Expert Systems with Applications 40 (2013) 3961–3969

Contents lists available at SciVerse ScienceDirect

Expert Systems with Applications

journal homepage: www.elsevier.com/locate/eswa

Web mining based extraction of problem solution ideas

D. Thorleuchter[a],*, D. Van den Poel[b]

[a] Fraunhofer INT, D-53879 Euskirchen, Appelsgarten 2, Germany
[b] Ghent University, B-9000 Gent, Tweekerkenstraat 2, Belgium

ARTICLE INFO

Keywords:
Web mining
R&D planning
Idea mining
Text mining

ABSTRACT

The internet is a valuable source of information where many ideas can be found dealing with different topics. A few numbers of ideas might be able to solve an existing problem. However, it is time-consuming to identify these ideas within the large amount of textual information in the internet. This paper introduces a new web mining approach that enables an automated identification of new technological ideas extracted from internet sources that are able to solve a given problem. It adapts and combines several existing approaches from literature: approaches that extract new technological ideas from a user given text, approaches that investigate the different idea characteristics in different technical domains, and multi-language web mining approaches. In contrast to previous work, the proposed approach enables the identification of problem solution ideas in the internet considering domain dependencies and language aspects. In a case study, new ideas are identified to solve existing technological problems as occurred in research and development (R&D) projects. This supports the process of research planning and technology development.

1. Introduction

The occurrence of technological problems during the running of research and development (R&D) projects is a well-known challenge for researchers and developers because they have to create problem solution ideas to solve this technological problem (Downey, 2005; Kim, 2012). This task normally is done by use of creativity methods where human experts identify new approaches based on their experiences (Geschka, 1986). A further strategy is to get inspiration from new solution ideas available in the internet. The internet contains a large amount of textual information where nearly each topic is considered and many problem solution ideas can be found there (Razzak, 2012). However, only a few of them are relevant for a current decision problem. The solution ideas that are relevant have to be identified among the large amount of textual information in the internet (Driscoll, 1997). For human experts, it is time consuming to scan all internet sources for these ideas and to identify relevant ideas among them (Shanteau, 1992). Thus, this work proposes a new methodology that enables an automated identification of solution ideas from the internet that are able to solve a given problem.

Some R&D-projects have the aim to find new solutions for existing technological problems (Thorleuchter, Van den Poel, & Prinzie, 2010c). Thus, researchers and developers can be supported with new problem solution ideas from the internet. Further, an ample need for technological problem solution ideas in strategic R&D-planning can be seen. This is because they can be used as starting point for future R&D-projects (Gupta, Raj, & Wilemon, 1986). Thus, strategic research planners also can be supported by establishing new R&D-projects.

Idea mining is introduced in literature to identify solution ideas based on an automatic process (Thorleuchter, Van den Poel, & Prinzie, 2010a). Ideas are extracted from a given text (e.g. scientific publications, patents, research projects). Idea mining is based on technique philosophy where an idea consists of a means and an end. Both, means and ends are represented by text patterns containing several terms. Means and ends that occur together in a description of the problem (furthermore it is named context) are known means and known ends because together, they represent a known idea. Otherwise, if means and ends can be found e.g. in the internet and they do not occur together in the context then they are unknown and this combination of means and ends represent a new idea. The known ideas probably do not work well because they cause problems as described in the context. Thus, idea mining should identify new ideas that can be used to replace the known ideas without causing problems. It is important to know that the new ideas are related to the known ideas because they are applied in the same context. In technique philosophy this relationship is satisfied by identifying a known means that occurs together with an unknown end or a known end that occurs together with an unknown means (Thorleuchter, 2008). Each means and each end consists of a set of technical terms. Similarity

* Corresponding author. Tel.: +49 2251 18305; fax: +49 2251 18 38 305.
E-mail addresses: dirk.thorleuchter@int.fraunhofer.de (D. Thorleuchter), dirk.vandenpoel@ugent.be (D. Van den Poel).

0957-4174/$ - see front matter © 2013 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.eswa.2013.01.013

measures can be used to compare means and ends from the context to means and ends identified in the internet. This enables to distinguish between known and unknown ideas. As a result, unknown ideas that are related to the problem are identified and can be used as problem solution ideas.

The idea mining approach automatically extracts ideas from the new text that is provided by the user but not from textual information from the internet. By transforming the idea mining approach one-to-one to a web mining approach, the new text, which is provided by the user, has to be replaced by textual information from the internet. Thus, a new text that consists of a file size of about several bytes up to several megabytes has to be replaced by textual information of many yottabytes (the total amount of textual information in the internet). This causes performance problems because it is not possible to analyze the total amount of textual information in the internet. Thus, a new idea web mining approach is proposed specifically for the extraction of new ideas in textual information from the internet.

As shown by research results of bibliometrical analyses, innovative problem solution approaches normally cannot be found inside a technological domain but in different domains where new approaches are taken over to solve a given problem (Rhoten & Pfirman, 2007; Schmickl & Kieser, 2008). Literature shows that the quality of the idea mining results strongly depends on the domain. This is because the formulation of ideas in different domains differs (Khare & Chougule, 2012). An example for this is that the number of terms that are used to describe an idea in the Microsystems technology domain is smaller than in the social behavior domain (Thorleuchter & Van den Poel, 2012a). A further example is that domain-specific terms in an idea from Microsystem technology occur more frequently than domain specific terms in an idea from the social behavior domain. Additionally, an equal distribution of known and unknown terms in the given text and in the context are more important for identifying ideas from Microsystem technology than for identifying ideas from the social behavior domain.

In contrast to previous work, this web mining approach considers these aspects by implementing an automated adaption of the different characteristics for idea identification based on the domain. For this, different characteristics are identified for a set of different domains by use of forward selection (Grechanovsky & Pinsker, 1995) as main approach in stepwise regression, by selecting predictive variables with a specific statistical significance and by using a grid search (Jiménez, Lázaro, & Dorronsoro, 2009).

