Finding new technological ideas and inventions with text mining and technique philosophy

Dirk Thorleuchter

Fraunhofer INT, D-53879 Euskirchen, Appelsgarten 2, Germany
Dirk.Thorleuchter@int.fraunhofer.de

Abstract. Text mining refers generally to the process of deriving high quality information from unstructured texts. Unstructured texts come in many shapes and sizes. It may be stored in research papers, articles in technical periodicals, reports, documents, web pages etc. Here we introduce a new approach for finding textual patterns representing new technological ideas and inventions in unstructured technological texts.

This text mining approach follows the statements of technique philosophy. Therefore a technological idea or invention represents not only a new mean, but a new purpose and mean combination. By systematic identification of the purposes, means and purpose-mean combinations in unstructured technological texts compared to specialized reference collections, a (semi-) automatic finding of ideas and inventions can be realized. Characteristics that are used to measure the quality of these patterns found in technological texts are comprehensibility and novelty to humans and usefulness for an application.

1 Introduction

The planning of technological and scientific research and development (R&D-) programs is a very demanding task, e.g. in the R&D-program of the German ministry of defense there are at least over 1000 different R&D-projects running simultaneously. They all refer to about 100 different technologies in the context of security and defense. There is always a lot of change in these programs - a lot of projects starting new and a lot of projects running out. One task of our research group is finding new R&D-areas for this program. New ideas or new inventions are a basis for a new R&D-area. That means for planning new R&D-areas it is necessary to identify a lot of new technological ideas and inventions from the scientific community (Ripke et al. (1972)). Up to now, the identification of new ideas and inventions in unstructured texts is done manually (that means by humans) without the support of text mining. Therefore in this paper we will describe the theoretical background of the text mining

approach to discover (semi-) automatically textual patterns representing new ideas and inventions in unstructured technological texts.

Hotho (2004) describes the characteristics that are used to measure the quality of these textual patterns extracted by knowledge discovery tasks. The characteristics are comprehensibility and novelty to the users and usefulness for a task. In this paper the users are program planers or researchers and the task is to find ideas and inventions which can be used as basis for new R&D-areas.

It is known from the cognition research that analysis and evaluation of textual information requires the knowledge of a context (Strube (2003)). The selection of the context depends on the users and the tasks. Referring to our users and our task, we have on one hand textual information about world wide existing technological R&D-projects (furthermore this is called "raw information"). This information contains a lot of new technological ideas and inventions. New means, that ideas and inventions are unknown to the user (Ipsen (2002)). On the other hand we have descriptions about own R&D-projects. This represents our knowledge base and furthermore this is called "context information". Ideas and inventions in the context information are already known to the user.

To create a text mining approach for finding ideas and inventions inside the raw information we have to create a common structure for raw and context information first. This is necessary for the comparison between raw and context information e.g. to distinguish new (that means unknown) ideas and inventions from known ideas and inventions.

In short we have to do 2 steps: 1. Create a common structure for raw and context information as a basis for the text mining approach. 2. Create a text mining approach for finding new, comprehensible and useful ideas and inventions inside the raw information. Below we describe step 1 and 2 in detail.

2 A common structure for raw and context information

In order to perform knowledge discovery tasks (e.g. finding ideas and inventions) it is required that raw information and context information have to be structured and formatted in a common way as described above. In general the structure should be rich enough to allow for interesting knowledge discovery operations and it should be simple enough to allow an automatically converting of all kind of textual information in a reasonable cost as described by Feldman et al.(1995).

Raw information is stored in research papers, articles in technical periodicals, reports, documents, databases, web pages etc. That means raw information contains a lot of different structures and formats. Normally context information also contains different structures and formats. Converting all structures and formats to a common structure and format for raw and context information by keeping all structure information available costs plenty

of work. Therefore our structure approach is to convert all information into plain text format. That means firstly we destroy all existing structures and secondly build up a new common structure for raw and context information.

