2011 International Conference on Uncertainty Reasoning and Knowledge Engineering

Semantic Technology Classification - A Defence and Security Case Study

Dirk Thorleuchter
Fraunhofer INT
Euskirchen, Germany
dirk.thorleuchter@int.fraunhofer.de

Dirk Van den Poel
Faculty of Economics and Business Administration
Ghent University, Department of Marketing
Gent, Belgium
dirk.vandenpoel@ugent.be

***Abstract*— In the last years, an increasing collaboration between defense and (civil) security, especially in technological areas can be observed. Here, an approach that automatically extracts relationships among defense - based technologies and security - based technologies is introduced. Information about these relationships can be used as planning support to defense and security - based technological research planners specifically for collaboration decisions.**

This approach uses machine learning techniques as supervised learning methods and a multi-label text classification algorithm to identify related technologies in different technological lists or taxonomies. Additionally, a web mining approach is used to create training examples. Similarities are computed by use of Jaccard's coefficient and by use of the fuzzy alpha cut method. Further, this approach uses standard text mining methods to prepare unstructured textual information.

Text Mining, Web Mining, Text Classification, Defense, Security, Technology, Taxonomy, Machine Learning, Multi Label Classification

I. INTRODUCTION

In the last years, the rising asymmetrical threat is causing governments to pay more attention to defense and security (D&S), especially in technological areas. New and ever more complex tasks in areas concerned with defense against these new types of threats require additional research and development of new techniques [5]. For this reason, European governments and the European Union are increasingly funding D&S - based technological research. For example, the European Defense Agency (EDA) was established in 2004 and coordinates defense - based research between Member States of the European Union. Further, the current European Framework Research Program contains security research as a central point [4]. As result of growing budgets in the field of D&S research, one can monitor an increasing collaboration especially between defense - based research and (civil) security - based research. This leads to a continuous change of the D&S related technological landscape.

It also has an effect on the planning of the technological and scientific research program e.g. of the German ministry of defense, running over 1000 different technological research projects simultaneously. Concerning these defense - based research projects, one important task is to identify new research projects (e.g. defense - based research projects from other Member States of the European Union or security - based research projects) for potential research collaboration. Here, research collaboration is possible only if a new research project is assigned to a similar technology or to a similar application as an existing German defense - based research project.

German defense research projects are assigned to the WEAG taxonomy of technologies. However, defense - based research projects from other Member States of the European Union as well as security - based research projects are assigned to different technological lists or taxonomies. Today, the identification of similar technologies or applications is done manually (that means by humans) without the support of text mining. Therefore, this paper describes a text mining approach that automatically assigns research projects to technological lists or taxonomies and that identifies relationships among D&S - based technologies and applications. Information about these relationships can be used as planning support to D&S - based technological research planners specifically for collaboration decisions.

This approach focuses on D&S - based technological lists and taxonomies, where items or taxonomy objects represent textual labels of D&S - based technologies (classes). It uses machine-learning techniques as supervised learning methods. For this, descriptions of the technology and of potential applications are extracted from the internet by use of a web mining approach and term vectors are built based on these descriptions (training examples). Then, each class is represented by a set of training examples. Additionally, the approach uses term vectors from descriptions of D&S - based research projects as test examples and it assigns them to several classes by use of multi-label text classification. Classes that are assigned by the same test example are presented to the user (e.g. research planner) as similar technologies.

II. METHODOLOGY

Labels of Technologies are published as items in technological 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 labels of D&S - based technologies and the objects on the first level represent manually created labels for technology fields or areas. Examples for

technology labels are "passive radar technologies" and "active radar technologies". Examples for the corresponding technology field / area label is “radar technologies”. For this approach, objects on the first level are not of interest because it focuses on technologies but not on technology fields / areas. Therefore, in Sect. III, technological lists and taxonomies are presented. They are used to extract technological labels.