Many case studies have been done to evaluate the idea mining approach. However, they all use the same language for the context and for the given text where a problem solution idea probably can be found. This restricts the approach to the usage of a single language. However, textual information in the internet consists of many different languages and the restriction of a single language forces a researcher and developer to discard a large amount of websites written in further languages (Chau & Yeh, 2004). In contrast to previous work, this web mining approach uses a multi-language approach where problem solution ideas can be identified even if the problem is described in a different language (Aytac, 2005).

In a case study, the proposed web mining approach is applied to improve the strategic planning of the R&D program of the German Ministry of Defense (GE MoD). The program contains about 600 projects with a wide technological scope that are processed to bridge technological gaps. Descriptions about the technological problems standing behind the projects are used to identify new ideas in the internet. An evaluation shows that some of these new ideas can be used as problem solution ideas or as starting point for new projects. This supports the R&D planning of GE MoD.

Overall, the proposed idea web mining approach enables the identification of problem solution ideas from the internet. It considers performance aspects because of the large amount of information in the internet, it considers domain specific characteristics for the identification of ideas, and it considers textual information written in different languages. A web based application is implemented based on this approach that enables an automated identification of new solution ideas. Thus, decision makers can be supported with this application to solve their existing technological problems.

2. Background

2.1. Manual identification of ideas on the internet

The rationale of this approach is based on how persons find technological problem solution ideas on the internet:

After the appearance of a technological problem – e.g. during the realization of a R&D-project – a R&D professional analyzes it so that (s)he recognizes all problem dimensions (Van der Lugt, 2000). After this, the professional searches the internet for documents, which also focus on the same technological problem (Coiera & Vickland, 2008). In this context, (s)he specifically searches for new ideas that probably can be used as problem solution approaches (Willoughby, Anderson, Wood, Mueller, & Ross, 2009).

Normally searching the internet is done by using an internet search engine (Taghavi, Patel, Schmidt, Wills, & Tew, 2012). Here, the professional uses search queries that consist of several domain-specific technical terms, which describe the technological problem (Brand-Gruwel, Wopereis, & Walraven, 2009). Literature estimates the number of these technical terms in a search query that are sufficient to describe the problem at three to five (Koenemann & Belkin, 1996). This is because the use of less than three terms leads to a huge amount of search results where most of them are probably not relevant (Bar-Ilan, 2005). The use of more than five terms leads to a very small number of search results and many relevant documents are probably not found (Clarke, Cormack, & Tudhope, 2000).

After executing these search queries, the internet search engine sorts the query results. The professional expects the ranking algorithm of the search engine to list the relevant documents, which (s)he can use for problem solving with high priority e.g. on the first result page (Mizzaro, 1997). This is because normally a user of a search engine only focuses on the first 10 query results that (s)he finds on the first result page (Jansen, Spink, & Pedersen, 2005; Silverstein, Henzinger, Marais, & Moricz, 1999). If (s)he does not find a new and useful idea as solution approach then (s)he normally modifies her/his search query and starts searching again (Banerji & Magarkar, 2012). Only in seldom cases, (s)he focuses on results of subsequent result pages (Jansen & Spink, 2004).

Each query result consists of a document title, a short description of the document that normally contains search terms from the search query in bold print, and an internet link that refers to the full text. The professional checks these query results for new and useful technological ideas inside the title and inside the short description. If (s)he identifies such a new and useful idea then (s)he follows the corresponding query result link to the full text of the retrieved document. There (s)he extracts the idea manually.

2.2. Idea mining approach

For the extraction of technological ideas, an existing approach is available (Thorleuchter et al., 2010a). It is named idea mining and it is described below.

Idea mining is an approach that has the aim to identify technological ideas from user-provided textual information. The user has to provide two texts: A new text where (s)he supposes the existence of new and useful ideas and a context that consists of a description of technological problems that (s)he wants to solve. Idea mining extracts the ideas from the new text and it evaluates them concerning their ability to solve the problems described in the context. The extraction of ideas is supported by a specific idea mining measure. Idea mining is proposed based on an idea definition from technique philosophy (Thorleuchter, 2008). A web-based application is developed that enables an easy usage of idea mining. The results of idea mining are highlighted text phrases in bold print from the new text. These text phrases represent new and useful problem solution ideas concerning the problems described in the context.

As mentioned before, the idea mining approach uses a new text for the identification of new ideas. Text patterns are built around each word (term) in the new text that is not a stop word and that occurs in the context. The length of text patterns is calculated by a term-weighting schema based on the difference between stop words and non-stop words. For each text pattern, all corresponding text patterns from the context are identified. These are text patterns around the same selected term. By comparing text patterns from new text with specially selected corresponding text patterns from context, known terms can be identified, which occur both in a text pattern from new text and in one of its corresponding text patterns.

Additionally unknown terms can be identified, which only occur in a text pattern from new text but never in one of its corresponding text patterns. An idea mining measure is used to enable a classification decision. This decision selects a text pattern as a new and useful idea if

– the number of known terms and the number of unknown terms are well balanced,
– known terms occur more frequently in context than other terms,
– unknown terms occur more frequently in new text than other terms, and
– specific terms occur, which are characteristic for a new and useful idea.

