Mining Innovative Ideas to Support New Product Research and Development

Dirk Thorleuchter[a], Dirk Van den Poel[b], and Anita Prinzie[b]

[a] Fraunhofer INT, D-53879 Euskirchen, Appelsgarten 2, Germany & PhD Candidate, Ghent University, dirk.thorleuchter@int.fraunhofer.de
[b] Ghent University, Faculty of Economics and Business Administration, B-9000 Gent, Tweekerkenstraat 2, Belgium, dirk.vandenpoel@ugent.be, anita.prinzie@ugent.be

Abstract

Here, we present an approach for automatically identifying the innovative potential of new technological ideas extracted from textual information. The starting point of each innovation is a good and new idea. Unfortunately, a high percentage of innovations fail, which means many ideas do not have the potential to become an innovation in future. The innovation process from a new idea as starting point via research, development, and production activities through to an innovative product is very cost- and time-consuming. Thus, the objective of our work is to identify the innovative potential of new technological ideas to improve the performance of the innovation process.

We extract new technological ideas from provided textual information. We also identify innovative technology fields by analysing relationships among technologies. All identified ideas are assigned to innovative technology fields by using text mining and text classification methods. Technological ideas in these fields are presented to the user as innovative ideas.

Key Words
Idea Mining, Text Mining, Text Classification, Innovation

1. Introduction

The word innovation refers to the latin terms novus (that means new) and innovatio (that means something is newly created). An innovation includes a new idea [9] as well as its realization e.g. as innovative product that is successful in market. Thus in economical sense, we talk about innovations if the newly created object increases producer or customer value [13].

To create an innovation, an innovation process can be used. It has the aim to lead a new idea to an innovative product. Thus, the starting point of the innovation process is a new technological idea [14]. Based on this idea, a research process starts. The result is probably a prototype that is developed further in a developing process. After this developing process a production process starts and it leads to a product [4]. If this product is successful in market that means it increases producer or customer value then it is an innovative product and the idea standing behind this innovation can be defined as innovative idea. However, by use of this economical definition, we only can identify innovative ideas subsequent to the innovation process that means after they become successful products in market.

Unfortunately, the innovation process is very cost- and time-consuming [5] and a high percentage of innovations fail. Thus, the objective of our work is to identify the innovative potential of new technological ideas before selecting them as starting ideas. This probably can improve the performance of the innovation process.

2. Background

Our definition of a technological innovation is based on bibliometrical analyses as described in [15]. There, it is shown that innovations normally do not occur alone but together with several further innovations. These groups of innovations are based on a common innovation field. Innovation fields are newly appeared technologies or scientific disciplines that occur on the borders of established technologies or scientific disciplines. This means they occur between at least two technologies or scientific disciplines that are not related. A definition of possible relationships is given in Sect. 6. Thus, innovations can be classified as interdisciplinary products. The (innovative) ideas behind these innovations also are of an interdisciplinary nature and they also occur together in an innovation field.

Our idea definition derived from technique philosophy [17]. There, a technological idea consists of two things: a means and an appertaining purpose [2]. Thus, we define an 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 means and the second subset should represent a purpose. An example for an idea is a nanomagnet (the means) that can be used to switch electronic signals (the appertaining purpose). This definition is used to identify interdisciplinary ideas by assigning means and purpose of an idea to different non-related, established technologies or scientific disciplines.