The new structure should refer to the relationship between terms or term-combinations (Kamphusmann (2002)). In this paper we realize this by creating sets of domain specific terms which occur in the context of a term or a combination of terms. For the structure formulation we define the term unit as word.

First we create a set of domain specific terms.

Definition 1. *Let (a text)* $T = [\omega_1, .., \omega_n]$ *be a list of terms (words)* ω_i *in order of appearance and let* $n \in N$ *be the number of terms in* T *and* $i \in [1, .., n]$*. Let* $\Sigma = \{\tilde{\omega}_1, .., \tilde{\omega}_m\}$ *be a set of domain specific stop terms (Lustig (1986)) and let* $m \in N$ *be the number of terms in* Σ*.* Ω *- the set of domain specific terms in text* T *- is defined as the relative complement* $T without \Sigma$*. Therefore:*

$$\Omega = T \backslash \Sigma \tag{1}$$

For each $\omega_i \in \Omega$ we create a set of domain specific terms which occur in the context of term ω_i.

Definition 2. *Let* $l \in N$ *be a context length of term* ω_i *that means the maximum distance between* ω_i *and a term* ω_j *in text* T*. Let the distance be the number of terms (words) which occur between* ω_i *and* ω_j *including the term* ω_j *and let* $j \in [1, .., n]$*.* Φ_i *is defined as a set of those domain specific terms which occur in an* l*-length context of term* ω_i *in text* T*:*

$$\Phi_i = \{\omega_j | (\omega_j \in \Omega) \wedge (|i - j| \leq l) \wedge (\omega_i \neq \omega_j)\} \tag{2}$$

For each combination of terms in Φ_i we create a set of domain specific terms which occur in the context of this combination of terms.

Definition 3. *Let* $\delta_p \in \Omega$ *be a term in a list of terms with number* $p \in [1, .., \mu]$*. Let* $\delta_1, .., \delta_\mu$ *be a list of terms - in further this will be called term-combination - with* $\delta_p \neq \delta_q \forall p \neq q \in [1, .., \mu]$ *that occurs together in an* l*-length context of term* δ_1 *in text* T *. Let* $\mu \in N$ *be the number of terms in the term-combination* $\delta_1, .., \delta_\mu$*.* $\Xi^T_{\delta_1,..,\delta_\mu}$ *is defined as the set of domain specific terms which occur together with the term-combination* $\delta_1, .., \delta_\mu$ *in an l-length context of term* δ_1 *in text* T*:*

$$\Xi^T_{\delta_1,..,\delta_\mu} = \bigcup \Phi_i \backslash \bigcup_{p=2}^{\mu} \delta_p \Big| \delta_1 = \omega_i \wedge \bigcup_{p=2}^{\mu} \delta_p \subset \Phi_i \tag{3}$$

In the Figure 1 an example for the relationships in set $\Xi^T_{\delta_1,..,\delta_\mu}$ is presented. The term-combination (sensor, infrared, uncooled) has a relationship to the term-combination (focal, array, plane) because uncooled infrared sensors can be built by using the focal plane array technology.

The text T could be a) the textual raw information or b) the textual context information. As result we get in case of a) $\Xi^{raw}_{\delta_1,..,\delta_\mu}$ and in case of b)$\Xi^{context}_{\delta_1,..,\delta_\mu}$.

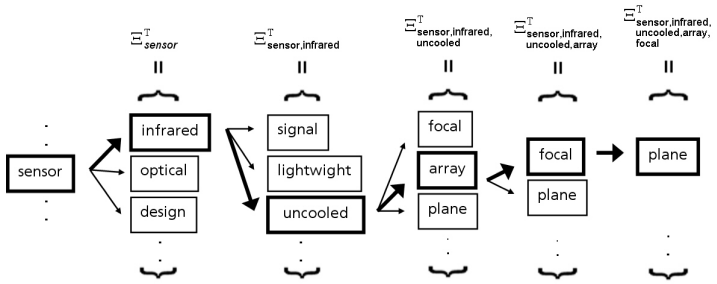

Fig. 1. Example for the relationships in $\Xi^{T}_{\delta_1,..,\delta_\mu}$: Uncooled infrared sensors can be build by using the focal plane array technology.