Several parameters are used for this classification decision:

Let α be defined as a set of stemmed terms from a text pattern of the new text, let β be defined as a set of stemmed terms from a text pattern of the context. To fulfill a restriction as mentioned in Thorleuchter et al. (2010a), it is assumed for further processing that α and β have at least one term in common. Let $p = |\alpha|$ be defined as the number of terms in α and let $q = |\alpha \bigcap \beta| \neq \varnothing$ be defined as the number of terms existing in both sets. Then, the first sub measure m_1 is defined as

$$m_1 = \begin{cases} \frac{2\cdot(p-q)}{p} & (q \geqslant \frac{p}{2}) \\ \frac{2\cdot q}{p} & (q < \frac{p}{2}) \end{cases} \tag{1}$$

For the calculation of m_2, the z% most frequent terms from the context are selected as set δ. The number of selected terms that are both in α and β are used as numerator ($r = |\alpha \bigcap \beta \bigcap \delta|$) and the number of all terms in α and β is used as denominator.

$$m_2 = \frac{r}{q} \tag{2}$$

In contrast to set δ, the calculation of m_3 uses set φ that contains the z% most frequent terms from the new text. The number of these terms that are not in β is used in the numerator ($s = |\alpha \cap \bar{\beta} \cap \varphi|$) and the number of all terms in α that are not in β is used as denominator.

$$m_3 = \frac{s}{p-q} \tag{3}$$

A set of characteristic terms is used to calculate the fourth sub measure. The occurrence of one of these terms in α (as indicated by $t = |\alpha \bigcap \lambda|$) sets m_4 to one while the non-occurrence of these terms sets m_4 to zero.

$$m_4 = \begin{cases} 1 & (t > 0) \\ 0 & (t = 0) \end{cases} \tag{4}$$

Weighting factors are used for each sub measure to calculate the idea mining measure m:

$$m = \begin{cases} g_1 m_1 + g_2 m_2 + g_3 m_3 + g_4 m_4 & (p \neq q) \\ 0 & (p = q) \end{cases} \tag{5}$$

A threshold $\hat{a}$ is defined that classifies a text pattern as a new idea if the corresponding idea mining measure m is greater than or equal to $\hat{a}$. For the calculation of the text patterns, the integer value l and the percentages u and v are used. Each non-informative term (e.g. a stop word) is assigned to a term weight u and all further terms are assigned to a term weight v. Text patterns are built around each term that occurs in the new text as well as in the context. Terms that appear directly before or after the terms in a text patterns also are included if the sum of all term weights from all terms in the text pattern is smaller than l.

2.3. Idea characteristics for different technological domains

Literature shows that the characteristics of an idea depend on the domain and it introduces approaches for identifying these characteristics specifically for idea mining (Thorleuchter, Herberz, & Van den Poel, 2012a; Thorleuchter & Van den Poel, 2012a). The different idea characteristics are represented by parameters of the idea mining approach. Parameters can be distinguished between two categories: parameters that are used to identify a text pattern and parameters that are used to calculate the idea mining measure. Text patterns are built based on the length l and the term weights u and v. The idea mining measure is applied using the six parameters (g_1, g_2, g_3, g_4, $\hat{a}$, and z) as indicated in Section 2.2. All nine parameters impact the performance of the idea mining approach.

A set of parameter values has to be identified that results in the highest classification performance. For this, variable selection (Coussement & Van den Poel, 2008) is used to order all parameters based on their impact on the performance as calculated by their $\chi 2$-statistic. The parameters (highest impact first) are selected one-by-one until parameter's impact is below a specific threshold (Van den Poel & Buckinx, 2005). This forward-selection procedure reduces the number of parameters. This enables calculating an optimized value for each of the selected parameters without high computationally costs. A grid search (Basrak, 1987) is applied where discrete sequences for the parameter values are used. These values are selected that results in the highest cross-validated F-measure (Guns, Lioma, & Larsen, 2012).

3. Methodology

3.1. Overview

This idea web mining approach has the aim to realize the rationale in Section 2.1 by use of a methodology (see Fig. 1). We use the existing idea mining approach to extract new and useful technological ideas from provided textual information as described in Section 2.2. This idea mining approach is extended specifically for identifying new ideas from the internet. In contrast to the extraction of ideas from a given text, relevant ideas from the

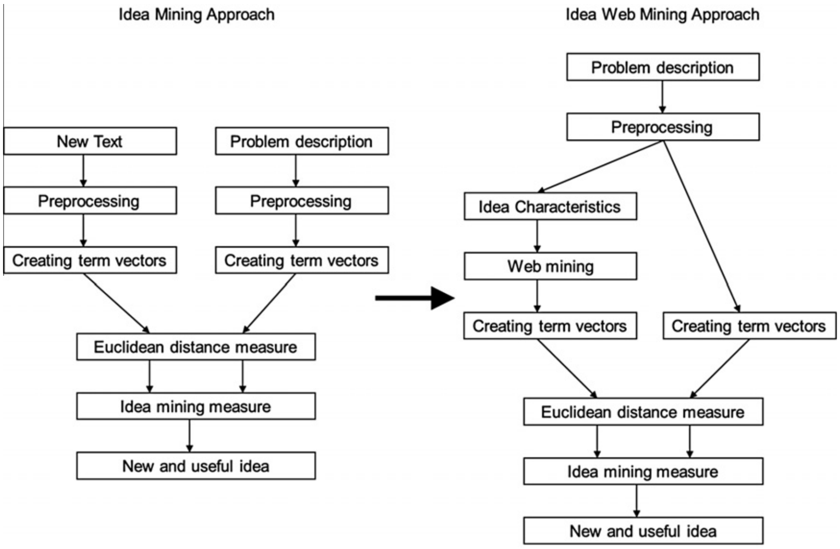


Fig. 1. The processing of this idea web mining approach (right figure) based on the processing of the idea mining approach (left figure). The user provides a context (problem description). After preprocessing, term vectors are created and search queries are built representing the problem. The queries are executed by use of a web search engine and term vectors are created from the query results based on the domain specific idea characteristics. They are compared to term vectors from the context by use of the Euclidean distance measure and the idea mining measure. As a result, new and useful problem solution ideas are extracted from the internet.

internet probably stem from different domains. The relevant domains are identified and research that identifies the idea characteristics in different domains is considered as described in Section 2.3. Further, ideas written in a different language also are extracted based on research results of multi-language web mining.