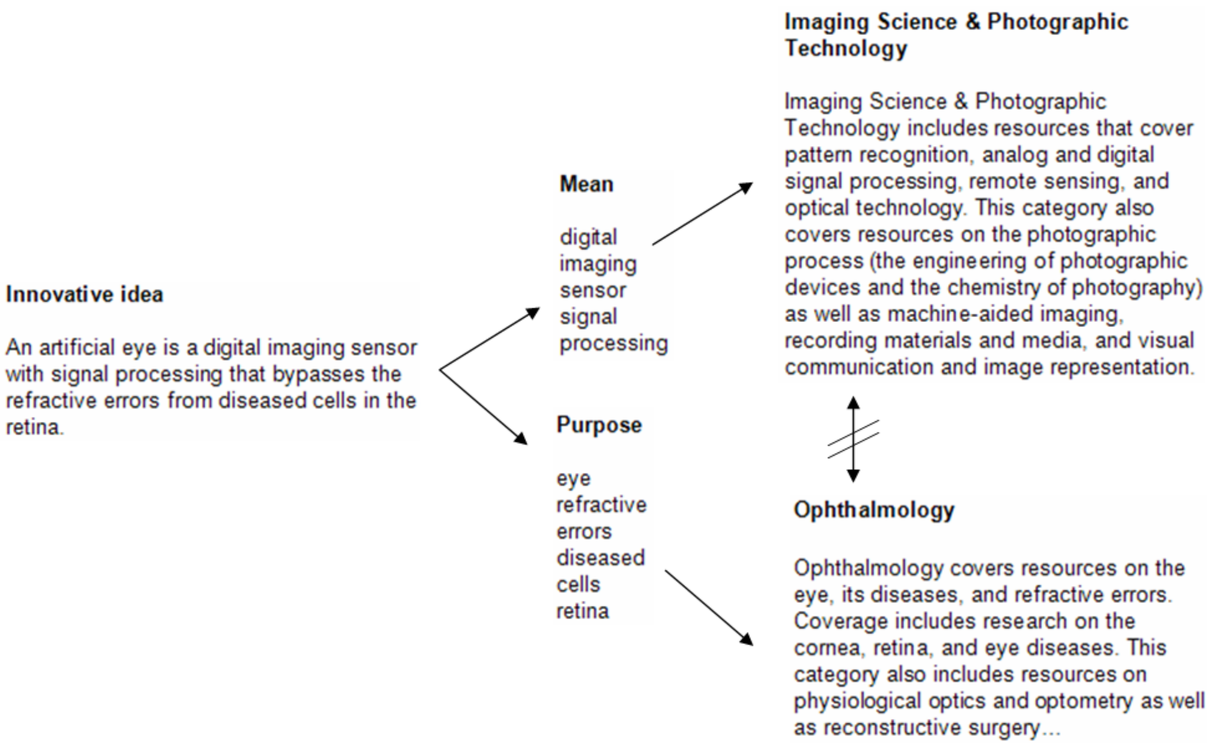
To classify ideas as innovative, we have to identify several interdisciplinary ideas that occur together in an innovation field. For this, we firstly have to provide technological context information containing descriptions of established technologies or scientific disciplines and we have to define their relations.

Secondly, we have to classify ideas as interdisciplinary by assigning means and purposes to established technologies or scientific disciplines that are not related. For example, if a means from a bionic idea can be assigned to biology and the appertaining purpose can be assigned to technological engineering then the bionic idea is interdisciplinary. This gives a hint that the combination of biology and technological engineering is probably an innovation field.

To be sure that it is really an innovation field, we thirdly have to find several further interdisciplinary ideas that can be assigned to the same non-related technologies or scientific disciplines combination and classify all the interdisciplinary ideas in this field as innovative ideas.

3. Process of Mining Innovative Ideas

This approach uses an existing idea mining approach [18] that supports users to identify means and purposes in text phrases (see Sect. 4). Then, we provide descriptions of scientific categories as context information (see Sect. 5). Both the means and the purpose of extracted new and useful ideas are assigned to several scientific categories by use of multi-label classification (see Sect. 7). After this, we compare each scientific category from means to each scientific category from purpose to find out relationships between them (see Sect. 6). Fig. 1 shows an example for the processing of this approach.



An artificial eye is a digital imaging sensor with signal processing that bypasses the refractive errors from diseased cells in the retina.

Imaging Science & Photographic Technology includes resources that cover pattern recognition, analog and digital signal processing, remote sensing, and optical technology. This category also covers resources on the photographic process (the engineering of photographic devices and the chemistry of photography) as well as machine-aided imaging, recording materials and media, and visual communication and image representation.