Definition 4. *To identify terms or term-combinations in the raw information which also occur in the context information - that means the terms or term-combinations are known to the user - we define* $\Xi^{known}_{\delta_1,..,\delta_\mu}$ *as the set of terms which occur in* $\Xi^{raw}_{\delta_1,..,\delta_\mu}$ *and* $\Xi^{context}_{\delta_1,..,\delta_\mu}$*:*

$$\Xi^{known}_{\delta_1,..,\delta_\mu} = \Xi^{raw}_{\delta_1,..,\delta_\mu} \cap \Xi^{context}_{\delta_1,..,\delta_\mu} \tag{4}$$

3 Relevant aspects for the text mining approach from technique philosophy

The text mining approach follows the statements of technique philosophy (Rohpohl (1996)). Below we describe some relevant aspects of the statements and some specific conclusions for our text mining approach.

a) A technological idea or invention represents not only a new mean, but a new purpose and mean combination. That means to find an idea or invention it is necessary to identify a mean and an appertaining purpose in the raw information. Appertaining means that purpose and mean shall occur together in an l-length context. Therefore for our text mining approach we firstly want to identify a mean and secondly we want to identify an appertaining purpose or vice versa.

b) Purposes and means can be exchanged. That means a purpose can become a mean in a specific context and vice versa. Example: A raw material (mean) is used to create an intermediate product (purpose). The intermediate product (mean) is then used to produce a product (purpose). In this example the intermediate product changes from purpose to mean because of the different context. Therefore for our text mining approach it is possible to identify textual patterns representing means or purposes. But it is not possible to distinguish between means and purposes without the

knowledge of the specific context.

c) A purpose or a mean is represented by a technical term or by several technical terms. Therefore purposes or means can be represented by a combination of domain specific terms (e.g. $\delta_1, .., \delta_\mu$) which occur together in an l-length context. The purpose-mean combination is a combination of 2 term-combinations and it also occurs in an l-length context as described in 3 a). For the formulation a term-combination $\delta_1, .., \delta_\mu$ represents a mean (a purpose) only if $\Xi^{raw}_{\delta_1,..,\delta_\mu} \neq \oslash$, which means there are further domain-specific terms representing a purpose (a mean) which occur in an l-length context together with the term-combination $\delta_1, .., \delta_\mu$ in the raw information.

d) To find an idea or invention that is really new to the user, the purpose-mean combination must be unknown to the user. That means a mean and an appertaining purpose in the raw information must not occur as mean and as appertaining purpose in the context information. For the formulation the term-combination $\delta_1, .., \delta_\mu$ from 3 c) represents a mean (a purpose) in a new idea or invention only if $\Xi^{known}_{\delta_1,..,\delta_\mu} = \oslash$, which means there are no further domain-specific terms which occur in an l-length context together with the term-combination $\delta_1, .., \delta_\mu$ in the raw and in the context information.

e) To find an idea or invention that is comprehensible to the user, either the purpose or the mean must be known to the user. That means one part (a purpose or a mean) of the new idea or invention is known to the user and the other part is unknown. The user understand the known part because it is also a part of a known idea or invention that occurs in the context information and therefore he gets an access to the new idea or invention in the raw information.
That means the terms representing either the purpose or the mean in the raw information must occur as purpose or mean in the context information. For the formulation the term-combination $\delta_1, .., \delta_\mu$ from 3 d) represents a mean (a purpose) in a comprehensible idea or invention only if $\Xi^{context}_{\delta_1,..,\delta_\mu} \neq \oslash$, which means $\delta_1, .., \delta_\mu$ is known to the user and there are further domain-specific terms representing a purpose (a mean) which occur in an l-length context together with the term-combination $\delta_1, .., \delta_\mu$ in the context information.