The first step in the introduced approach is to provide a context that consists of a description of a technological problem. In this description, domain specific terms occur that belong to specific technologies and to application fields (technological domains) (Thorleuchter & Van den Poel, 2012f). These terms are identified by a preprocessing step where text preparation and term filtering is done. Technologies and application fields that are related to the technological domains of the context are identified and described. The idea characteristics for the related domains are calculated. Groups of domain specific terms are built considering the co-occurrences of these terms. Each group is used as a search query for web mining. The search queries are translated to several languages, executed, and a text pattern is extracted from the full text of each query result item. The text patterns are translated to the target language. Terms from the text patterns – excluding the terms from the corresponding search query – are used to identify the technological domain of the new idea. This is done by comparing these terms to terms from the description of the related domains. The idea characteristics from the domain with the largest similarity are selected and the term vectors from the new ideas are created. They are compared to its most similar term vectors that represent text patterns from the problem context by use of Euclidean distance measure and the idea mining measure. As a result, new and useful problem solution ideas are extracted from the internet.

3.2. Preprocessing

The aim of the preprocessing step is to identify relevant terms within a provided text that could be used to represent the text.

Based on the provided text e.g. a description of a technological problem, the raw text is cleaned by deleting images as well as tags, scripting code, punctuation, and specific characters (Thorleuchter & Van den Poel, 2013c). Tokenization is applied to separate different terms where a term unit normally is defined as a word. Terms are compared to a dictionary and identified typographical errors are corrected. The used case conversion aims to a capitalized first sign and all further signs of a word are written in lower case (Thorleuchter & Van den Poel, 2012c).

Term filtering methods are applied to distinguish between relevant terms in contrast to non-informative terms. Part-of-speech tagging enables this by considering the syntactic category of terms e.g. retaining nouns and verbs and discarding adjectives and adverbs. Stop word filtering also enables this by comparing terms to a stop word list.

Further term filtering methods group related terms together by considering their stems (Thorleuchter & Van den Poel, 2012d). Stemming enables this by applying a set of production rules while lemmatizing uses a dictionary based assignment of terms to their stems. The proposed approach uses stemming because lemmatizing is error prone and computationally expensive.

A last step in term filtering could be a manual evaluation of the selected terms by human experts (Gericke, Thorleuchter, Weck, Reiländer, & Loß, 2009). This improves quality of the results however it is time consuming.

3.3. Idea characteristics

The technological problem described in the context contains terms that belong to a technological domain. However, many solution ideas for this problem stem from different technological domains. This is because approaches that have been approved in a different domain probably are suited to solve the existing problem. Literature shows that the characteristics of ideas in different domains are different (Thorleuchter et al., 2012a). The characteristics are represented by the parameters of the idea mining approach. Thus, this step identifies the idea characteristics from different domains.

Relevant domains that have to be considered for this task are domains that are related to the domain from the technological problem (Thorleuchter, Van den Poel, & Prinzie, 2010b). The selection of related technological domains is done by considering existing lists of technologies and application fields. A description for each selected domain is created that contains the corresponding domain specific terms. The description is used to assign the web mining results to domains. Term vectors are created for the descriptions based on vector space model (D'Haen, Van den Poel, & Thorleuchter, 2013).

For each domain, the nine parameters (l, u, v, $g1$, $g2$, $g3$, $g4$, $\hat{a}$, and z) are calculated based on the methodology described in Section 2.3 that consists of variable selection, grid search, and n-fold cross-validation.

3.4. Web mining

The web mining step is used to identify internet websites where relevant ideas are described. To access this textual information, the corpora of internet search engines are used. In general, internet search engines offer web services in form of advanced programming interfaces (Thorleuchter & Van den Poel, 2013b). By use of these interfaces, queries are executed automatically and the results that consists of a title, a short description, and a hyperlink are integrated into own applications.

The search queries are created based on the identified relevant terms from the preprocessing step. Section 2.1 shows that the numbers of terms in a search query that are sufficient to describe the problem are three to five. Thus, four terms as the mean value are selected for creating each query. The selection of terms is done by considering co-occurrences of terms: The distance between two terms is measured by the smallest number of relevant terms that occur between the appearances of the two terms. Terms that are discarded in the preprocessing step are not considered for this calculation. Then, two terms co-occur if the distance between them is below a specific threshold. For each relevant term, all its co-occurrences are identified and all combinations of four relevant terms where each term co-occurs with all three further terms are selected as search query (Thorleuchter & Van den Poel, 2011b).

Searching for documents in different language normally requires the use of terms in a search query that are written in the corresponding language. Google translate advanced programming interface (API) is used that offers an automated translation of terms (Thorleuchter, Schulze, & Van den Poel, 2012b). The interface contains a large number of languages and it translates terms from the target language (English) to these languages. For each translated query, the search results are restricted to websites written in the corresponding language.

All created queries are executed and the short descriptions of the results are sorted by different languages. Google translate API is used to translate the text patterns to the target language. Terms from the short description are selected excluding translated terms from the search query and excluding terms from a stop word list. Then, a term vector is created and it is compared to term vectors from the domain descriptions.

As a result, the technological domain standing behind each query result can be identified and the corresponding set of parameter values is selected. The full texts of the query results are crawled and the set of parameter values is used to extract the text pattern. Terms in the pattern are translated by use of Google translate API.

3.5. Creating and comparing term vectors

Term vectors are created from both, the text patterns from the preprocessed context and the text patterns from the extracted full text of the query results. Comparing is done by use of the Euclidean distance measure and the idea mining measure. The corresponding set of parameter values is considered. A detailed description of this step is described in Thorleuchter et al. (2010a).

As a result, new ideas are identified that can be used to solve a specific problem.

