term (a word) that is characteristic for a technology does not have to be in the description of a project even if this project processes the technology but a semantic textual pattern of several terms probably will be. Thus for the text classification approach proposed in this paper, it is more important to compare the aspects of meaning between a project and technologies than to compare the aspects of words between them (Park, Kim, Choi, & Kim, 2012). The aspects of meaning can be identified by calculating the semantic textual patterns.

2.2.1 Knowledge structure approaches

The most frequently used approaches in text classification are knowledge structure approaches. Examples for standard algorithms are k nearest neighbor (k-NN) classification as instance-based learning algorithm, C4.5 as decision tree model, naive Bayes (NB) as a simple probabilistic algorithm, and support vector machine (SVM) (Shi & Setchi, 2012; Lee & Wang, 2012). These approaches are not able to identify hidden semantic textual patterns. Despite this weakness, a knowledge structure approach is selected as baseline for the evaluation to show the success of the used semantic approach.

The centroid-based approach is in contrast to some standard categorization algorithms in text classification where example classes are not described by one centroid vector, but by a number of training examples. We select this approach as baseline. Below, we give detailed explanations for using a centroid-based text classification. Our explanations are based on the results of (Han, 2000) where extensive evaluations of centroid-based classifications and comparisons with other classifiers are described.

With a centroid-based scheme, the characteristics of each class can be summarized. By use of this summarization, several prevalent features are joined together. This is very important for our approach because terms that represent these technology-characteristic features are not simultaneously present in research project descriptions that belong to the technology as shown in Sec. 2.1. Therefore, comparing a term vector from a project (as test example) to a centroid vector leads to better performance than comparing it to term vectors from projects (training examples) that describe a class. We can find a similar summarization in the naive Bayes algorithm where for each class a distribution function is created that represents the term probabilities. Further

algorithms (k-NN, C4.5, SVM etc.) describe a class by a number of training examples and therefore, they do not use summarizations.

Further, a problem in text classification is the appearance of synonyms. Synonyms are different words with identical or at least similar meanings. In technological texts (e.g. in an applied science research project description) we can find them (assign, associate, classify, correlate etc.). By using a summarization, commonly used synonyms also are summarized that means, we can find them in the centroid vector. Therefore, comparing a term vector from a research project to a centroid vector also considers synonyms. Here, we also see that the centroid-based scheme and the naive Bayes algorithm outperform k-NN, C4.5, and SVM that do not use summarization.

Additionally, we focus on the computational complexity of this centroid-based approach. This is relevant because as shown above, we will select 2.850 technologies in our case study that leads to 2.850 classes and that also will lead to a time consuming training and classification phase. In the training phase, we see a linear-time complexity that depends on the number of training examples for the centroid-based approach. We also see a linear complexity in the classification phase that depends on the number of classes. Therefore, the computational complexity in total is very low and it equals the complexity of the naive Bayes algorithm. Thus, the centroid-based scheme and the naive Bayes have a better performance concerning the computational complexity than k-NN, C4.5, and SVM.

We also see advantages of the centroid-based algorithm concerning the naive Bayes algorithm that applies the Bayes theorem with strong (naive) independence assumptions. Conditioned on each class, this means that different terms independently occur. However, as shown in Sec. 2.1 the independence assumption is not true by using project description as training and test examples. Therefore, we think that the centroid-based algorithm also outperforms the Bayes algorithm.

Thus, we use the centroid-based algorithm for the evaluation to compare results of the selected semantic approach to this knowledge based approach.

2.2.2 Semantic approaches

As mentioned above, computational techniques are needed that are able to identify the aspect of meaning by calculating the semantic textual patterns. These techniques use eigenvectors in different variations and apply them on statistical procedures. (Jiang, Berry, Donato, Ostrouchov, & Grady, 1999; Luo, Chen, & Xiong, 2011). With these techniques, words that occur in project or technology descriptions are used in the hidden semantic patterns but also words, that might be in these descriptions (Thorleuchter & Van den Poel, 2012d). This enables the identification of a similarity between a project and a set of technologies even if the words in the project description are completely different than the words in the technology descriptions (Tsai, 2012; Christidis, Mentzas, & Apostolou, 2012). This approach uses LSI as well-known representative of these techniques. It extracts a large number of semantic textual patterns and it reduces their number by considering the values of the eigenvectors (Thorleuchter & Van den Poel, 2013).

LSI is a good mean for the identification of application based technological relationships because it fulfills the requirements from Sect. 2.1 as described below.

The paragraph a) in Sect. 2.1 indicates that the approach should be able to compute textual similarity in project and technology descriptions. LSI assigns project and technology descriptions to semantic textual patterns. Textual similarity between a project and a technology description can be assumed if both descriptions are assigned to the same semantic textual pattern. In the paragraph b) in Sect. 2.1, it is shown that prevalent features that are characteristic for a technology are not simultaneously present in all projects that belong to one technology. LSI as a semantic classification approach always considers this fact by using a semantic indexing that also consists of terms that are not mentioned explicitly in a text but that are related to the corresponding topic. Different terms so not occur independently in the technology or project descriptions as indicated by the paragraph c) in Sect. 2.1. LSI considers this by calculation relationships between projects and technologies based on semantic textual patterns. LSI groups several technologies that are related during the process of creating an application. This means it considers the fact that a project description consists of features from several different technologies as mentioned in the paragraph d) in Sect. 2.1. The paragraphs e), f), and g) indicate that similar,

substitutive, integrative, predecessor, and successor technologies have to be identified by considering terms that also describe the application field (beside the technology area). LSI as semantic classification approach considers all related terms (describing a technology as well as describing an application field).

