Expert Systems with Applications 37 (2010) 7182–7188

Contents lists available at ScienceDirect

Expert Systems with Applications

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

Mining ideas from textual information

Dirk Thorleuchter [a,*,1], Dirk Van den Poel [b], Anita Prinzie [b,c]

[a] *Fraunhofer INT, D-53879 Euskirchen, Appelsgarten 2, Germany*
[b] *Ghent University, B-9000 Gent, Tweekerkenstraat 2, Belgium*
[c] *Visiting Researcher, Manchester Business School, The University of Manchester, Manchester M15-6PB, Booth Street West, UK*

ARTICLE INFO

Keywords:
Idea mining
Text mining
Text classification
Technology

ABSTRACT

This approach introduces idea mining as process of extracting new and useful ideas from unstructured text. We use an idea definition from technique philosophy and we focus on ideas that can be used to solve technological problems.

The rationale for the idea mining approach is taken over from psychology and cognitive science and follows how persons create ideas. To realize the processing, we use methods from text mining and text classification (tokenization, term filtering methods, Euclidean distance measure etc.) and combine them with a new heuristic measure for mining ideas.

As a result, the idea mining approach extracts automatically new and useful ideas from an user given text. We present these problem solution ideas in a comprehensible way to support users in problem solving. This approach is evaluated with patent data and it is realized as a web-based application, named ‘Technological Idea Miner’ that can be used for further testing and evaluation.

1. Introduction

1.1. Overview

An idea is an image existing or formed in the mind but it can be written down as textual information. In the last years, we see a continually increasing amount of information. About 80% off all this information is stored in textual form (Gentsch & Hänlein, 1999). Examples are research papers, articles in technical periodicals, reports, documents, web pages etc. These texts possibly contain many new ideas. A new idea is often needed to discover unconventional approaches e.g. to create a technological breakthrough. However, a manual extraction of new ideas from these masses of texts is time consuming and costly. Therefore, it is useful to search for new problem solution ideas automatically.

Text mining or knowledge discovery from texts refers generally to the process of extracting interesting information and knowledge from unstructured text (Hotho, Nürnberger, & Paaß, 2005). Referring to this, we introduce idea mining as an automatically process of extracting new and useful ideas from unstructured text using text mining methods.

Creating ideas is a well-known topic that is related to psychology and cognitive science. There, we find many approaches dealing with how persons create ideas especially for problem solution. Therefore, in Section 2 we focus on a general process of creating problem solution ideas and use it as rationale for the idea mining approach.

In recent years, data and text mining techniques explore and analyze huge amounts of available textual data (Coussement & Van den Poel, 2009). Idea mining uses known methods from these techniques and combine them with a new method to create text patterns and a new heuristic measure for mining ideas to realize the rationale. Therefore, we present the processing of the idea mining approach in Section 3 and we introduce this new idea mining measure in Section 4.

A further task of idea mining is to present the extracted ideas in a comprehensible way to the user. Therefore, we focus on results of comprehensibility research and their relations to our task (see Section 5). Additionally, we provide an extensive evaluation to show the success of the idea mining approach and specifically the heuristic idea mining measure (see Section 7).

1.2. Idea definition

We limit our approach to the technological language because of two reasons. Firstly, the technological language is much more standardized than the colloquial language (Hoffmann, Kalverkämper, & Wiegand, 1998; Martin-Bautista, Sanches, Serrano, & Vila, 2004). Therefore, we get better results by analyzing technological texts with text mining approaches. Secondly, our idea definition is taken over from technique philosophy (Rohpohl, 1996). There, an idea is

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

0957-4174/$ - see front matter © 2010 Elsevier Ltd. All rights reserved.
doi:10.1016/j.eswa.2010.04.013

defined as a combination of two things: a mean and an appertaining purpose. An example for an idea is a transistor. A transistor is a semiconductor device. It can be used to amplify or switch electronic signals. Here, we have a mean (a semiconductor device) and an appertaining purpose (to amplify or switch electronic signals).

In general, we talk about a new idea if a know mean is related to an unknown purpose or if a known purpose is related to an unknown mean (Thorleuchter, Van den Poel, & Prinzie, 2010c). Then, a new idea is a nanomagnet because a nanomagnet is a miniaturized magnet that also can be used to amplify or switch electronic signals. Here we have an unknown mean (a miniaturized magnet) appearing together with a known purpose. This new idea could be useful to humans who are working in the field of electronic signals because in future nanomagnetic technology possibly could replace transistor technology.

Therefore, we define a new and probably useful idea as a text phrase. This text phrase consists of domain specific terms that occur together in textual information. These terms can be divided up into two subsets. The first subset should represent a known mean (or a known purpose) and the second subset should represent an unknown purpose (or an unknown mean). Additionally, all terms in the first subset should occur together in a text phrase of the technological problem description.