Ophthalmology covers resources on the eye, its diseases, and refractive errors. Coverage includes research on the cornea, retina, and eye diseases. This category also includes resources on physiological optics and optometry as well as reconstructive surgery…

Fig. 1. Means and purpose are extracted from an idea and are assigned to different scientific categories. If the categories that are assigned by the means are not related to categories that are assigned by the purpose then the idea is of an interdisciplinary nature. If several ideas also are of an interdisciplinary nature concerning these categories then the combination of both categories is defined as innovation field and ideas from this field are presented as innovative ideas.

If a means is assigned to several scientific categories and the appertaining purpose is assigned to further scientific categories that are not related to any scientific category of the means then the corresponding idea is of an interdisciplinary nature. If several further ideas also are interdisciplinary concerning at least two of the above mentioned scientific categories then we define the combination of these scientific categories as innovation field. For this, the user provides the smallest number of interdisciplinary ideas that is sufficient to define such an innovation field. Ideas from these innovation fields are classified as innovative ideas.

4. Acquisition of Ideas

Our approach based on technological ideas. The user extracts them from provided textual information e.g. patent data. He is supported by a further approach that automatically extracts new and useful ideas from textual information as presented in [18]. This approach extracts textual phrases that represent new and useful ideas. Additionally for each idea, it identifies terms that represent a means as well as terms that represent the appertaining purpose. This information is used as input for our approach.

5. Acquisition of Technological Context Information

To provide technological context information, we focus on scientific categories. We can find an overview of current scientific categories in the science citation index (SCI). This index is built on bibliographic information, author abstracts, and cited references from about 3,700 science and technical journals. The content of these highly cited journals is assigned to 172 scientific categories. The official description of all categories in the SCI is available in scope notes [11] that is manually created, of good quality, and up to date. We use this description as technological context information for our approach.

6. Relationship among Scientific Categories

After providing descriptions of scientific categories that represent technological context information, the next step is to identify relationships among these scientific categories.

In general, we identify two different kinds of relationships [8]. One kind of relationship is that technologies can be similar to other technologies. They deal with the same technology field but have a different focus. The descriptions of two similar technologies also are similar because they both contain the same domain specific terms by describing the technological field.

A further kind of relationship is that technologies are related in a substitutive, integrative, predecessor or successor way. If technologies are related in this way then they deal with the same application field. Their descriptions also are similar because they both contain the same domain specific terms representing the application field.

The descriptions of the scientific categories in scope notes contain terms representing the technological field as well as terms representing potential application fields. If we identify similar terms in descriptions of two different scientific categories then both categories are related according to at least one kind of relationship. Therefore, related categories are identified by comparing category descriptions among each other.

Comparing is done by transforming category description to term vectors in vector space model. For this, terms in the descriptions are tokenized [6] by using the term unit as word, stop word filtered by using a standard stop word list [12], and stemmed [10] using a dictionary-based stemmer combined with a set of production rules [16] to give each term a correct stem. The production rules are used when a term is unrecognizable in the dictionary. Vectors representing scientific categories can be compared using similarity measures in combination with the alpha cut method [1] and two categories are classified as related if the corresponding similarity measure result value is greater than or equal to alpha. For comparing, we prefer the well-known Jaccard's coefficient measure [7] because it considers well the different sizes of both vectors.

7. Classification of Ideas

Each selected idea consists of a set of terms that represents a means and of a set of terms, which represents an appertaining purpose. To identify an interdisciplinary technological idea we have to assign both sets to scientific categories. Both sets of terms are stop word filtered and stemmed as described in Sect. 6. For multi-label classification, we transform these sets to term vectors in vector space model and compare them with term vectors from each scientific category. For comparing, we also use Jaccard's coefficient measure in combination with the alpha cut method. As a result, means and purposes are assigned to scientific categories only if the appertaining Jaccard's coefficient result value is greater than or equal to alpha.

Each means and each purpose of a new idea is probably assigned to several scientific categories. To identify relations, we compare each scientific category from means to every single scientific category from purpose as described in Sect. 6. If we cannot find any relationships then the new idea is of interdisciplinary nature and each of these scientific category combinations from means and purpose is probably an innovation field. If we identify at least n interdisciplinary ideas that can be assigned to one specific scientific category combination then we define an innovation field on this basis. The user provides the smallest number n of interdisciplinary ideas that are sufficient to define such an innovation field.