3 Methodology

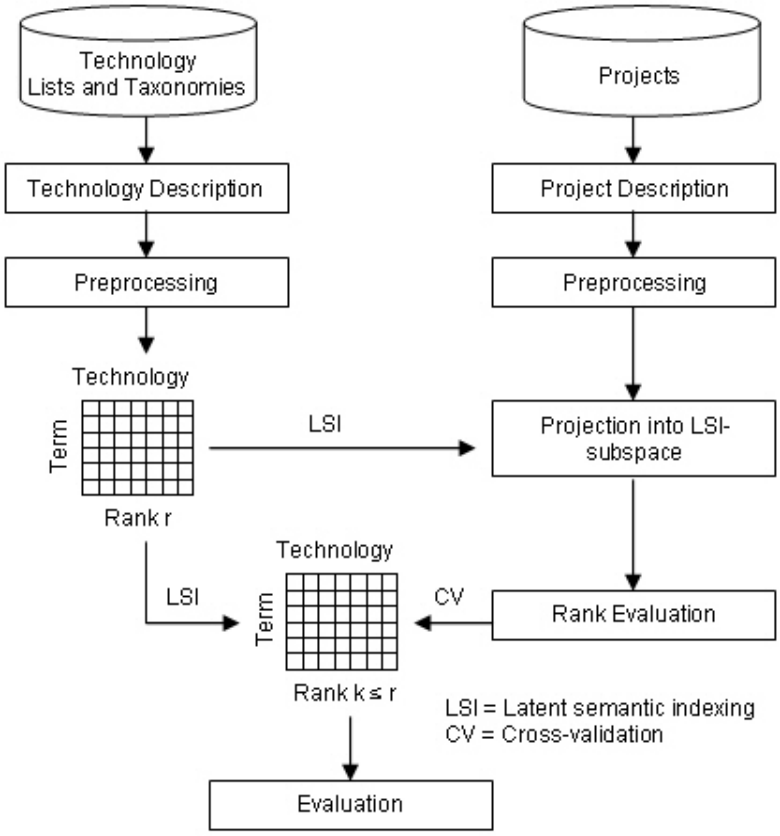

Fig. 1 shows the processing of our approach in different steps.

The methodology selects technology lists or taxonomies as well as information about research projects. Technology descriptions are extracted from the technology lists or taxonomies. Further, projects descriptions are identified or created from the research projects. The technology and project descriptions consist of terms, which describe the technology area as well as the application fields as assumed in Sect. 2.1. They are pre-processed by using tokenization, stop word filtering and stemming. Further, term vectors in a vector space model are created for each technology description and for each project description. LSI is applied to create the semantic textual patterns within the technology descriptions, where the impact of each technology on each semantic textual pattern is calculated. This impact is used to identify related technologies. Technologies with high impact on a specific semantic textual pattern are grouped together in a set of technologies. Projects descriptions are projected into the created LSI subspace where LSI calculated the impact of each project on each semantic textual pattern and thus, on each set of technologies. To determine the optimal value of the rank k as the number of semantic textual patterns, a cross-validation procedure is applied on test and training data from the project descriptions. An evaluation is used to compare the assignment of projects to the related technologies by this LSI based approach to the assignment by a knowledge structure based classification approach (centroid based approach).

3.1 Pre-processing

The extracted textual information (technology and project description) has to be pre-processed. The aim of this step is to create term vectors in vector-space model. This is because textual information in term vectors can be

used for further processing e.g. as input for a singular value decomposition. The textual information has to be prepared in a first step.
This consists of raw text cleaning where specific objects e.g. images or xml-tags are removed. A dictionary is used to identify and correct typographical errors in the raw text. Tokenization is applied that splits the text in terms where the term unit is defined as words. A conversion of terms to lower case is done (case conversion).

In a second step, the text is filtered to reduce the number of distinct terms. Different filtering methods are applied (Thorleuchter, Schulze, & Van den Poel, 2012): Part-of-speech tagging is used to identify the syntactic category of each term (e.g. nouns and verbs) and based on the category, non-informative terms are identified. Stop word filtering is also used to identify the content information of terms. Non-informative terms are discarded (Thorleuchter & Van den Poel, 2012a).

As further filtering method, stemming is applied. While words occur in different forms, stemming use a basic form of words to map related words to this basic form. In contrast to lemmatization, stemming does not consider the context of a word. This leads to problems by processing words with the same spelling but with a different meaning. However, after the preprocessing step, latent semantic indexing is applied on the terms where the aspect of meaning is considered. Thus, at this time, it is not necessary to use lemmatization. The basic form of words is taken over from a dictionary. If a term is not in the dictionary then a set of production rules are applied to transform the word to its basic form. Terms that appear once or twice are discarded as stated in Zipf distribution (Zipf, 1949; Zeng et al., 2012).