f) Normally an idea or an invention is useful for a specific task. Transferring an idea or an invention to a different task makes it sometimes necessary that the idea or invention has to be changed to become useful for the new task. To change an idea or invention you have to change either the purpose or the mean. That is because the known term-combination $\delta_1, .., \delta_\mu$ from 3

e) must not be changed, otherwise it will become unknown to the user and then the idea or invention is not comprehensible to the user as described in 3 e).

g) After some evaluation we get the experience that for finding ideas and inventions the number of known terms (e.g. representing a mean) and the number of unknown terms (e.g. representing the appertaining purposes) shall be well balanced. Example: one unknown term among many known terms often indicates that an old idea got a new name. Therefore the unknown term is probably not a mean or a purpose. That means the probability that $\delta_1, .., \delta_\mu$ is a mean or a purpose increases when μ is close to the cardinality of $\Xi^{raw}_{\delta_1,..,\delta_\mu}$.

h) There are often domain specific stop terms (like better, higher, quicker, integrated, minimized etc.) which occur with ideas and inventions. They point to a changing purpose or a changing mean and can be indicators for ideas and inventions.

i) An identified new idea or invention can be a basis for further new ideas and inventions. That means all ideas and inventions that are similar to the identified new idea and invention are also possible new ideas and inventions.

4 A text mining approach for finding new ideas and inventions

In this paper we want to create a text mining approach by applying point 3 a) to 3 g). Further we want to prove the feasibility of our text mining approach.

Firstly we want to identify a mean and secondly we want to identify an appertaining purpose below as described in 3 a). The other case - firstly identify a purpose and secondly identify an appertaining mean - is trivial because of the purpose-mean dualism described in 3 b).

Definition 5. *We define $p(\Xi^{raw}_{\delta_1,..,\delta_\mu})$ as the probability that the term-combination $\delta_1, .., \delta_\mu$ in the raw information is a mean. That means whether μ is close to the cardinality of $\Xi^{raw}_{\delta_1,..,\delta_\mu}$ or not as described in 3 g):*

$$p(\Xi^{raw}_{\delta_1,..,\delta_\mu}) = \begin{cases} \frac{\left|\Xi^{raw}_{\delta_1,..,\delta_\mu}\right|}{\mu} & \mu > \left|\Xi^{raw}_{\delta_1,..,\delta_\mu}\right| \\ \frac{\mu}{\left|\Xi^{raw}_{\delta_1,..,\delta_\mu}\right|} & \mu \leq \left|\Xi^{raw}_{\delta_1,..,\delta_\mu}\right| \end{cases} \tag{5}$$

The user determines a minimum probability p_{min}. For the text mining approach the term-combinations $\delta_1, .., \delta_\mu$ are means only if

a) $\Xi^{raw}_{\delta_1,..,\delta_\mu} \neq \oslash$ as described in 3 c),

b) $\Xi^{known}_{\delta_1,..,\delta_\mu} = \oslash$ as described in 3 d) to get a new idea or invention,

c) $\Xi^{context}_{\delta_1,..,\delta_\mu} \neq \oslash$ as described in 3 e) to get a comprehensible idea or invention and

d) $p(\Xi^{raw}_{\delta_1,..,\delta_\mu}) \geq p_{min}$ as described in 3 g).

For each of these term-combinations we collect all appertaining purposes (that means the combinations of all further terms) which occur in an l-length context together with $\delta_1, .., \delta_\mu$ in the raw information.

We present each $\delta_1, .., \delta_\mu$ as a known mean and all appertaining unknown purposes to the user. The user selects the suited purposes for his task or he combines some purposes to a new purpose. That means he changes the purpose to become useful for his task as described in 3 f). Additionally it is possible that the user changes known means to known purposes and appertaining purposes to appertaining means as described in 3 b) because at this point the user gets the knowledge of the specific context.