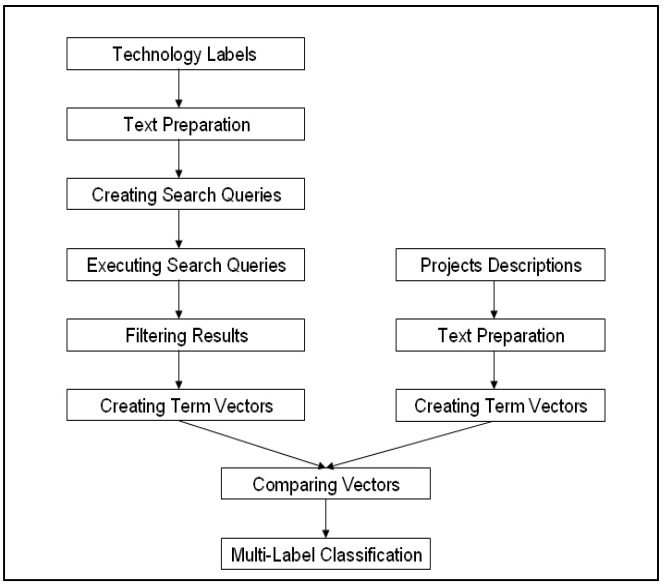

Figure 1. This figure shows the processing of the web mining approach in different steps.

However, very often, technological lists or taxonomies do not contain detailed descriptions. Additionally, if one can find a technology description then probably it is not up-to-date, it consists of heterogeneous quality, or terms in this description only focus on the technology but not on applications that can be realized by use of the technology.

For this approach, it is necessary to get current and homogeneous descriptions of technologies for a more precise association with research project descriptions. Additionally, descriptions are needed that consist of terms, which describe the technology as well as possible applications.

This is because technologies substitute each other e.g. electrical fuel cells, electrical batteries and solar cells in the context of energy supply. Then, a research project that focuses on the energy supply for the electronic equipment of the soldier by use of electrical battery technologies can collaborate on a research project that focuses on the same application field by use of electrical fuel cell technology. If current descriptions of technologies are used that also include possible applications then these substitutive relationships among technologies can be discovered by use of text classification methods.

To get a current description in homogeneous quality with terms that represent a technology as well as applications, a web mining approach is used in the training phase. This is because the approach identifies 2900 technology labels from different technological lists and taxonomies. To create descriptions for each technology and to keep them up-to-date is time and cost consuming.

In the internet, one can find a large amount of textual information for nearly each topic [11]. If one specifically focuses on information about technological topics in the domain of D&S then many patents, research papers, technical reports etc. can be identified. Using this information leads to a description that consists of current research activities and that is up-to-date as demanded above. Therefore, in a training phase, text phrases are extracted from the internet as corpus - based term co-occurrence data (see Sect. III).

Technology labels are tokenized, stop word filtered and stemmed to use them as search queries (see Sect. IV). The search queries are executed by use of an internet search engine and query results are filtered to reduce the number of result items (see Sect. V). Then, for each filtered query result, a term vector is built. It is assigned to its corresponding technology (see Sect. VI). This means, each technology is represented by several term vectors (training examples).

For classification, descriptions of current D&S - based research projects are acquired (see Sect. III). These project descriptions are used to create test examples. They are prepared by use of tokenization, stop word filtering and stemming. Additionally, terms and term frequencies are extracted for each research project. Then, term vectors (test examples) are created based on these extracted terms and on term frequencies. Each term vector from the test examples is compared to every term vector from the training examples by use of Jaccard’s coefficient and the fuzzy alpha cut method to realize a multi-label text classification.

As a result, each research project can be assigned to none, one or several technologies. Additionally, two technologies can be defined as related if they are assigned by the same research project.

III. TEXT ACQUISITION

We use technological lists and taxonomies that consist of D&S - based technology labels from the European Defense Agency (EDA), the Western European Armaments Group (WEAG), the stakeholder's platform for supply chain mapping, market condition analysis and technologies opportunities (STACCATO), the European Security Research Advisory Board (ESRAB) [3], the Militarily Critical Technologies List (MCTL) [8], and the Developing Science and Technologies List (DSTL) [2].

For the training set, D&S - based research projects are selected from the United States Small Business Innovation Research (SBIR) Program and the Small Business Technology Transfer (STTR) Program. These SBIR and STTR research projects are founded by United States Department of Homeland Security, the Environmental Protection Agency, the United States Department of Defense, and further D&S – agencies. The projects are published as non-proprietary textual data with title and abstract. The abstract consists of terms that describe the used technology as well as possible applications. Therefore, they can be used to evaluate this approach.

IV. Text Preparation and Search Query Creation

In a pre-processing phase, the provided technology labels are tokenized [9] by using the term unit as word. To build the queries, stemmed terms [6] are used. This is because searching in a web search engine with a stemmed term leads to query results that contain several different terms, which all have the same stem. Further, search queries should not contain stop-words [12] because normally search engines delete these terms automatically. Therefore, stop-words are deleted by use of a standard stop word list and all further terms are stemmed using the well known Porter stemmer [7]. Then, search queries are built that consist of stemmed and stop word filtered terms from each technology label.