Literature shows that term vectors of weighted frequencies outperform term vectors of raw frequencies (Thorleuchter, Van den Poel, & Prinzie, 2010d). Thus, vectors of weighted frequencies are created for each description in a third step. Based on the calculated weights, the importance of a term within the collection of all descriptions can be estimated (Sparck Jones, 1973). A term is assigned to a large weight if it occurs frequently in a small number of descriptions and seldom in further descriptions (Salton & Buckley, 1988). Based on the proposed weighting scheme from Salton, Allan, & Buckley (1994), the a weight $w_{i,j}$ for a term i in description j is calculated by

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

where n is the number of descriptions, m the number of the term vector dimension, df_i is the number of all descriptions containing term i, $tf_{i,j}$ is the term frequency, and idf_i, the inverse descriptions frequency (Chen, Chiu, & Chang, 2005). The different length of the descriptions is considered by using a length normalization factor in the divisor of the formula.

3.2 Identification of hidden semantic textual patterns with singular value decomposition

Based on the calculated vectors of weighted frequencies, a term-by-description matrix can be created. The dimensionality of this matrix is large because of the large number of distinct terms. Most of the terms only occur frequently in a few numbers of descriptions but not in the further descriptions. This leads to many zero values in the matrix and thus, to a small matrix rank. To reduce the dimensionality of the matrix, LSI is used together with a matrix factorization technique. LSI summarizes terms with respect to their semantics (Deerwester et al., 1990). Singular value decomposition as matrix factorization technique identifies the relationships between terms based on their co-occurrences in the descriptions. All related terms are grouped into a semantic textual pattern and each semantic textual pattern has high discriminatory power to other patterns (Thorleuchter & Van den Poel, 2012c).

Each semantic textual pattern is assigned to a singular value by processing the singular value decomposition algorithm. The singular value is calculated by splitting the term-by-description matrix A in a product of the matrices U, Σ, and V^t.

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

Matrix A consists of m terms and n descriptions (m x n matrix) and a rank r ($r \leq \min(m,n)$) because of many zero values in the matrix. Matrix U consists of m terms and r semantic patterns (m x r matrix), matrix V consists of n descriptions and r semantic patterns (n x r matrix), and matrix Σ consists of the r singular values of matrix A. Thus, Σ is a diagonal (r x r) matrix and the singular values are sorted in descending order.

For processing the singular value decomposition, the rank r is important. A large value of r leads to an unmanageable high number of semantic textual patterns. In this case, many semantic textual patterns only occur in a single description but not in several descriptions. For a technology classification, it is important to identify the relationships between different technologies as represented by the technology descriptions. Thus, semantic textual patterns are relevant for this task by considering the relationships between terms based on their co-occurrences in the collection of descriptions. These semantic textual patterns can be identified by reducing the rank r to a parameter k.

As shown above, if k is too large e.g. k = r then too many semantic textual patterns are build that are not relevant. Otherwise, if k is too small then many relevant semantic textual patterns are not considered. Chen et al. (2010) proposes the use of an operational criterion to get an optimal value of k. We satisfy this by calculating the cross-validated area under the ROC (receiver operating characteristics) curve (AUC) for each k (DeLong, DeLong, & Clarke-Pearson, 1988; Hanley & McNeil, 1982; Halpern et al., 1996; Van Erkel & Pattynama, 1998). For this, we construct several rank-k models as described below.

Based on the selection of a specific k, three matrices U_k, Σ_k and V_k are calculated where the first k columns of U, Σ, and V are retained while from k+1 on, the columns are discarded. Thus, the new term-by-description matrix A_k is based on the reduced matrix rank k<r.

$$A_k = U_k \Sigma_k V_k^t \quad (3)$$

The new term-by-description matrix A_k contains the k relevant semantic textual patterns. The matrix U_k shows the impact of each term from the descriptions on each semantic textual pattern from A_k. The matrix V_k shows the impact of each (technology or product) description on each of the k patterns. This enables the identification of related technologies on one hand as well as the assigning of projects to a set of related technologies on the other hand.

Then, project descriptions have to be projected in the same LSI-subspace (Zhong & Li, 2010). This is because based on the corresponding vector of each project description from V_k, a project description can be assigned to a semantic textual pattern and thus, to a set of related technologies. To create such a vector, a term-by-description vector A_d has to be created for each description d that is based on the terms from matrix A. Then, the vector V_d for matrix V_k can be calculated by

$$V_d = A_d' \cdot U_k \cdot \Sigma_k^{-1} \quad (4)$$

The project descriptions are split in test and training examples and a fivefold cross-validation is used on training and test examples (Thorleuchter, Herberz, & Van den Poel, 2012). The training examples are used to identify the rank-k model with the best AUC performance and the test examples are used to evaluate the model as described in Sect. 3.3.