8. Results and Evaluation

We present a heuristic approach for automatically identifying the innovative potential of new technological ideas. The extraction of ideas and the identification of terms that represent means and purposes is already evaluated in [18]. Thus, the evaluation is limited to the further steps of our approach and it is based on current context information. For this, scientific categories in the science citation index as current technological information described in scope notes [11] are used.

The approach extracts 1000 new ideas from randomly selected patents because patent descriptions consist of new ideas that also are innovative. However, not all new ideas are innovative in terms of the technological innovation definition in Sect. 2. 500 ideas are used as training examples to obtain the optimal parameter values and 500 ideas are used as test set to validate and compare the model. To evaluate the results of the approach, we use precision and recall measures commonly used in information retrieval based on true positives, false positives, and false negatives. For this, the ground truth for our evaluation is defined. Therefore, a human expert classifies the 1000 new ideas as innovative or as non-innovative.

The approach depends on three parameters (n, α_1, α_2). The smallest number (n) of interdisciplinary ideas that is sufficient to define an innovation field gives a hint concerning the innovative potential of the new idea. If the number n is large then we only obtain ideas as result items that probably consist of a very high innovative potential. This is because we identify many ideas that are classified concerning a specific non-related combination of scientific categories. Here, we have a high probability that this category combination represents an innovation field. If the number n is small e.g. it equals one then we get all interdisciplinary ideas as result items regardless weather they consists of high or low innovative potential. This is because every idea - that is classified concerning a specific non-related combination of scientific categories - is presented as innovative idea. We estimate that an optimal value of n is between $4 \leq n \leq 8$.

After this, the alpha cut of Jaccard's coefficient results are estimated. The first alpha cut is the set of all terms that represents a means or a purpose such that the corresponding result value by comparing this set to a scientific category is greater than or equal to α_1. With the second alpha cut we identify two related scientific categories only if the appertaining Jaccard's coefficient result value is greater than or equal to α_2. If α_1 is too small or too large then means and purposes are not classified correctly. If α_2 is too small or too large then the identification of relationships among scientific categories fails. This leads both to a small precision and to a small recall value. An optimal value of α_1 and α_2 is estimated between
$5\% \leq \alpha_1, \alpha_2 \leq 20\%$.

To investigate the dependency of the approach on the parameters, we explicitly check if the parameter values are identifiable on the training set. These values are used to compute precision and recall on the test set. For this, we use the estimations for n $\in$ {4, 5, ..., 8} and the percentages $\alpha_1 \in$ {5%, 6%, ..., 20%} and $\alpha_2 \in$ {5%, 6%, ..., 20%}. We identify 5 •16 • 16 = 1280 different parameter combinations of (n, α_1, α_2). The training set is used to compute average precision and recall for each parameter combination to identify the optimal parameter values with a maximal F-measure. The F-measure is used because precision and recall are equally important. As a result, parameter values n = 5, α_1 = 14%, and α_2 = 16% are identified. These parameter values are used to compute precision and recall for each test example and the average precision and recall values for all test examples. We get a precision value of 38% and a recall value of 30%. A precision value of 38% means that if this approach predicts 100 ideas as innovative ideas then 38 of them are innovative. A recall value of 30% means that if there are 10 innovative ideas in the provided text then this approach identifies three of them.

We compare this approach to a baseline model because we are not aware of other approaches for identifying the innovative potential of ideas at the present time. A positive class probability of 5% is already calculated by human experts. This leads to a 5% precision at 30% recall for a random prediction and it shows that this approach is much better than random. We think that the results are sufficient to proof the feasibility of our approach.

Using the 500 new ideas from the test set, the approach automatically computes several innovation fields. We present examples for these innovation fields. They can be found between 'Health Care Sciences and Services' and 'Computer Science, Artificial Intelligence' (e.g. the use of methods from artificial intelligence for health care applications), between 'Imaging Science and Photographic Technology' and 'Medical Informatics', between 'Remote Sensing' and 'Tropical Medicine', and between 'Computer Science, Theory and Methods' and 'Psychiatry'. Then, the approach identifies ideas from these innovation fields as innovative ideas.