4. Case study

4.1. Overview

In a case study, the German Ministry of Defense (GE MoD) is supported by this web mining based idea extraction. The case study identifies new technological ideas for the German defense research program based on existing R&D projects (Thorleuchter & Van den Poel, 2012e). Thus as context information, we use 100 descriptions of R&D projects dealing with different technologies in different application fields (see Table 1). All projects are granted in 2006 by the GE MoD. Thus, the descriptions contain the technological limitations at that time. At present, many new technological approaches have been proposed in literature and thus, in the internet. Some of them are able to improve R&D results from 2006. The proposed approach is used to identify these new technologies from the internet. An evaluation is done by human experts that compare the identified new technologies to the currently used technologies in successive R&D projects.

Idea characteristics are identified for a list of related technologies in Section 4.2. Web mining is processed (see Section 4.3) based on the idea characteristics. The results are converted to term vectors and they are compared to term vectors from the context. As a result of this case study, several new ideas are extracted from the internet that are useful for German defense research planners because they can be used to solve technological problems in the selected R&D projects from 2006 or they can be used as starting point for successive R&D projects. Some successful examples of the results are shown in Section 4.4. The results are evaluated in Section 4.5 by use of performance measures from information retrieval.

4.2. Idea characteristics

The selected R&D projects investigate some technologies to use them in the application field defense. New ideas might stem from further technologies that also contribute to the application field defense. A collection of technologies related to defense is published as taxonomy by the European Defense Agency (EDA) (Thorleuchter & Van den Poel, 2011a). A list of 12 interesting technologies is selected from the EDA taxonomy (see Table 2). For each technology, several descriptions are available (Thorleuchter & Van den Poel, 2013a) and used for the calculation of an optimized set of parameter values for each technology.

The methodology as described in Sections 2.3 and 3.3 is applied for each technology separately. It shows that the parameters u, v, and g_4 are of low predictive performance for all technologies. They are discarded for further processing and they are set to a standard value ($u = v = 100\%$; $g_4 = 0$) that makes them comparable to results of further studies.

The results of this idea characteristics step are the optimized values for the idea mining parameters as presented in Table 3.

The results show idea that characteristics in the technologies 3 and 13 are related to idea characteristics in the social behavior domain because the corresponding parameter values are similar. Further the technology eight lies between the social behavior domain and the medical domain and technology 10 lies between the social behavior domain and the Microsystem technology domain. The further nine technologies are related to the Microsystems technology domain rather than to the social behavior domain and the medical domain. This is because they all contribute to material sciences as a common basis and material sciences is more similar to Microsystem technology than to the other two domains.

The results are evaluated using the F-measure that is obtained by applying the corresponding parameter values in the idea mining

Table 1
Characteristics of the data.

Number of R&D project descriptions	100
Average number of search queries per project	30
Number of search queries in total	2.985
Search query results per search query	10
Search query results in total (after validating e.g. deleting double entries)	24.184

Table 2
List of technology areas from EDA taxonomy of defense-based technologies.

Number	Technology area
01	Lethality & platform protection
02	Propulsion and power plants
03	Design technologies for platforms and weapons
04	Electronic warfare and directed energy technologies
05	Signature control and signature reduction
06	Sensor systems
07	Guidance and control systems for weapons and platforms
08	Simulators, trainers and synthetic environments
09	Integrated systems technology
10	Communications and CIS-related technologies
11	Personnel protection systems
12	Manufacturing processes/design tools/techniques
13	Human sciences

Table 3
Results of the idea characteristics step.

Number	g_1	g_2	g_3	$\hat{a}$	l	z	F-measure (%)
01	0.40	0.30	0.30	0.50	6	0.05	32
02	0.40	0.30	0.30	0.40	8	0.10	38
03	0.20	0.40	0.40	0.30	9	0.15	35
04	0.40	0.30	0.30	0.40	8	0.20	34
05	0.40	0.30	0.30	0.40	8	0.10	36
06	0.40	0.30	0.30	0.50	8	0.20	32
07	0.40	0.30	0.30	0.50	7	0.10	30
08	0.30	0.35	0.35	0.20	10	0.20	29
09	0.40	0.30	0.30	0.50	8	0.10	34
10	0.30	0.35	0.35	0.30	8	0.10	32
11	0.40	0.30	0.30	0.50	8	0.15	30
12	0.40	0.30	0.30	0.40	10	0.10	34
13	0.20	0.40	0.40	0.20	11	0.10	32

approach. Further, a baseline is used that is already introduced in the evaluation of Thorleuchter et al. (2012a). There, the baseline is calculated by using a heuristic similarity measure that replaces the idea mining measure. Based on a precision value of 30% and based on a recall value of 20%, the F-measure for the baseline is 25%. Thus, the F-measure results in Table 3 outperform the baseline.

4.3. Web mining

This web mining approach creates even from a large context a very large number of text patterns. For each of these text patterns, several search queries are created. Additionally, each search query leads to several query results. Building term vectors for all these query results and comparing them to all term vectors from the context by using the Euclidean distance measure and the idea mining measure is computationally expensive. To prevent these performance problems, this approach aims to reduce the number of the selected query results in total.

The methodology is applied as described in Section 3.4. For each search query, the first 10 query results are considered. This is because as described in Section 2.1, a person focuses on the first ten query results that (s)he finds on the first result page.

For each query result from the search engine, we do not focus on title and hyperlink but on the short description (see Fig. 2). 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 (Thorleuchter, Van den Poel, & Prinzie, 2010d). One aim of this web mining step is to identify a text pattern that contains all terms from the search query. Thus, these short descriptions are selected that contain the terms from the search query in one text pattern. The others are discarded.

For each selected short description, the different terms are extracted using methods from the preprocessing step. The terms (excluding the terms from the search query) are compared to the terms from the description of the technology area. As a result, each short description is assigned to a technology area and the corresponding parameter values are used for further processing.