3.3 Evaluation criteria

The evaluation focuses on comparing the performance of the semantic classification approach to a knowledge structure approach. Based on the vector V_d, the impact of a project description from the test example on a set of related technologies as represented by a specific semantic textual pattern is given and evaluated by human experts.

For each set of related technologies, the number of examples that are correctly identified as related to this set are the true positives (TP) and the number of correctly identified non-related examples are the true negatives (TN). The number of not correctly identified related examples are the false negative (FN) and the number of not correctly identified non-related examples are the false positive (FP). Based on these four values, the commonly used evaluation criteria: the precision, the recall, the sensitivity, and the specificity can be calculated by

TP/(TP+FP) (Precision), by TP/(TP+TN) (Recall), by TP/(TP+FN) (Sensitivity), and by TN/(TN + FP) (Specificity). The well-known two dimensional plot of sensitivity versus (1-specificity) is named the receiver operating characteristic curve (ROC) and the AUC is the area under the ROC curve.

4 Empirical verification

4.1 Application Field Defense and Security

For the evaluation, we use technology lists or taxonomies as well as current research projects from the application field D&S. The explanation for the selection of this application field is described below.

D&S is a field where governments are forced to pay more attention because of the rising asymmetrical threat e.g. terrorism (Greenberg, Irving, & Zimmerman, 2009). A possible solution is the use of new techniques based on results of technological research and development. Thus, an increased funding of D&S based technological research and development can be seen by national and European governments. An example is the European Defence Agency (EDA) that was established in 2004. One important task of this organization is the coordination of defence based research between EU Member States (Hoerber, 2012). Further the European Framework Research Program (FP7) contains security research as a central point. As result of growing budgets in the field of D&S research we can monitor a continuous change of the D&S related technological landscape (Thorleuchter, 2008).

D&S is not a technology like laser technology or fuel cell technology but it is an application field. Projects in this field are assigned to applied science research and they combine several technologies (e.g. III-V compounds, stealth technologies, human protection technologies, radar technologies) to create an application (a prototype or a demonstrator) (Thorleuchter, Van den Poel, & Prinzie, 2010a).

Therefore, the technological landscape of D&S is characterized by national and international research funding organizations. Many of these organizations have defined relevant technologies for future D&S applications on their own. These technologies are published as lists or as objects in hierarchical taxonomies, which means normally a two-level tree structure of classifications for a given set of objects. The objects on the second level represent names of D&S related technologies and the objects on the first level represent manually created labels for these technologies. Technology names are described by few technical words e.g. "passive radar technologies" or "active radar technologies" labelled by "radar technologies". Additionally, descriptions of technologies that consist of terms describing the technology itself as well as the corresponding application fields are given (Thorleuchter & Van den Poel, 2011).

The technologies are the basis for research funding activities. That means proposals are manually classified by research funding organizations according to their own technologies. If proposals do not fit with these technologies, they normally are not accepted (McLeish & Nightingale, 2007). In order to acquire funding in this area, researchers should have knowledge about national and international research funding organizations and their appertaining technologies. Thus, this approach considers the relevant organizations and their technologies as mentioned in Sect. 4.1.1.

Beside this overview, a further aspect is to consider. Sometimes D&S research is sensitive concerning technological proliferation (Perry, 2004). That means some technologies have the potential to significantly enhance or degrade national D&S capabilities in the future or to permit significant advances of military capabilities of potential adversaries. Research planners and researchers should have knowledge about the sensibility of their research. Therefore, the overview in Sect. 4.1.1 also includes technologies with proliferation control aspects. Additionally in Sect. 4.1.2, we focus on the acquisition of research projects in the D&S field.

4.1.1 Technologies in D&S

In this section we describe examples for taxonomies and lists of D&S related technologies.

The European Defence Agency (EDA) has been created to help member states of European Union (EU) develop their defence capabilities for crisis-management operations under the "European Security and Defence Policy" (Oikonomou, 2012). One aim of the EDA is to stimulate European research and technology collaboration, focused on improving defence capabilities. The EDA taxonomy of technologies is the basis for this funding and contains about 200 technologies in defence context.

The Western European Armaments Group (WEAG) is a forum for armaments cooperation established by defence ministers of the European NATO nations. It coordinates defence-related research and development projects inside the European Union (Te Kulve & Smit, 2003). The coordination activities of the WEAG are transferred to EDA in 2005 but the WEAG taxonomy of technologies is still in use by many national ministries of defence for defence research funding. The WEAG taxonomy of technologies contains about 200 technologies including underpinning defence technologies, weapon systems related technologies and technologies for (military) products.

The stakeholder's platform for supply chain mapping, market condition analysis and technologies opportunities (STACCATO) is a European Commission-financed activity with the objective to prepare a proposal for a strategic research plan for European security. The STACCATO taxonomy of technologies builds on technology taxonomies from WEAG and United Kingdom and contains about 800 technologies in security context.

The European Security Research Advisory Board (ESRAB) shall make recommendations to the European Commission in the field of strategic missions, focus areas and priorities setting for future security research programs (Remuss, 2010). The ESRAB technology list therefore is a basis for the security part of the European framework research program (FP7). It consists of 150 technologies in security context.