2. Rationale behind idea mining

Creating ideas is a well-known topic that is related to creativity in psychology and cognitive science. One of the first descriptions of the creative process was published by Wallas (1926). His stage model explains creative insights and illuminations for finding a problem solution. This model consists of a four stages process. In stage one 'preparation', the problem is analyzed so that a person recognizes the problem's dimensions. The stage two 'incubation/intimation' and the stage three 'illumination' transfer the problem from the conscious to the unconscious mind. The unconscious mind works on the problem continuously and it probably finds a solution by creative insights and illuminations. This solution is transferred to the conscious mind, which means after some time the person suddenly gets an idea that is new for him and that probably solves the problem. In the last stage 'verification', the idea is tested for novelty and usefulness.

One of the best-known pragmatic approaches of using practical creativity is brainstorming from Osborn (1948). The first step in brainstorming is to define the problem e.g. by creating descriptions of the problem. Then, persons generate new ideas using creativity methods like idea association etc. The last step in the brainstorming process is to cluster the generated ideas and to evaluate it for novelty and usefulness.

Beside this, there are several further approaches dealing with the creation of new ideas. We can learn from all these approaches that for creating ideas three steps are necessary. The first step is to focus on a problem, the second step is to generate some new ideas specific for this problem with creative methods and the third step is to evaluate the generated ideas for novelty and usefulness concerning the problem.

Referring to these approaches, we build an adequate rationale for the idea mining process. Therefore, idea mining also consists of three steps. In the first step, we focus on the problem. Here, the user of our idea mining approach has to provide textual information where he describes his specific problem (a problem description). In the second step, the user has to provide further textual information where he supposes the existence of new and useful ideas (a new text) that probably can solve his problem (Ripke & Stöber, 1972). Ideas are contained in text phrases inside this new text as described in Section 1.2. Therefore, with an automatically process, we extract a very large number of overlapping text phrases from the new text. In the remainder of this paper, text phrases will be named text patterns. In the third step, all extracted text patterns are evaluated for novelty and usefulness. This means, they are compared to the problem description by using a specific idea mining measure. With this measure, text patterns can be classified as new and useful idea. Therefore, idea mining identifies new and useful ideas in three steps:

- Preparation of a problem description
- Extraction of text patterns from a new text and
- Evaluation of text patterns for novelty and usefulness concerning problem description.

3. Idea mining process

Fig. 1 shows the processing of the idea mining approach in different steps based on the rationale for the idea mining process (see Section 2).

With tokenization (Coussement & Van den Poel, 2008), texts are separated in terms and the term unit is word. The set of different terms in a text is reduced by using stop word filtering methods and stemming (Hotho et al., 2005). For this, a general list of stop words is used as well as the well-known Porter stemming algorithm (Porter, 1980).

A related problem to the use of stemming is to identify synonyms and homonyms. Synonyms are different words with identical or at least similar meanings. Homonyms are groups of words with the same spelling but with different meanings. With stemming synonyms and homonyms cannot be identified because stemming does not use knowledge of the context of a term. In this idea mining approach, we do not identify synonyms and homonyms. This is because the approach always considers the context of a term by working on text patterns containing several co-occurring terms as described below.

Here, we show how to create these text patterns automatically. Around each appearance of each term in the new text, we create a text pattern containing the selected term and all terms, which occur in the left and right context of the selected term. To reduce the number of text patterns, we only create text patterns around non-

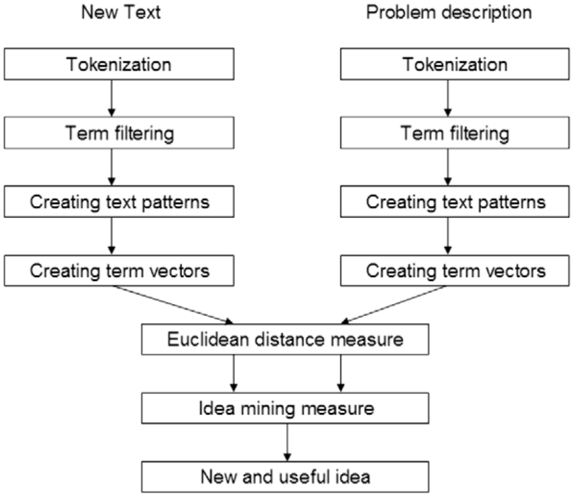


Fig. 1. Processing of our idea mining approach in different steps: After tokenization and term filtering, text patterns are created and term vectors are built representing these text patterns. Term vectors from the new text are compared to term vectors from the problem description using the Euclidean distance measure. Then, term vectors from the new text are compared to their most similar term vectors from the problem description using the idea mining measure. As a result, we get term vectors from the new text that represent new and useful ideas.

stop words and around terms that occur both in the new text and in the problem description.