V. Search Query Executing and Result Filtering

Web services are used to execute the created search queries. A web service is a software system that is designed to support interoperable machine-to-machine interaction over a network. Frequently, web services are just web based advanced programming interfaces. An access to these interfaces is possible over the internet. Then, the requested service is executed and result data is transferred back to an application that requested the service [13]. A lot of internet search engines offer web services. In this approach, Google is used as internet search engine. The query results of this search engine consist of a title, a short description that contains terms from the search query in bold print, and a hyperlink that leads to the full text. In this approach, the short description from each query result is used for further processing.

The short description consists of one or several text patterns. If there are several text patterns then the text patterns are separated by several dots: e.g. '...'. If these dots occur between the terms from the search query then they only occur together in a document but not in the same text pattern. For this web mining approach, all terms from the search query should occur in the same text pattern to ensure that the text pattern is related to the corresponding technology. These text patterns are selected for further processing; the other text patterns are discarded. Then the selected text patterns are used to represent the technologies as described below.

VI. Multi-Label Classification

In a training phase, technology representations are built. For each technology label, the web mining approach extracts several assigned text patterns. Terms in each text pattern are tokenized, stop word filtered and stemmed. These stop word filtered and stemmed terms and their term frequencies are used to build term vectors concerning vector space model. A term vector component equals the term frequency if the term occurs in the corresponding text pattern and it is zero if the term does not occur. The term frequency is simply the number of times a given term appears in a text pattern. Because of this training phase, each technology (class) is represented by several term vectors (training examples) belonging to the same class.

For classification, the similarity is computed between the training examples and the vectors from D&S - based research projects (test examples). To create these research project vectors, descriptions of research projects are tokenized, stop word filtered, and stemmed. Then, they are transformed to term vectors as described above. Each research project is probably related to several technologies from different technological lists and taxonomies. Therefore, a multi-label classification is used to assign the test examples to several classes.

The similarity between the training examples and the test examples is measured by use of Jaccard's coefficient [14]. This is because Jaccard's coefficient measure considers the different sizes of both vectors in contrast to other similarity measures. The Jaccard's result values are always between 0% and 100% for each combination of training and test examples. Additionally, the alpha-cut method [1] is used to identify similar training and test examples and to decide whether a test example is assigned to a class or not. An alpha-cut of the Jaccard's coefficient is the set of all combinations of training and test examples such that the appertaining Jaccard's result value is greater than or equal to the value of alpha [10]. Then, a new test example is assigned to all classes that correspond to all its similar training examples. Additionally, related technologies can be identified that are assigned by the same research project.

VII. Results (Examples)

Here, an example for the results of the multi-level classification approach is presented. A research project is used as described below:

Research project title: “Tunable diode-pumped IR laser source”

TABLE I. Example of related technologies from different technological lists and taxonomies

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
MCTL	Space Laser Diodes
MCTL	Tunable Solid-State Lasers
DSTL	Excimer Lasers (LELs), Excimer
DSTL	Free Electron Laser (FEL) (HPM NB Sources)

Research project 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."

After classification, the approach automatically identifies the following related technologies from different technological lists and taxonomies:

VIII. EVALUATION

In the evaluation, the technology collection of all taxonomies and lists is used as described above, which means all technologies from EDA, WEAG, STACCATO, ESRAB, MCTL and DSTL are aggregated to a collection of 2900 technologies. They are represented by term vectors from the web mining approach. Additionally, project descriptions from 600 research projects are used to create term vectors. Then, projects are classified according to several technologies by use of multi-label classification.

For classification, the alpha cut value has to be determined. If the percentage alpha is small, then one obtains many result items and therefore, many research projects are assigned to technologies in the test phase. This leads to a small precision value because many of these projects are not assigned correctly to these technologies. If alpha is large, then one only obtains a very small number of results and probably the recall value is small because most of the projects that are related to technologies are not identified. A human expert checks several project descriptions for an optimal value of alpha. He concludes that 15% is a good compromise. Therefore, we set alpha to 15% as default value.

To measure the performance one uses precision and recall measures commonly used in information retrieval. Here two classes (A means a research project is related to a technology, B means a project is not related to a technology) are in our data, and each project - technology combination is classified as either A or B. A human expert also assigns projects to technologies. These results are the ground truth for our evaluation because they refer to information that is collected by experts. The number of projects and technologies is limited in our evaluation because the manual assignment is time consuming.