Each European member state has own technology collections (lists or taxonomies) for D&S. In general they are created by ministries of defence and unfortunately they are very often classified as restricted information but most of these technological collections are based on WEAG taxonomy like described above.

The Militarily Critical Technologies List (MCTL) is a compendium of existing goods and technologies that would permit significant advances in the development, production and use of military capabilities of potential adversaries (Bradley, 1989). This technology list contains about 600 technologies in proliferation control context.

The Developing Science and Technologies List (DSTL) is a compendium of scientific and technological capabilities being developed worldwide that have the potential to significantly enhance or degrade US military capabilities in the future. This list includes technologies from basic research, applied research and advanced technology development and it contains about 900 technologies in proliferation control context.

Further DSTL and MCTL are a basis for the technological part of the Waasmar List, the armaments export control list for conventional arms and dual-use goods and technologies.

We have extracted technologies from EDA, WEAG, STACCATO, ESRAB, MCTL and DSTL as structured documents (XML). In our web application, they can be selected by users to get a user defined technology collection. The XML structure consists of an identifier and a technology label.

4.1.2 Research Projects from D&S

For our test and training set, we need D&S related and innovative projects. They can be found in the United States Small Business Innovation Research (SBIR) Program and the Small Business Technology Transfer (STTR) Program. SBIR and STTR ensure that small, high-tech and innovative businesses are a significant part of the United States federal government's research and development efforts. Eleven federal departments participate in SBIR and STTR programs awarding $ 2 billion to small high-tech businesses (Lockett, Siegel, Wright, & Ensley, 2005). The central point in SBIR and STTR research is D&S because of the height award amount of the Department of Homeland Security, the Environmental Protection Agency and the Department of Defense divided in Air Force, Army, Chemical and Biological Defense Program (CBD), Defense Advanced

Research Projects Agency (DARPA), Defense Logistics Agency (DLA), Defense Microelectronics Activity (DMEA), Defense Technical Information Center (DTIC), Defense Threat Reduction Agency (DTRA), Missile Defense Agency (formerly BMDO), National Geospatial-Intelligence Agency (NGA) (formerly NIMA), Navy, Special Operations Acquisition and Logistics Center (SOCOM).

The projects are published as non-proprietary textual data with title and abstract. The abstract consists of terms that represent the technological field as well as the application field. Therefore, we use them to evaluate the approach.

4.2 Data Characteristics

In this study, we use the technology and project descriptions from Sect. 4.1. All descriptions are in English language.

Table 1 provides summary information of the (randomly-selected) training and test set. The optimal SVD dimension is calculated using the training set and a regression model is estimated. The test set is used to show the success of the regression model compared to the frequent baseline as calculated from the relative percentage in Table 1.

	Number of items	Relative percentage
Taxonomies / Technology lists	6	
Technology descriptions	2900	
Training set: Project descriptions	480	80
Test set: Project descriptions	120	20
Total	600	

Table 1: Overview on the data characteristics

4.3 Optimal dimension selection

The high number of 2900 technology description can be reduced to a small number of sets of related technologies because many of these descriptions describe equal technologies or similar technologies while other describe substitutive, integrative, predecessor, or successor technologies. A human based evaluation identifies the AUC for specific selected values of k. For other values of k, regression based interpolation is used to construct new data points between two known data points.

The number of selected semantic textual patterns (SVD dimension or rank k) is represented by the x-axis. The y-axis represents the cross-validated AUC (see Fig. 2). It can be seen that the AUC increases up to the use of 200 semantic textual patterns. Using more than 200 semantic textual patterns in the SVD model leads to a higher complexity of the model however, the cross-validated AUC performance does not increase. Thus, the parameter k is set to 200.