One important decision to be taken is to determine the length of a text pattern. Text patterns should not be too small so that they contain all terms representing a new idea. Further text patterns should not be too large so that only terms occur in the text patterns that are related to the new idea. For example if we set the length of the text patterns to l then a text pattern contains the selected term, l terms from its left context and also l terms from its right context. The cardinality of the set of stop word filtered and stemmed terms from this pattern is normally smaller than $2 * l + 1$ because some terms are stop words, some terms occur twice and some terms have the same stem.

In this paper, we do not use a constant length l for all patterns but a variable length of text patterns based on a dynamic adaptation of its context. This is realized by using a term weighting scheme based on the difference between stop words and non-stop words because the importance of a stop word in a text pattern is not as high as the importance of a non-stop word. If an author formulates an idea very briefly by joining catchwords together then he normally does not use many stop words and the text pattern length can be small. If an author formulates an idea in a flowery style that means his writing is not expressed in a clear and simple way then he normally uses more stop words and the text pattern length has to be larger. In the idea mining application the value of text pattern length l and the percentage of the importance of stop words u and of non-stop words v can be provided by the user.

To compute the variable length of a text pattern, we firstly define the term weighting scheme.

Definition 1. Let (a text) $T = [w_1, \ldots, w_n]$ be a list of terms (words) w_i in order of appearance and let $n \in N$ be the number of terms in T and $i \in [1, \ldots, n]$. Let $\Sigma = \{\tilde{w}_1, \ldots, \tilde{w}_m\}$ be a set of domain specific stop terms (Thorleuchter, Van den Poel, & Prinzie, 2010b) and let $m \in N$ be the number of terms in Σ. Let the percentage $u \in N$ be a term weighting coefficient for stop words. Let the percentage $v \in N$ be a term weighting coefficient for non-stop words. Then, we define $f_g(w_i) \in N$ as term weighting scheme:

$$f_g(w_i) = \begin{cases} u | w_i \in \Sigma \\ v | w_i \notin \Sigma \end{cases} (\forall i \in \{1, \ldots, n\}). \tag{1}$$

We give an example for this. The text pattern 'components for frequency conversion of infrared lasers' is built around the word 'conversion'. It contains the word conversion itself, three terms from its left context (components for frequency), and three terms from its right context (of infrared lasers). Here, we use a constant length $l = 3$ and a term weighting scheme with $\alpha = \beta = 100\%$. This means the importance of a stop word is equal to the importance of a non-stop word. The next text pattern is an example for a variable length: 'In a 1st phase, known but so far not available materials and technologies such as layer systems and crystals'. This text pattern is built around the word 'technologies'. Here we use a constant length $l = 3$ and a term weighting scheme with $u = 10\%$ and $v = 100\%$. As a result, this text pattern contains six terms from the right context and eleven terms from the left context of the term 'technologies'. In this example, non-stop words are phase, materials, technologies, layer, systems, and crystal. We compute the number of terms from the left and right context as described below:

Definition 2. Let $l \in N$ be a constant length of text patterns. Let l_i^{left} be the number of terms from the left context of a text pattern that is built around the term w_i. Let l_i^{right} be the number of terms from the right context of a text pattern that is built around the term w_i. Then, we define $l_i^{left} \in N$ and $l_i^{right} \in N$ as:

$$l_i^{right} = \min_j \left| \left(\sum_{k=1}^{j} f_g(w_{i+k}) \geqslant l \right) \vee (i + j = n) \quad \forall i \in \{1, \ldots, n\}, \right. \tag{2}$$

$$l_i^{left} = \min_j \left| \left(\sum_{k=1}^{j} f_g(w_{i-k}) \geqslant l \right) \vee (i - j = 1) \quad \forall i \in \{1, \ldots, n\}. \right. \tag{3}$$

After computing l_i^{left} and l_i^{right}, we can build a text pattern T_i around the term w_i from the text $T = [w_1, \ldots, w_n]$.

$$T_i = \left[w_{i - l_i^{left}}, \ldots, w_i, \ldots, w_{i + l_i^{right}} \right]. \tag{4}$$

For each text pattern from the new text, we create a term vector in vector space model. The size of the vector is defined by the number of different stemmed and stop word filtered terms in the new text. For text pattern encoding, we use binary term vectors that means a vector element is set to one if the corresponding unstemmed term is used in the text pattern and to zero if the term is not. We also build text patterns from the problem description and create term vectors as described above.

To identify new and useful ideas, we create a specific idea mining measure. This idea mining measure is described in Section 4. By comparing a vector from the new text to one from the problem description, we can compute a result value always between 0% and 100% using this measure. The greater the result value the more is the probability that the vector from the new text represents a new and useful idea concerning a vector from the problem description.