The human expert randomly selects a subset of 20 projects and a subset of 10 technologies. Then he assigns all of these 200 project - technology combinations to A or B. He compares these results to the results of the approach and computes the precision and recall values. The F1 - measure is used because precision and recall are equally important. The abovementioned procedure results in a precision value of 70%, a recall value of 45% and an F1 - measure value of 55%.

In contrast to this, the precision and recall for the baseline is computed. The chance baseline, which assigns a classification randomly, is not used because the distribution of the data concerning A and B is skewed. Therefore, the frequency baseline is used, which means each project is classified with a specific percentage as either A or B.

To compute this percentage, further (thirty) research projects are used and are assigned to the technologies. As a result, projects are identified that can be assigned to about 50 technologies as well as projects that can be assigned to less than ten technologies. Therefore, the mean value is estimated at about 29 from 2900 technologies that mean, the frequency baseline is set to 1%. As further baseline, a constant model is used that predicts the class A for all projects regardless of its textual description.

For the frequency baseline, a precision value of 1% is obtained, a recall value of 1%, and a F1 - measure of 1%. Additionally, the use of the constant baseline results in a precision value of 1%, a recall value of 100%, and a F1 - measure of 1.98%.

Furthermore, the result values are compared to another approach that uses further term vectors from 200 D&S - based research projects as training examples but that do not use text patterns from the internet. This approach is evaluated by use of cross validation. A precision value of 45% is obtained, a recall value of 41% and an F1 - measure value of 43%. This is because the number of 200 training examples is probably too small to represent the technologies.

[1] A. J Abebe, V. Guinot, and D. P. Solomatine, "Fuzzy alpha-cut vs. Monte Carlo techniques in assessing uncertainty in model parameters," Proc. 4-th International Conference on Hydroinformatics, Iowa City, USA, 2000.

[2] Defense Threat Reduction Agency, "Developing Science and Technologies List," Ft. Belvoir, USA, 2000.

[3] European Commission, "Meeting the challenge - the European Security Research Agenda," Luxembourg, p. 27, 2006.

[4] Eropean Council and European Parliament, "DECISION No 1982/2006/EC of 18 December 2006 concerning the 7th Framework Programme of the European Community for research, technological development and demonstration activities," Bruxelles, p. 5, 2006.

[5] W. Gericke, D. Thorleuchter, G. Weck, F. Reilaender, and D. Loss, „Vertrauliche Verarbeitung staatlich eingestufter Information - die Informationstechnologie im Geheimschutz," Informatik Spektrum, vol. 32 (2), pp. 102-109, 2009.

[6] A. Hotho, A Nuernberger, and G. Paass, "A Brief Survey of Text Mining," LDV Forum, vol. 20 (1), pp. 19-26, 2005.

[7] M. F. Porter, "An algorithm for suffix stripping," Program, vol. 14 (3), pp. 130-37, 1980.

[8] Secretary of Defense, Acquisition, Technology, and Logistics, "Militarily Critical Technologies List," Pentagon, VA, USA, 2006.

[9] D. Thorleuchter., D. Van den Poel, "Companies Website Optimising concerning Consumer's searching for new Products," in: Proc. URKE 2011, IEEE Press, Los Alamitos, CA, 2011.

[10] D. Thorleuchter, "Finding new technological ideas and inventions with text mining and technique philosophy" in: Data analysis, machine learning and applications, C. Preisach Ed. Berlin, Springer, 2008, pp. S.413-420.

[11] D. Thorleuchter, D. Van den Poel, and A. Prinzie, "A compared R&D-based and patent-based cross impact analysis for identifying relationships between technologies," Technological Forecasting and Social Change, vol 77 (7), pp. 1037-1050, 2010.

[12] D. Thorleuchter, D. Van den Poel, and A. Prinzie, "Mining Ideas from Textual Information," Expert Systems with Applications, vol. 37 (10), pp. 7182-7188, 2010.

[13] D. Thorleuchter, D. Van den Poel, and A. Prinzie, "Extracting Consumers Needs for New Products," in Proceedings of WKDD 2010, IEEE Computer Society, CA: Los Alamitos, 2010, pp. 440-443.

[14] D. Thorleuchter, D. Van den Poel, and A. Prinzie, "Mining Innovative Ideas to Support new Product Research and Development," in: Classification as a Tool for Research, H. Locarek-Junge and C. Weihs Eds. Berlin, Springer, 2010, pp. 587-594.