The parameter values contain information about the length of a text pattern (variable l, u, and v). It is important to know that the length of the text pattern from the short description is often not large enough. Thus, it is important to extract the extended text pattern from the full text. For each of the selected short description, a self-developed web crawler uses the corresponding hyperlink to crawl the full text of the query result (Thorleuchter & Van den Poel, 2012b). The position of the text pattern from the short description in the full text is identified. The text pattern is extracted from the full text by using the variable l, u, and v. Then, the text pattern is converted to a term vector and it is compared to term vectors from the context by use of the Euclidean distance and the idea mining measure. For this, the corresponding variable values for g_1, g_2, g_3, g_4, $\hat{a}$, and z are used.

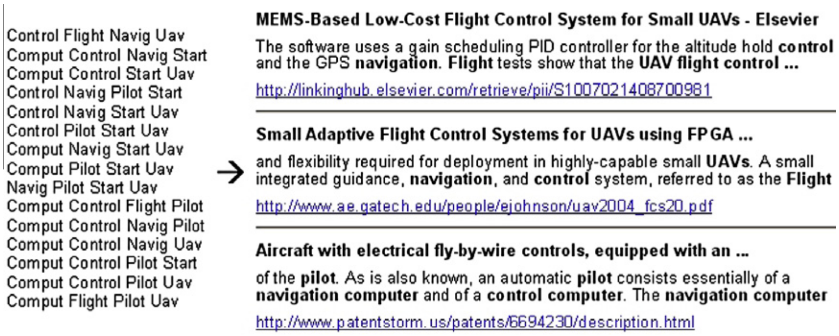
4.4. Examples

Some example results that are able to advance the selected R&D projects from 2006 are presented below:

High-strength aluminum alloys for the use in military aircrafts are investigated to reduce aircraft's weight. The proposed idea web mining approach identifies magnesium based light metals as new solution idea. Magnesium has a one third lower density than aluminum. Current scientific developments from material sciences e.g. an extremely fine grain, processes of severe plastic deformation or rapid solidification lead to an overcome of existing limitations (low strength, low toughness and poor forming capability). Thus, a recommendation for the R&D planning can be given to start a successive R&D project that investigates the use of magnesium in military aircrafts instead of aluminum alloys.

For high frequency (HF) communication (between 3 and 30 MHz) the used antennas have to be at a specific length. Especially on naval ships where the ship size is limited, the antennas have to be dampened by resistors to become smaller. The resistors commonly consist of discrete circuits (in contrast to integrated circuits). Existing limitations e.g. the large size of the discrete circuits, low weather resistance, low power consumption, and low resonance suppression are investigated. Idea web mining identifies the use of resistors with amorphous materials (e.g. cellulose or fiberglass) as possible problem solution idea. In the internet, current advances in amorphous material are described that show that these materials possibly can be used to overcome the described limitations.

A communication system for the use in space e.g. between satellites is investigated. The use of conventional communication systems in space leads to several disadvantages. A high data rate (up to 10 Gbit/s) over a high link distance (up to 8000 km) is necessary



Control Flight Navig Uav
Comput Control Navig Start
Comput Control Start Uav
Control Navig Pilot Start
Control Navig Start Uav
Control Pilot Start Uav
Comput Navig Start Uav
Comput Pilot Start Uav
Navig Pilot Start Uav
Comput Control Flight Pilot
Comput Control Navig Pilot
Comput Control Navig Uav
Comput Control Pilot Start
Comput Control Pilot Uav
Comput Flight Pilot Uav

→

MEMS-Based Low-Cost Flight Control System for Small UAVs - Elsevier
The software uses a gain scheduling PID controller for the altitude hold **control** and the GPS **navigation**. **Flight** tests show that the **UAV flight control** ...
http://linkinghub.elsevier.com/retrieve/pii/S1007021408700981

Small Adaptive Flight Control Systems for UAVs using FPGA ...
and flexibility required for deployment in highly-capable small **UAVs**. A small integrated guidance, **navigation**, and **control** system, referred to as the **Flight**
http://www.ae.gatech.edu/people/ejohnson/uav2004_fcs20.pdf

Aircraft with electrical fly-by-wire controls, equipped with an ...
of the **pilot**. As is also known, an automatic **pilot** consists essentially of a **navigation computer** and of a **control computer**. The **navigation computer**
http://www.patentstorm.us/patents/6694230/description.html

Fig. 2. Example for the processing of the web mining step. Search queries are built based on the co-occurrence of relevant terms from the context. They are executed by Google Web APIs. Query results consist of a title in bold print, a hyperlink for access to the full text, and a short description where terms from the search query are also printed in bold.

and the communication components have to be transported to space. Thus, a low mass, a small volume and low power dissipation is also important. Idea web mining identifies laser communication technology as solution idea. Current advances in laser communication technology show that sizes and energy consumption of newly developed components can be reduced significantly. As laser communication technology is always able to realize a high data rate over a high link distance, it might be a valuable solution approach.

4.5. Evaluation

This paper proposes a web mining approach that enables an automatic identification of new and useful technological ideas from the internet. It is related to the idea mining approach that is already evaluated in Thorleuchter et al. (2010a). Thus, this evaluation focuses specifically on the performance of the extensions of the approach: the impact of the web mining step and the impact of the idea characteristics step on the idea mining results (the identification of new and useful ideas).

This idea web mining approach is compared to a baseline model. It uses the frequency baseline because a high percentage of extracted query results do not represent a new and useful idea. Two classes are defined (*A:* a query result represents a new and useful idea, *B:* a query result does not represent a new and useful idea) and each instance (query result) with a specific percentage is classified as either *A* or *B*. This comparing enables to show the feasibility of the approach. A comparing with further models is not done because we are not aware of other approaches for an automated identification of new and useful ideas from the internet at the present time.