We use this measure for comparing vectors from the new text to their most similar vectors from the problem description but not to all vectors. This is because result values from comparing a vector to its most similar vectors predominate result values from comparing a vector to its further vectors. For example if a vector from the new text is similar to one from the problem description then the idea is not new to the user regardless whether result values from comparing this vector to further vectors from the problem description are greater than zero. Therefore, we can be sure that a vector represents a new and useful idea only if it gets a great result value from idea mining measure concerning one of its most similar vectors. Further, the computing of the idea mining measure is time consuming. Therefore, it is necessary to limit the number of comparisons with idea mining measure for implementing an idea mining application.

We choose a two-step classification way. In the first step, we compare each vector from the new text to all vectors from the problem description by using the well-known Euclidean distance measure. Fortunately, the computing of the Euclidean distance measure is not time consuming so that it is suited for implementing in an idea mining application. In detail, for each vector from the new text, we identify all vectors from the problem description where the Euclidean distance result value is the lowest that means we identify the most similar vectors. In the second step, we compare each vector from the new text to its most similar vectors using the idea mining measure.

Each vector from the new text – that is compared to several similar vectors – gets the highest result value from idea mining measure as result value. To identify a new and useful idea we use alpha-cut method. An alpha-cut of the idea mining measure result value is the set of all vectors from the new text such that the appertaining result value is greater than or equal to alpha ($\tilde{\alpha}$). In the idea mining application, the user can provide the value of $\tilde{\alpha}$.

4. Idea mining measure

With the idea mining measure, we compare a vector that represents a text pattern from the new text to its most similar vectors from the problem description to identify a new and useful idea inside the text pattern from the new text. In detail, we have to find text pattern from the new text where all terms representing a mean (purpose) and no terms representing a purpose (mean) occur in a text pattern from the problem description.

If all terms in the text pattern from the new text are known, which means all terms also occur in a text pattern from the problem description then the idea is not new to the user. Furthermore, the idea is not useful if all terms in the text pattern from the new text are unknown because there is no relation to the problem. It is shown in Thorleuchter (2008) that to find new and useful ideas the number of known terms (e.g. representing a mean) and the number of unknown terms (e.g. representing an appertaining purpose) shall be well balanced.

Definition 3. Let α_i be a the set of stemmed and stop word filtered terms representing a text pattern with number i from the new text. Let β_j be a set of stemmed and stop word filtered terms representing a text pattern with number j from the problem description. Let γ be the set of all stemmed and stop word filtered terms from the new text. Let $x = |\gamma|$ be the cardinality of γ. Let $\omega_i \in \{0,1\}^x$ be a term vector in vector space model concerning α_i. Let $\rho_j \in \{0,1\}^x$ be a term vector in vector space model concerning β_j. Let $p = |\alpha_i| = \sum_{k=1}^{x} \omega_{i,k}$ be the number of all (known and unknown) terms in text pattern with number i. Let $q = |\alpha_i \cap \beta_j| = \sum_{k=1}^{x} \omega_{i,k} \cdot \rho_{j,k}$ be the number of known terms in text pattern with number i concerning a text pattern with number j from the problem description. Then, we define m_1 as measure for well-balanced known and unknown term distribution.

$$m_1 = \begin{cases} \frac{2*(p-q)}{p} & (q \geqslant \frac{p}{2}), \\ \frac{2*q}{p} & (q < \frac{p}{2}). \end{cases} \tag{5}$$

The known terms in the text pattern from the new text should occur in the problem description more frequently than other terms. This is because they represent a known mean or a known purpose that is a central part of the problem. In the problem description, terms that represent the problem occur more frequently than other terms. For this, we define these frequent terms by using a percentage z as parameter and we compute m_2 as the number of known and frequent terms over the number of all known terms.

Definition 4. Let z be a percentage. Let δ be a set of z% most frequently stemmed and stop word filtered terms in the problem description. Let $\xi \in \{0,1\}^x$ be a term vector in vector space model concerning δ. Let $r = |\alpha_i \cap \beta_j \cap \delta| = \sum_{k=1}^{x} \omega_{i,k} \cdot \rho_{j,k} \cdot \xi_k$ be the number of known terms which occur frequently in the problem description. We define m_2 as measure for frequently occurrence of known terms in the problem description.

$$m_2 = \frac{r}{q}. \tag{6}$$

The unknown terms in the text pattern from the new text represent a new approach (an unknown mean or purpose), which is a central part of the new idea. These terms normally occur more frequently than other terms in the new text because this text deals about the new idea. For this, we also define these frequent terms by using a percentage z as parameter and we compute m_3 as the number of unknown and frequent terms over the number of all unknown terms.