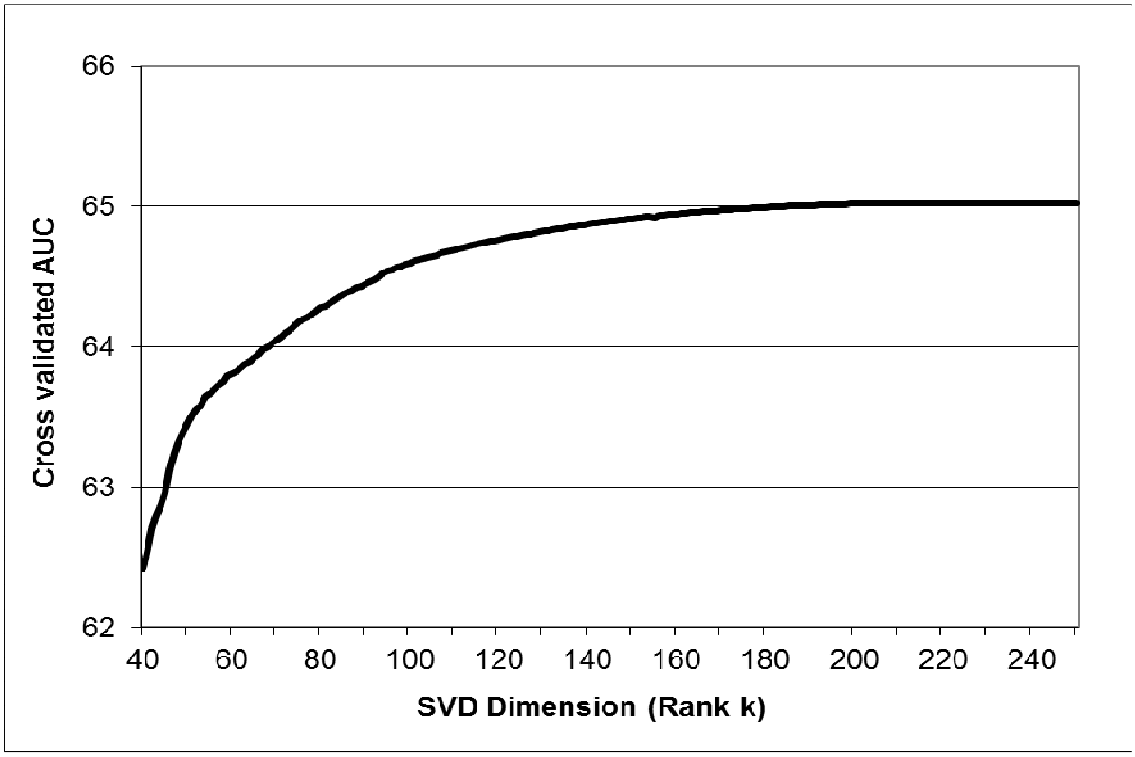

Figure 2: Calculating an optimal SVD dimension
The baseline has an AUC value of 50. It is also important to know that this approach outperforms the baseline even if a small value of k (at about 40) is selected.

4.4 Case study results

Here, an example for the results of this approach is presented. These example results confirm results of a further knowledge structure based approach (Thorleuchter and Poel 2011). Unfortunately, that approach could not be used for the evaluation because the precision and recall values are calculated based on a very small subset of technologies and projects.

A research project is used as described below: The 2002 SBIR / STTR research project 57405 with the title: "Tunable diode-pumped IR laser source" has the following abstract: "The Space Based Laser (SBL) requires a Low Energy Laser (LEL) system to serve as a high fidelity surrogate during startup and optical alignment portions of test operations. In this proposal, we will develop a CW, diode-pumped solid state laser that can meet the requirements for the LEL, namely a CW power level in the 1-10 W range, and wavelengths in the 2600-2900-nm region. The device, based on a direct diode-pumped Er:YLF crystal, is rugged, compact, tunable, and well suited for space - based systems."
The semantic approach has assigned this project to a set of 12 related technologies as mentioned in Table 2.

Technology list / Taxonomy	Technology label
EDA	Communications Systems – IR / Visible / UV
ESRAB	Space Systems
WEAG	Laser Sensors
STACCATO	Space Based Lasers
STACCATO	Communications systems - IR / Visible / UV laser
STACCATO	IR / Visible / UV laser
MCTL	Laser Location Systems
MCTL	Multispectral and Hyperspectral Space Sensor Systems

Technology list / Taxonomy	Technology label
EDA	Communications Systems – IR / Visible / UV
MCTL	Space Laser Diodes
MCTL	Tunable Solid-State Lasers
DSTL	Excimer Lasers (LELs), Excimer
DSTL	Free Electron Laser (FEL) (HPM NB Sources)

Table 2: Example of related technologies from different technological lists and taxonomies

4.5 Comparing the performance of the approach

Besides evaluating the optimal performance of this approach as based on the value of k, the overall performance of this semantic approach is also compared to a centroid approach by use of precision and recall measures. Both approaches are comparable because they assign a project to a set of technologies.

For each of the 200 sets of technologies, the precision and recall values for the semantic approach are estimated by human experts as described in 3.3. Then, the average precision and recall values are calculated.

We have defined a centroid classifier that assigns a project to a technology if at least z% of all stemmed and stop word filtered terms from one technology description appear in the project description. If z is too large then probably we do not get many projects assigned to a technology in the learning phase. This decreases the quality of the corresponding centroid vector. If z is too small then probably we get many projects assigned to a technology that normally are not related to this technology. This also decreases the quality of the corresponding centroid vector. The value of z is estimated by a human expert. Each centroid vectors represents a set of technologies. This is used for the calculation of precision and recall values as described above.

For comparing, the F-measure is used because precision and recall are equally important. The semantic approach gets a precision of 76 % at a recall of 48 % while the centroid approach gets a precision of 69 % at a recall of 44 %. This leads to an F-measure of 59 % for the semantic approach in contrast to an F-measure of 54 % for the centroid approach.

5 Conclusions

This paper proposes a new approach that classifies applied science research projects according to corresponding technologies of research-funding organizations. It considers that technologies are related during the process of creating an application. LSI is used to identify these related technologies based on semantic textual patterns occurring in the technology descriptions. Project descriptions - divided in training and test examples - are projected into the LSI subspace. They are also used to estimate the parameters and to evaluate this approach.