To compute the percentage for the frequency baseline, further descriptions from R&D projects are used. The proposed methodology is applied to get the number of queries from each description. For each query, the first 10 query results are considered. Thus, the number of queries is multiplied by 10 to get the number of query results for each description. Then, for each description, the number of new and useful ideas is divided by the number of query results. After this, a percentage *x* is obtained. It says that *x*% of all query results represent a new and useful problem solution idea. Then, the average percentage for all R&D projects is calculated. As a result, 3% of all query results represents a new and useful idea. Therefore, we set the frequency baseline to 3%.

This idea web mining approach is processed twice: The first way is to use the proposed methodology without applying the idea characteristics step. The second way is to use the methodology as presented. This enables to show the impact of the idea characteristics step on the proposed methodology.

A web mining application is created that realizes the proposed methodology. It is available at http://www.text-mining.info ("Idea Web Miner") and it can be used for further evaluation, too. The application automatically extracts ideas from the internet and presents them to the user and it is used in two versions (with an enabled idea characteristics step and with a disabled idea characteristics step).

The evaluation of the results is based on precision and recall measures that are commonly used in information retrieval (Thorleuchter, Van den Poel, & Prinzie, 2012a). These measures use the number of true positives, false positives, and false negatives results. The ground truth in this evaluation is received by human experts in two different ways. First, they are aware of currently used technologies in the successive R&D projects of the selected R&D projects from 2006. These technologies are successful problem solution approaches from point of view of 2006. Second, human experts use the context to search manually in the internet for new problem solution ideas. Additionally, (s)he uses the results of this idea web mining approach to find further new and useful ideas in the internet. This enables human experts to calculate the value of true positives, false positives, and false negatives for the idea web mining results.

For the results of the proposed methodology, we get a precision value of 15% at a recall value of 30%. A precision value of 15% means that if this approach predicts 100 ideas as new and useful than 15 of them are really new and useful ideas. A recall value of 30% means that if there are 10 new and useful ideas in the internet then this approach identifies three of them. For the results of the proposed methodology without using the idea characteristic step, a precision value of 10% at a recall value of 30% is obtained. The difference in the F-score (22.5–20%) confirms research results from literature where it is stated that the characteristics of ideas depend on the domain.

To see whether these results are good or bad, they are compared to the frequency baseline. The baseline has a precision value of 3% at a recall value of 30% and an F-score of 16.5%. Thus, the values of the baseline are below that of the proposed methodology. This shows that the web idea mining approach can be used to support persons by finding technological problem solution ideas from the internet.

5. Conclusion

This paper introduces a new idea web mining approach that automatically identifies new technological ideas extracted from the internet. Existing approaches for identifying ideas from texts are adapted to the use of the internet. An automated assignment of ideas to different domains is realized that enables the extraction of ideas from the full text of internet documents by use of an optimized set of parameter values.

In a case study, it is shown that this idea extraction from the internet is successful and that the results are valuable for R&D professionals to improve research planning and technological development.

For future work, the used knowledge based classification approach should be replaced with a semantic classification approach. Well-known methods e.g. latent semantic indexing or non-negative matrix factorization probably are suited to identify new ideas with a better performance than the proposed approach because they consider aspects of meaning rather than aspects of words. A further avenue of research focuses on the automated translation of extracted text patterns from websites. This translation has a large impact on the performance of the approach. The proposed approach uses Google translate API. For an automated translation, the results of Google translate API are of good quality. However compared to a manual translation by human experts, the results sometimes are very poor. Further work should focus on improving the quality of the automated translation by considering domain-specific knowledge. This is possible because related domains are identified for each text pattern that should be translated. Technical terms from the different domains can be translated by human experts in advance and they can be used to improve results of the automated translation.

References

Aytac, S. (2005). Multilingual information retrieval on the internet: A case study of Turkish users. *The International Information & Library Review, 37*(4), 275–284.

Banerji, A., & Magarkar, A. (2012). How happy is your web browsing? A model to quantify satisfaction of an Internet user searching for desired information. *Physica A: Statistical Mechanics and its Applications, 391*(17), 4215–4224.

Bar-Ilan, J. (2005). Comparing rankings of search results on the Web. *Information Processing & Management, 41*(6), 1511–1519.

Basrak, Z. (1987). A routine for parameter optimization using an accelerated grid-search method. *Computer Physics Communications, 46*(1), 149–154.

Brand-Gruwel, S., Wopereis, I., & Walraven, A. (2009). A descriptive model of information problem solving while using internet. *Computers & Education, 53*(4), 1207–1217.

Chau, R., & Yeh, C. (2004). A multilingual text mining approach to web cross-lingual text retrieval. *Knowledge-Based Systems, 17*(5–6), 219–227.

Clarke, C. L. A., Cormack, G. V., & Tudhope, E. A. (2000). Relevance ranking for one to three term queries. *Information Processing & Management, 36*(2), 291–311.

Coiera, E. W., & Vickland, V. (2008). Is relevance relevant? User relevance ratings may not predict the impact of internet search on decision outcomes. *Journal of the American Medical Informatics Association, 15*(4), 542–545.

Coussement, K., & Van den Poel, D. (2008). Integrating the voice of customers through call center emails into a decision support system for churn prediction. *Information & Management, 45*(3), 164–174.

D'Haen, J., Van den Poel, D., & Thorleuchter, D. (2013). Predicting customer profitability during acquisition: finding the optimal combination of data source and data mining technique. *Expert Systems with Applications, 40*(6), 2007–2012.

Downey, G. (2005). Are engineers losing control of technology? From 'problem solving' to 'problem definition and solution' in engineering education. *Chemical Engineering Research and Design, 83*(6), 583–595.