Definition 5. Let ϕ be a set of z% most frequently stemmed and stop word filtered terms in the new text. Let $\tau \in \{0,1\}^x$ be a term vector in vector space model concerning ϕ. Let $s = |\alpha_i \cap \overline{\beta_j} \cap \phi| = \sum_{k=1}^{x} \omega_{i,k} \cdot \tau_k - \sum_{k=1}^{x} \omega_{i,k} \cdot \rho_{j,k} \cdot \tau_k$ be the number of unknown terms which occur frequently in the new text. We define m_3 as measure for frequently occurrence of unknown terms in the new text.

$$m_3 = \frac{s}{p-q}. \tag{7}$$

There are often characteristic terms (higher, quicker, integrated, minimized etc.) that occur together with new ideas. They point to a changing purpose or a changing mean and can be an indicator for new ideas.

Definition 6. Let λ be a set of these characteristic terms (stemmed and stop word filtered). Let $\theta \in \{0,1\}^x$ be a term vector in vector space model concerning λ. Let $t = |\alpha_i \cap \lambda| = \sum_{k=1}^{x} \omega_{i,k} \cdot \theta_k$ be the number of these characteristic terms in text pattern with number i. We define m_4 as measure for changing means and purposes.

$$m_4 = \begin{cases} 1 & (t > 0), \\ 0 & (t = 0). \end{cases} \tag{8}$$

The idea mining measure bases on all four heuristic sub measures.

Definition 7. Let $h \in \{1, \ldots, 4\}$ and let $g_h \geqslant 0$ be weighting factors with $\sum_{h=1}^{4} g_h = 1$. Let the idea mining measure be the sum of all four sub measures multiplied by weighting factors g_h in case of $p \neq q$.

$$m = \begin{cases} g_1 m_1 + g_2 m_2 + g_3 m_3 + g_4 m_4 & (p \neq q), \\ 0 & (p = q). \end{cases} \tag{9}$$

5. Idea mining and comprehensibility research

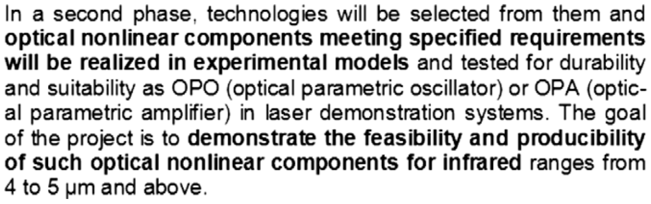
The aim of idea mining is to find new and useful ideas but also to present these ideas in a comprehensible way to the user. To realize this, we focus on comprehensibility research.

Up to the 1960's comprehensibility was a property of the text. It was measured in an objective way by analyzing text parameters like word length, sentence length, word-usability, relationship between number of different words and number of words. The well-known approach in this time was the 'Reading Ease'-formula from Flesch (1948).

Later research in this field focuses on cognitive effects by doing textual production and reception. The results of this research are presented by two approaches: the 'Hamburger Verständlichkeitsmodell' (Langer, Schulz v. Thun, & Tausch, 1974) and the 'Groebener Modell' (Groeben, 1982). Both approaches describe four dimensions of comprehensibility: simplicity, structure-organization, brevity-shortness and interest-liveliness.

A further approach from cognition research is named text excerption. If a human expert finds new and useful ideas in texts he highlights all corresponding text phrases e.g. with text marking. This behaviour is described by Puppe, Stoyan, and Studer (2003).

In the idea mining application, text excerption is used to present the extracted ideas to the user (Fig. 2 shows an example). For



In a second phase, technologies will be selected from them and **optical nonlinear components meeting specified requirements will be realized in experimental models** and tested for durability and suitability as OPO (optical parametric oscillator) or OPA (optical parametric amplifier) in laser demonstration systems. The goal of the project is to **demonstrate the feasibility and producibility of such optical nonlinear components for infrared** ranges from 4 to 5 μm and above.

Fig. 2. We present the new text back to the user with text patterns in bold print that represent new and useful ideas.

the 'Groebener Modell' marking text pattern is important for structure-organization and this leads directly to comprehensibility. In this point there are differences between the 'Groebener Modell' and the 'Hamburger Verständlichkeitsmodell' in which structure-organization is not so important for comprehensibility.

As a result, the presentation of ideas in the idea mining application based on text excerption. It is comprehensible after the 'Groebener Modell' and it is less comprehensible after the 'Hamburger Verständlichkeitsmodell'.