As a result, it is shown that LSI as semantic classification approach is suited to identify the relationships among technologies because it considers well terms from the technology area as well as terms from the application field. The automated identification of these semantic relationships is not possible with knowledge structure based approaches. Thus, the results contribute to the existing literature concerning the application based technological relationships.

Further, LSI is suited to assign projects to a specific set of related technologies as represented by a semantic textual pattern. Here, LSI also outperforms knowledge structure based approaches that assign projects to each technology separately. Thus, the results are helpful for researchers and research-funding planners.

Future avenues of research could be the use of this approach in the field of explorative scenario-based technological roadmapping. This is because this specific roadmapping approach considers the relationships between the investigated technologies during the process of creating applications. Up to now, the identification of relationships is done manually by human experts. That restricts this roadmapping approach to a small number

of investigated technologies. However, using the automated LSI based approach is helpful for identifying relationships and it probably enables the investigating of a large number of technologies.

For the case study, technologies from the field of D&S are selected. A further direction of research is the use of different application fields to show the success of this approach.

Acknowledgments
This project is realized using SAS v9.1.3, SAS Text Miner v5.2, Fraunhofer Idea Web Miner v1.0, and Matlab v7.0.4. Further, a self-developed program is used for web content mining to collect the data.

References

Beaudry, C., & Allaoui, S. (2012). Impact of public and private research funding on scientific production: The case of nanotechnology. *Research Policy*, In Press.
Bradley, V.G. (1989). National strategies for technology trade: A response to Chris Hill. *Technology in Society*, 11(2), 181-188.
Bradshaw, D.H., Empy, C., Davis, P., Lipschitz, D., Nakamura, Y., & Chapman, C.R. (2008). Trends in Funding for Research on Pain: A Report on the National Institutes of Health Grant Awards Over the Years 2003 to 2007. *The Journal of Pain*, 9(12), 1077-1087.
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.
Chen, M.-Y., Chu, H.-C., & Chen, Y.-M. (2010). Developing a semantic-enable information retrieval mechanism. *Expert Systems with Applications*, 37(1), 322-340.
Choi, J.Y., Lee, J.H., & Sohn, S.J. (2009). Impact analysis for national R&D funding in science and technology using quantification method II. *Research Policy*, 38(10), 1534-1544.
Choi, S., Park, H., Kang, D., Lee, J.Y., & Kim, K. (2012). An SAO-based text mining approach to building a technology tree for technology planning. *Expert Systems with Applications*, 39(13), 11443-11455.
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.
Finzen, J., Kintz, M., & Kaufmann, S. (2012). Aggregating web–based ideation platforms. *International Journal of Technology Intelligence and Planning*, 8(1), 32-46.
Fleck, J., & Howells, J. (2001). Technology, the Technology Complex and the Paradox of Technological Determinism. *Technology Analysis & Strategic Management*, 13(4), 523-531.
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.
Geschka, H., Lenk, T., & Vietor, J. (2002). The idea and project database of WELLA AG. *International Journal of Technology Management*, 23(5), 410-416.
Geschka, H. (1983). Creativity techniques in product planning and development: A view from West Germany. *R&D Management*, 13(3), 169-183.
Greenberg, M., Irving,W., & Zimmerman, R. (2009). Allocating U.S. Department of Homeland Security funds to States with explicit equity, population and energy facility security criteria. *Socio-Economic Planning Sciences*, 43(4), 229-239.
Grimpe, C. (2012). Extramural research grants and scientists' funding strategies: Beggars cannot be choosers? *Research Policy*, In Press.
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.
Han, E.S., & Karypis, G. (2000). Centroid-Based Document Classification: Analysis and Experimental Results. In: *Principles of Data Mining and Knowledge Discovery* (pp. 116-123). Berlin: Springer.
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.
Herstatt, C., & Geschka, H. (2002). Need assessment in practice - methods, experiences and trends. *International Journal of Entrepreneurship and Innovation Management*, 2(1), 56-68.