With this selection the user gets the purpose-mean combination that means he gets an idea or invention. This idea or invention is novel to him because of 3 d) and it is comprehensible to him because of 3 e). Further it is useful for his application because the user selects the suited purposes for his task.

5 Evaluation and Outlook

We have done a first evaluation with a text about R&D-projects from the USA as raw information (Fenner et al. (2006)), a text about own R&D-projects as context information (Thorleuchter (2007)), a stop word list created for the raw information and the parameter values $l = 8$ and $p_{min} = 50\%$. The aim is to find new, comprehensible and useful ideas and inventions in the raw information. According to human experts the number of these relevant elements - the so-called "ground truth" for the evaluation - is eighteen. That means eighteen ideas or inventions can be used as basis for new R&D-areas. With the text mining approach we extracted about fifty patterns (retrieved elements) from the raw information. The patterns have been evaluated by the experts. Thirteen patterns are new, comprehensible and useful ideas or inventions that means thirteen from fifty patterns are relevant elements. Five new, comprehensible and useful ideas or inventions are not found by the text mining approach. Therefore, as result we get a precision value of about 26% and a recall value of about 72%. This is not representative because of the small number of relevant elements but we think this is above chance and it is sufficient to prove the feasibility of the approach.

For future work firstly we will enlarge the stop word list to a general stop word list for technological texts and optimize the parameters concerning the precision and recall value. Secondly we will enlarge the text mining approach with further thoughts e.g. the two thoughts described in 3 h) and 3 i). The aim of this work shall be to get better results for the precision and recall value. Thirdly we will implement the text mining approach to a web based application. That will help the users to find new, comprehensible and useful ideas and inventions with this text mining approach. Additionally with this application it will be easier for us to do a representative evaluation.

6 Acknowledge

This work was supported by the German Ministry of Defense. We thank Joachim Schulze for his constructive technical comments and Jörg Fenner for helping collect the raw and context information and evaluate the text mining approach.

References

FELDMAN, R. and DAGAN, I. (1995): Kdt - knowledge discovery in texts. In: *Proceedings of the First International Conference on Knowledge Discovery (KDD)*. Montreal, 112–113.

FENNER, J. and THORLEUCHTER, D. (2006): Strukturen und Themengebiete der mittelstandsorientierten Forschungsprogramme in den USA. Fraunhofer INT's edition, Euskirchen, 2.

HOTHO, A. (2004): *Clustern mit Hintergrundwissen*. Univ. Diss., Karlsruhe, 29.

IPSEN, C. (2002): *F&E-Programmplanung bei variabler Entwicklungsdauer*. Verlag Dr. Kovac, Hamburg, 10.

KAMPHUSMANN, T. (2002): *Text-Mining*. Symposion Publishing, Düsseldorf, 28.

LUSTIG, G. (1986): *Automatische Indexierung zwischen Forschung und Anwendung*. Georg Olms Verlag, Hildesheim, 92.

RIPKE, M. and STÖBER, G. (1972): Probleme und Methoden der Identifizierung potentieller Objekte der Forschungsförderung. In: H. Paschen and H. Krauch (Eds.): *Methoden und Probleme der Forschungs- und Entwicklungsplanung*. Oldenbourg, München, 47.

ROHPOHL, G. (1996): Das Ende der Natur. In: L. Schäfer and E. Sträker (Eds.): *Naturauffassungen in Philosophie, Wissenschaft und Technik*. Bd. 4, Freiburg, München, 151.

STRUBE, G. (2003): Menschliche Informationsverarbeitung. In: G. Görz, C.-R. Rollinger and J. Schneeberger (Eds.): *Handbuch der Künstlichen Intelligenz*. 4. Auflage, Oldenbourg, München, 23–28.

THORLEUCHTER, D. (2007): Überblick über F&T-Vorhaben und ihre Ansprechpartner im Bereich BMVg. Fraunhofer Publica, Euskirchen, 2–88.