6. Results and discussions

In a study for the German Ministry of Defence (MoD), we use this approach to identify new technological ideas for the German defence research program. In detail, we have to identify new solution ideas to solve current problems in German defence based research projects. We extract new ideas from 300 descriptions of research projects granted in 2006 by the National Institute of Standards and Technology (NIST) in the United States Small Business Innovation Research (SBIR) Program. We use textual information from current defence based research projects of the German MoD as problem description (Thorleuchter, Van den Poel, & Prinzie, 2010a). As a result, we extract several new ideas that are useful for German defence research planners and that now are used as starting point for collaboration projects or for new defence based research projects. A proper selection of these ideas is a strategic issue and - together with the weapon selection problem (Dagdeviren, Yavuz, & Kilinc, 2009) – it has significant impacts to the efficiency of future defence systems. The results are published in Fenner and Thorleuchter (2009). Here, we show some successful examples:

A modified focal plane array technology is identified that can be used to create a detector for the far ultraviolet spectrum. It leads to an improvement of military reconnaissance. This idea is new because up to now focal plane array technology is only used in the infrared, visual and near ultraviolet area.

Further, the approach identifies personnel ultrasonic locating equipment that was originally developed to make orientation possible for fire fighters in dense smoke. It also can be used to improve the location and navigation of soldiers in urban warfare (e.g. in buildings).

Additionally, the approach shows that the use of avalanche photodiode (APD) technology can improve the internal gain and the dark current of infrared detectors. This also leads to an improvement of military reconnaissance.

This study shows that some of the automatically extracted ideas are useful for technological research planners from the German MoD. Unfortunately, the used problem description (textual information about current defence based research projects) is classified as German restricted (Verschlusssache – Nur für den Dienstgebrauch) that means it is not allowed to distribute it to the scientific community. Therefore, we cannot use the results of this study to evaluate this idea mining approach. However, a separate evaluation (see Section 7) is done using (unclassified) patent data that allows re-computing of the evaluation.

7. Evaluation

The idea mining measure as central point in the idea mining approach consists of four heuristic sub measures that are not theoretically founded. Therefore, it is crucial to provide an extensive evaluation to show their success. We compare this approach to a baseline because we are not aware of other approaches for idea mining. As measure for the baseline, we use Jaccard's coefficient (Ferber, 2003) as well-known heuristic similarity measure.

The idea mining approach is evaluated by using our idea mining application (see Section 8). There the web-based application and all texts that are used for evaluation are presented. Additionally, we create an alternative idea mining application, based on Jaccard's coefficient instead of the idea mining measure for the sole purpose of comparison to the baseline.

For evaluation, we use patent data because in patent descriptions, we normally can find new ideas, which include a considerable part of scientific and technological knowledge (Li, Wang, & Hong, 2009). We use the abstract of a patent as new text. A patent often bases on further patents. We aggregate abstracts of theses references as problem description. Then we identify new and useful ideas from this patent concerning its patent references using the idea mining applications.

We use abstracts from 40 randomly selected patents and from their references, a general stop word list and Porter stemmer for evaluation. Then we determine the parameters of the idea mining measure (g_1, g_2, g_3, g_4, $\tilde{\alpha}$, and z) as well as the parameters for the length of the text patterns (l, u, and v).

For this, we use further patent data and their references as new text and as problem description. The results are evaluated by a human expert and compared to each single sub measure m_1, m_2, m_3 and m_4 alone. We find out that using the first sub measure alone is successful. If this sub measure is small then the corresponding text pattern normally does not contain a new and useful idea. If this sub measure is large then the probability that the text pattern contains a new idea is also high. We also find out that using the further sub measures alone is not successful. This means, they are successful only if the result value of the first sub measure is medium to high. Therefore, they only can be used in addition to the first sub measure.

The results of the second and third sub measures depend on the parameter z. This parameter is used to define frequent terms by building a set of z% most frequently stemmed and stop word filtered terms. We heuristically think that this parameter should be between 10% and 30% to get good sub measures. This is because if z is greater than 30% then we probably classify several terms, which only occur once as frequent terms. If z is smaller than 10% then we only identify high frequently terms for the set. In this case, the result values of the second and third sub measures are small regardless weather known terms occur frequently in the problem description or unknown terms occur frequently in the new text. Therefore, we determine z to the mean value (20%). Additionally, we see that the second and third sub measure is nearly equally successful and that the fourth sub measure is less successful. Therefore, we heuristically determine the parameters of g_1 to 50%, g_2 to 20%, g_3 to 20% and g_4 to 10%.

We also have used other values to optimize the combination of these four sub measures. However, we do not find a combination that is generally superior to the selected combination. This is because the success of these value combinations depends on the quality of the user given textual information.