Hicks, D. (2012). Performance-based university research funding systems. *Research Policy*, 41(2), 251-261.
Hoerber, T. (2012). New horizons for Europe – A European Studies perspective on European space policy. *Space Policy*, 28(2), 77-80.
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.
Jiménez, C.H.O., Garrido-Vega, P., Díez de los Ríos, J.L.P., & González, S.G. (2011). Manufacturing strategy–technology relationship among auto suppliers. *International Journal of Production Economics*, 133(2), 508-517.
Jiricka, A., & Pröbstl, U. (2012). The role of SEA in integrating and balancing high policy objectives in European cohesion funding programmes. *Environmental Impact Assessment Review*, In Press.
Kim, Y., Toh, K.A., Teoh, A.B.J., Eng, H.L., & Yau, W.Y. (2012). An online AUC formulation for binary classification. *Pattern Recognition*, 45(6), 2266-2279.
Ko, Y., & Seo, J. (2009). Text classification from unlabeled documents with bootstrapping and feature projection techniques. *Information Processing & Management*, 45(1), 70-83.
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.
Lepori, B. (2011). Coordination modes in public funding systems. *Research Policy*, 40(3), 355-367.
Lin, M-H., & Hong, C-F. (2011). Opportunities for Crossing the Chasm between Early Adopters and the Early Majority through New Uses of Innovative Products. *The Review of Socionetwork Strategies*, 5(2), 27-42.
Lockett, A., Siegel, D., Wright, M., & Ensley, M.D. (2005). The creation of spin-off firms at public research institutions: Managerial and policy implications. *Research Policy*, 34(7), 981-993.
Ludwig, R., Roson, R., Zografos, C., & Kallis, G. (2011). Towards an inter-disciplinary research agenda on climate change, water and security in Southern Europe and neighboring countries. *Environmental Science & Policy*, 14(7), 794-803.
Luo, Q., Chen, E., & Xiong, H. (2011). A semantic term weighting scheme for text categorization. *Expert Systems with Applications*, 38(10), 12708-12716.
Madjarov, G., Kocev, D., Gjorgjevikj, D., & Džeroski, S. (2012). An extensive experimental comparison of methods for multi-label learning. *Pattern Recognition*, 45(9), 3084-3104.
McLeish, C., & Nightingale, P. (2007). Biosecurity, bioterrorism and the governance of science: The increasing convergence of science and security policy. *Research Policy*, 36(10), 1635-1654.
Mobjörk, M., & Linnér, B.O. (2006). Sustainable funding? How funding agencies frame science for sustainable development. *Environmental Science & Policy*, 9(1), 67-77.
Oikonomou, I. (2012). The European Defence Agency and EU military space policy: Whose space odyssey? *Space Policy*, 28(2), 102-109.
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.
Perry, W.J. (2004). Military technology: an historical perspective. *Technology in Society*, 26(2), 235-243.
Radder, H. (2009). Science, Technology and the Science-Technology Relationship. *Philosophy of Technology and Engineering Sciences*, 2009, 65-91.
Remuss, N.L. (2010). Creating a European internal security strategy involving space applications. *Space Policy*, 26(1), 9-14.
Rubenstein, A.H., Douds, C.F., Geschka, H., Kawase, T., Miller, J.P., Saintpaul, R., & Watkins, D. (1977). Management perceptions of government incentives to technological innovation in England, France, West Germany and Japan. *Research Policy*, 6(4), 324-357.
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.
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.
Sudhamathy, G. & Jothi Venkateswaran, C. (2012). Fuzzy Temporal Clustering Approach for E-Commerce Websites. *International Journal of Engineering and Technology*, 4(3), 119-132.
Subramanian, A.M., & Soh, P.H. (2010). An empirical examination of the science–technology relationship in the biotechnology industry. *Journal of Engineering and Technology Management*, 27(3-4), 160-171.
Takci, H. & Güngör, T. (2012). A High Performance Centroid-based Classification Approach for Language Identification. *Pattern Recognition Letters*, In Press.
Te Kulve, H., & Smit, W.A. (2003). Civilian–military co-operation strategies in developing new technologies. *Research Policy*, 32(6), 955-970.

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). Mining innovative ideas to support new product research and development. In H. Locarek-Junge, & C. Weihs (Eds.), *Classification as a Tool for Research* (pp. 587-594). Berlin: Springer.
Thorleuchter, D., Van den Poel, D., & Prinzie, A. (2010d). Extracting consumers needs for new products - A web mining approach. In *Proceedings WKDD 2010* (p. 441.). Los Alamitos: IEEE Computer Society.
Thorleuchter, D., & Van den Poel, D. (2011). 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., & 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. *Advances in Intelligent and Soft Computing*, 168, 563-568.
Thorleuchter, D., & Van den Poel, D. (2012b). Predicting e-commerce company success by mining the text of its publicly-accessible website. *Expert Systems with Applications*, 39(17), 13026-13034.
Thorleuchter, D., & Van den Poel, D. (2012c). Using NMF for Analyzing War Logs. *Communications in Computer and Information Science*, 318, 73-76.
Thorleuchter, D., Schulze, J., & Van den Poel, D. (2012). Improved Emergency Management by Loosely Coupled Logistic System. *Communications in Computer and Information Science*, 318, 5-8.
Thorleuchter, D., & Van den Poel, D. (2012d). Using Webcrawling of Publicly-Available Websites to Assess E-Commerce Relationships. In *SRII Global Conference* 2012. New York: IEEE press, in press.
Thorleuchter, D., & Van den Poel, D. (2013). Improved Multilevel Security with Latent Semantic Indexing. *Expert Systems with Applications*, 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.
Yu, L., Hurley, T., Kliebenstein, J., & Orazem, P. (2012). A test for complementarities among multiple technologies that avoids the curse of dimensionality. *Economics Letters*, 116(3), 354-357.
Zeng, J., Duan, J., Cao, W., & Wu C. (2012). Topics modeling based on selective Zipf distribution. *Expert Systems with Applications*, 39(7), 6541-6546.
Zhong, J., & Li, X. (2010). Unified collaborative filtering model based on combination of latent features. *Expert Systems with Applications*, 37(8), 5666-5672.
Zipf, G. K. (1949). *Human Behaviour and the Principle of Least Effort.* Cambridge: Addison-Wesley.