This approach can be re-evaluated by using our application for mining innovative ideas (see http://www.text-mining.info). There, the web based application that is programmed in perl/ruby and all texts that are used for evaluation are presented. The application extracts ideas from a provided text, creates terms representing means and purposes, identifies innovation fields, and classifies the ideas as (non-) innovative ideas.

9. Outlook

This work shows that the automatic identification of the innovative potential of new technological ideas is feasible using text classification and specific technological definitions. Further work should aim at enlarging and optimizing this approach e.g. by identifying further properties of innovative ideas. A second avenue of further research could take the granularity of the context information into account e.g. by using technologies rather than scientific categories. This also probably leads to an increasing precision and recall.

Bibliography

1. Abebe, A. J., Guinot, V., Solomatine, D. P. (2000). Fuzzy alpha-cut vs. Monte Carlo techniques in assessing uncertainty in model parameters. In: Proc. 4-th International Conference on Hydroinformatics. Iowa City, USA.
2. Albers, S., Gassmann, O. (2005). Handbuch Technologie- und Innovationsmanagement: Strategie- Umsetzung- Controlling. p. 196. Gabler Verlag.
3. Berth, R. (1997). Der große Innovations-Test: das Arbeitsbuch für Entscheider: Chancen erkennen, Flops vermeiden - Theorie und Praxis des Management of Change. Econ, Düsseldorf.
4. Bürgel, H.D., Haller, C., Binder, M. (1996). F&E-Management. p. 85. Vahlen, München.
5. Disselkamp, M. (2005). Innovationsmanagement: Instrumente und Methoden zur Umsetzung im Unternehmen. p. 179. Gabler Verlag.
6. Feldman, R., Sanger, J. (2007). The Text Mining Handbook: Advanced Approaches in Analyzing Unstructured Data. p. 318. Cambridge University Press.
7. Ferber, R. (2003). Information Retrieval. p. 78. dpunkt.verlag, Heidelberg.
8. Geschka, H., Schauffele, J., Zimmer, C. (2005). Explorative Technologie-Roadmaps - Eine Methodik zur Erkundung technologischer Entwicklugslinien und Potenziale. In Möhrle, M.G., Isenmann, R. (eds.) Technologie-Roadmapping, p. 165. Springer, Berlin, Heidelberg.
9. Guiltinan, J.P., Paul, G.W. (1991). Marketing Management: Strategies and Programs. p. 196. McGraw-Hill.
10. Hotho, A., Nürnberger, A., Paaß, G. (2005). A Brief Survey of Text Mining. LDV Forum 20 (1), 19-26.
11. Institute for Scientific Information ISI (eds.) (1997) SCI Journal Citation Reports.
12. Lustig, G. (1986). Automatische Indexierung zwischen Forschung und Anwendung. p. 92. Georg Olms Verlag, Hildesheim.

13. Mckeown M. (2008). The Truth About Innovation. Pearson Education, Harlow.

14. Möslein, K.M., Matthaei, E.E. (2008). Strategies for Innovators: A Case Book of the HHL Open School Initiative. p. 13. Gabler Verlag.

15. Reiß, T. (2006). Innovationssysteme im Wandel - Herausforderungen für die Innovationspolitik. In Müller, B., Glutsch, U. (eds.) Fraunhofer-Institut für System- und Innovationsforschung - Jahresbericht 2006, p. 10. Karlsruhe.

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

17. Rohpohl, G. (1996). Das Ende der Natur. In: Schäfer, L., Sträker, E. (eds.) Naturauffassungen in Philosophie, Wissenschaft und Technik, Bd. 4, pp. 143-163. Alber, Freiburg, München.

18. Thorleuchter, D. (2008). Finding Technological Ideas and Inventions with Text Mining and Technique Philosophy. In: Preisach, C., Burkhardt, H., Schmidt-Thieme, L., Decker R. (eds.) Data Analysis, Machine Learning, and Applications, pp. 413-420. Springer, Berlin, Heidelberg.