Then, we determine the alpha-cut value $\tilde{\alpha}$ of the idea mining measure m. If the percentage $\tilde{\alpha}$ is small then we get many result items. This leads to a small precision value because many extracted text patterns do not contain a new and useful idea. If $\tilde{\alpha}$ is large then we only get a very small number of results and probably our recall value is small because we do not find most of the new and useful ideas in the new text. A human expert checks the results of several patent descriptions for an optimal value of $\tilde{\alpha}$. He gets the experience that 60% is a good compromise. Therefore, we set $\tilde{\alpha}$ to 60%. We also determine the alpha-cut value of Jaccard's coefficient as measure for the baseline to 20% by using the same way of evaluation as described above.

After this, we determine the length of the text patterns. The length depends on the parameter l and on $f_g(w_i)$, a term weighting scheme that is based on the difference between stop words and non-stop words (see Section 3). Text patterns should not be too small so that they contain all terms representing a new and useful idea. Additionally, text patterns should not be too large so that further terms occur in the text patterns that are not related to the new and useful idea. To find out an optimal size of text patterns, we create text patterns from several patent descriptions by using different values for l and for the percentages u and v. A human expert checks the different length of these text patterns for an optimal size. He gets the best results by setting the value of text pattern length l to 7 terms and the percentage u to 50% and v to 100%.

Then, the approach extracts automatically about 200 new ideas from the 40 randomly selected patents. To cluster these results, means and purposes are assigned to scientific categories in the science citation index and examples are presented below. Several ideas are identified that uses methods from 'Artificial Intelligence' (mean) for applications in 'Health Care Sciences and Services' (purpose). We also identify new ideas using 'Imaging Science and Photographic Technology' (mean) for 'Medical Informatics' purposes. Further ideas use techniques from 'Remote Sensing' (mean) in the field of 'Tropical Medicine' (purpose). Additionally, several ideas use 'Computer Science, Theory and Methods' (mean) for applications in 'Psychiatry' (purpose). Furthermore, methods from 'Artificial Intelligence' (mean) are used for 'Automation and Control Systems' purposes.

To evaluate these results, we use precision and recall measures commonly used in information retrieval based on true positives, false positives and false negatives. For this, we have to define the ground truth for our evaluation. Therefore, a human expert also identifies new and useful ideas from these patents manually that means without using our idea mining approach. He uses the idea definition in Section 1.2. This means, he checks each text pattern for finding terms representing a known mean (purpose) and terms representing an unknown purpose (mean). These results are the ground truth for the evaluation.

For each patent, we compute its precision and recall values by using the idea mining measure and by using the Jaccard's coefficient. Then, we compute the average precision and recall values. As a result, we get a precision value of 40% and a recall value of 25% by using the idea mining approach with the idea mining measure. A precision value of 40% means that if the idea mining approach extracts ten text patterns then four of them represent a new and useful idea. A recall value of 25% means that if there are four new and useful ideas in the new text then the idea mining approach extracts only one of them. In contrast to this, we get a precision value of 30% and a recall value of 20% by using Jaccard's coefficient. This is because in some texts Jaccard's coefficient extracts text patterns from the new text that are similar to text patterns from the problem description. This represents probably a known idea but not a new idea.

Beside Jaccard's coefficient, we also test other well-known heuristic measures like overlap-index, cosine-similarity and dice-similarity (Ferber, 2003) as baseline. However, we get nearly the same results for the precision (30%) and for the recall (20%) value.

8. The idea mining application

The idea mining application focus on users without extensive knowledge in the text mining field as well as on text mining experts. We give them the possibility to extract specifically problem solution ideas for their own needs using this idea mining approach. They can access to the web-based application via the internet. It is available under http://www.text-mining.info and it is programmed in perl and ruby.

An user has to provide two textual files, a problem description and a new text that probably consists of problem solution ideas. These files can be formatted in various ways e.g. as plain text, html, xml etc. However, scripting code, (html- or xml-) tags, and images are discarded that means the application extracts plain text from the provided files. Then, the user has to select the language of these texts to integrate a general stop word list of this language. The application offers general stop word lists in English, German, Dutch, Spain and French. After determining the parameters of the application the automatically extraction of new and useful ideas from the new text starts as described in the idea mining process (see Sections 3 and 4). As a result, new ideas are presented as described in Section 5.

9. Conclusions and future research

This study shows the success of an automatic approach for finding new ideas from textual information. For this, the study transforms creativity approaches from psychology and cognitive science to text mining approaches. One main finding here is to redefine an abstract term (an idea) in a concrete way that it can be used for computing with text mining methods. In detail, it is shown that a technological idea represents a combination of a purpose and a mean and that purposes and means are defined by a combination of terms, which co-occur.