Driscoll, J. R. (1997). Method and system for searching for relevant documents from a text database collection, using statistical ranking, relevancy feedback and small pieces of text. *Laboratory Automation & Information Management, 33*(2), 150.

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. (1986). From experience. Creativity workshops in product innovation. *Journal of Product Innovation Management, 3*(1), 48–56.

Grechanovsky, E., & Pinsker, I. (1995). Conditional p-values for the F-statistic in a forward selection procedure. *Computational Statistics & Data Analysis, 20*(3), 239–263.

Guns, R., Lioma, C., & Larsen, B. (2012). The tipping point: F-score as a function of the number of retrieved items. *Information Processing & Management, 48*(6), 1171–1180.

Gupta, A. K., Raj, S. P., & Wilemon, D. (1986). A model for studying R&D-marketing interface in the product innovation process. *Journal of Marketing, 50*, 7–17.

Jansen, B. J., & Spink, A. (2004). An analysis of web searching by European Alltheweb.com users. *Information Processing and Management, 41*(6), 361–381.

Jansen, B. J., Spink, A., & Pedersen, J. (2005). A temporal comparison of AltaVista Web searching. *Journal of the American Society for Information Science and Technology, 56*(6), 559–570.

Jiménez, A. B., Lázaro, J. L., & Dorronsoro, J. R. (2009). Finding optimal model parameters by deterministic and annealed focused grid search. *Neurocomputing, 72*(13–15), 2824–2832.

Khare, V. R., & Chougule, R. (2012). Decision support for improved service effectiveness using domain aware text mining. *Knowledge-Based Systems, 33*, 29–40.

Kim, M. K. (2012). Cross-validation study of methods and technologies to assess mental models in a complex problem solving situation. *Computers in Human Behavior, 28*(2), 703–717.

Koenemann, J., & Belkin, N. J. (1996). A case for interaction: a study of interactive information retrieval behavior and effectiveness. In *ACM SIGCHI conference on human factors in computing systems* (pp. 205–212). Vancouver: ACM.

Mizzaro, S. (1997). Relevance. The whole history. *Journal of the American Society for Information Science, 48*(9), 810–832.

Razzak, F. (2012). Spamming the internet of things: A possibility and its probable solution. *Procedia Computer Science, 10*, 658–665.

Rhoten, D., & Pfirman, S. (2007). Women in interdisciplinary science. Exploring preferences and consequences. *Research Policy, 36*, 56–75.

Schmickl, C., & Kieser, A. (2008). How much do specialists have to learn from each other when they jointly develop radical product innovations? *Research Policy, 37*, 1147–1163.

Shanteau, J. (1992). How much information does an expert use? Is it relevant? *Acta Psychologica, 81*(1), 75–86.

Silverstein, C., Henzinger, M., Marais, H., & Moricz, M. (1999). Analysis of a very large web search engine query log. *ACM SIGIR Forum, 33*(1), 6–12.

Taghavi, M., Patel, A., Schmidt, N., Wills, C., & Tew, Y. (2012). An analysis of web proxy logs with query distribution pattern approach for search engines. *Computer Standards & Interfaces, 34*(1), 162–170.

Thorleuchter, D., & Van den Poel, D. (2013a). Technology classification with latent semantic indexing. *Expert Systems with Applications, 40*(5), 1786–1795.

Thorleuchter, D., & Van den Poel, D. (2013b). Protecting research and technology from espionage. *Expert Systems with Applications*, http://dx.doi.org/10.1016/j.eswa.2012.12.051.

Thorleuchter, D., & Van den Poel, D. (2013c). Analyzing Website Content for Improved R&T Collaboration Planning. In *World conference on information systems and technologies. Advances in intelligent systems and computing*. Berlin: Springer.

Thorleuchter, D., Herberz, S., & Van den Poel, D. (2012a). Mining social behavior ideas of przewalski horses. *Lecture Notes in Electrical Engineering, 121*, 649–656.

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., Schulze, J., & Van den Poel, D. (2012b). Improved emergency management by loosely coupled logistic system. *Communications in Computer and Information Science, 318*, 5–8.

Thorleuchter, D., & Van den Poel, D. (2011a). Semantic technology classification – A defence and security case study. In *Proceedings uncertainty reasoning and knowledge engineering* (pp. 36–39). New York: IEEE.

Thorleuchter, D., & Van den Poel, D. (2011b). Companies website optimising concerning consumer's searching for new products. In *Proceedings uncertainty reasoning and knowledge engineering* (pp. 40–43). New York: IEEE.

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., & Van den Poel, D. (2012d). Using webcrawling of publicly-available websites to assess e-commerce relationships. In *SRII global conference 2012* (pp. 402–410). San Jose, CA, USA: IEEE.

Thorleuchter, D., & Van den Poel, D. (2012e). Improved multilevel security with latent semantic indexing. *Expert Systems with Applications, 39*(18), 13462–13471.

Thorleuchter, D., & Van den Poel, D. (2012f). Rapid Scenario Generation with Generic Systems. In *International conference on management sciences and information technology. Lecture Notes in Information Technology* (pp. 87-91). Delaware: IERI.

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. (2010d). Extracting consumers needs for new products – A web mining approach. In *Proceedings WKDD 2010* (pp. 441). Los Alamitos: IEEE Computer Society.

Thorleuchter, D., Van den Poel, D., & Prinzie, A. (2012a). 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., 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.

Van den Poel, D., & Buckinx, W. (2005). Predicting online-purchasing behavior. *European Journal of Operational Research, 166*(2), 557–575.

Van der Lugt, R. (2000). Developing a graphic tool for creative problem solving in design groups. *Design Studies, 21*(5), 505–522.

Willoughby, T. S., Anderson, A., Wood, E., Mueller, J., & Ross, C. (2009). Fast searching for information on the internet to use in a learning context: The impact of domain knowledge. *Computers & Education, 52*(3), 640–648.