Additionally, it is shown that problems and problem solution ideas can be represented as term vectors in vector space model. For this, the study contributes a new (idea mining) measure. This measure identifies new ideas by comparing vectors that represent a problem to vectors that represent a problem solution idea. Last, it is shown that approaches from comprehensibility research can be adopted to this approach to present the new ideas in a comprehensible way to the user. As further main finding, it is demonstrated that this theoretical approach can be realized by a web-based application. The success of the idea mining measure is proved by comparing it to further heuristic measures (overlap-index, cosine-similarity and dice-similarity).

Directions for future research are given by the fact that nowadays there is a large amount of textual information available on the internet and this information probably contains many new technological ideas. Enlarging this approach to a web idea mining approach that automatically identifies problem solution ideas from the internet is an interesting topic for further research.

Additionally, the parameters of the approach can be optimized and the idea mining measure can probably be enlarged with further aspects to improve its quality that means to get better results for the precision and recall values.

A further aspect is to transform this idea mining approach to the colloquial language. For this, it is necessary that the idea definition also contains new product ideas from the consumers. Then, new product ideas can be identified to support marketing activities.

Last, the approach can be extended with innovation-related aspects. Then, extracted ideas can be classified as innovative ideas and might be used as starting point for the new product development.

Acknowledgements

We thank Joachim Schulze and Jörg Fenner for constructive technical comments.

References

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 and Management, 45*, 165.

Coussement, K., & Van den Poel, D. (2009). Improving customer attrition prediction by integrating emotions from client/company interaction emails and evaluating multiple classifiers. *Expert Systems with Applications, 36*, 6127–6134.

Dagdeviren, M., Yavuz, S., & Kilinc, N. (2009). Weapon selection using the AHP and TOPSIS methods under fuzzy environment. *Expert Systems with Applications, 36*, 8150.

Fenner, J., & Thorleuchter, D. (2009). Textmining-Analyse von Forschungsvorhaben des National Institute of Standards and Technology. Euskirchen: Fraunhofer INT Edition.

Ferber, R. (2003). *Information retrieval*. Heidelberg: dpunkt.verlag. pp. 74–80.

Flesch, R. (1948). A new readability yardstick. *Journal of Applied Psychology, 32*, 221–233.

Gentsch, P., & Hänlein, M. (1999). Text mining. *WISU, 12*, 1646.

Groeben, N. (1982). *Leserpsychologie: Textverständnis – Textverständlichkeit*. Münster: Aschendorff.

Hoffmann, L., Kalverkämper, H., & Wiegand, H. E. (1998). *Fachsprachen – Languages for special purposes: Ein internationales Handbuch zur Fachsprachenforschung und Terminologiewissenschaft – An international handbook of special-language and terminology research*. Berlin: Walter de Gruyter. p. 1602.

Hotho, A., Nürnberger, A., & Paaß, G. (2005). A brief survey of text mining. *LDV Forum, 20*(1), 19–26.

Langer, I., Schulz v. Thun, F., & Tausch, R. (1974). *Verständlichkeit in Schule und Verwaltung*. München: Ernst Reinhardt.

Li, Y. R., Wang, L. H., & Hong, C. F. (2009). Extracting the significant-rare keywords for patent analysis. *Expert Systems with Applications, 36*, 5200–5204.

Martin-Bautista, M. J., Sanches, D., Serrano, J. M., & Vila, M. A. (2004). Text mining using fuzzy association rules. In V. Loia, M. Nikravesh, & L. A. Zadeh (Eds.), *Fuzzy logic and the internet* (pp. 173). Berlin: Springer-Verlag.

Osborn, A.-F. (1948). *Your creative power*. New York: C. Scribner's sons.

Porter, M. F. (1980). An algorithm for suffix stripping. *Program, 14*(3), 130–137.

Puppe, F., Stoyan, H., & Studer, R. (2003). Knowledge engineering. In G. Görz, C. R. Rollinger, & J. Schneeberger (Eds.), *Handbuch der Künstlichen Intelligenz* (pp. 611). München: Oldenbourg.

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

Rohpohl, G. (1996). Das Ende der Natur. In L. Schäfer & E. Sträker (Eds.), *Naturauffassungen in Philosophie, Wissenschaft und Technik, Freiburg* (pp. 143–163). München: Alber.

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-Verlag.

Thorleuchter, D., Van den Poel, D., & Prinzie, A. (2010a). A compared R&D-based and patent-based cross impact analysis for identifying relationships between technologies. *Technological Forecasting and Social Change*, doi:10.1016/j.techfore.2010.03.002.

Thorleuchter, D., Van den Poel, D., & Prinzie, A. (2010b). Extracting consumers needs for new products - a web mining approach. In M. Gong, & Q. Luo (Eds.), *Proceedings WKDD 2010* (pp. 440). Los Alamitos: IEEE Computer Society.

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*. Berlin: Springer-Verlag.

Wallas, G. (1926). *The art of thought*. New York: Harcourt Brace.