MEASURING PATENT NOVELTY USING NATURAL LANGUAGE PROCESSING

Yassine, Ali;
Lipizzi, Carlo

Stevens Institute of Technology

ABSTRACT

This paper develops a novelty measure for patents. We devise a text-based novelty measure using natural language processing (NLP) techniques. The proposed method is applied on patents that belong to a common category, which represents a subset of patents under a specific patent class. We then extract the novelty-value profile of those patents and discuss a use case for product design and development (i.e., extracting patent novelty and predicting inventive value).

Keywords: New product development, Machine learning, Open source design

Contact:
Yassine, Ali
Stevens Institute of Technology
United States of America
ayassine@stevens.edu

Cite this article: Yassine, A., Lipizzi, C. (2023) 'Measuring Patent Novelty Using Natural Language Processing', in *Proceedings of the International Conference on Engineering Design (ICED23),* Bordeaux, France, 24-28 July 2023. DOI:10.1017/pds.2023.261

https://doi.org/10.1017/pds.2023.261 Published online by Cambridge University Press

1 INTRODUCTION

The prevailing industrial revolution, Industry 4.0, holds the promise of intricate customizability, better productivity, and faster workflow from conceptualization to delivery (Zhong et al., 2017). The need for a shorter time to market, coupled with improved quality and sustained profits, necessitates a revamped product design process (Brossard et al., 2018). The constant push for this has led product designers to utilize various data sources in order to optimize the design process (Bertoni, 2020). However, the volume of dynamically changing data generated throughout the lifecycle of products is growing (Kuo and Kusiak, 2018). Embedded within this data is vital information regarding the working of the product, in case the data was collected during the usage of the product, or spatial and temporal data generated by systems in the manufacturing phase. This data can then be analysed, and meaningful insights can be gleaned, which feeds back into the data-generating phase of the design process itself so it can be further enhanced (Bertoni, 2020).

In this paper, we select a design-relevant data source, patent data, which plays a significant role in the conceptual and detailed design phases of product development and apply machine learning techniques to extract useful information from patents. Patent data is essential in determining what is already in existence, which greatly influences the current design direction of new products and associated patents filed. As patent data gets bigger and more detailed as of recent years, the need to have better means of exploring this data, rather than manually paging through it, is becoming more relevant, if not necessary. It is important for product designers to understand patent data and how it affects their design decisions without investing too much time and effort sifting through the patents (Aristodemou and Tietze, 2018).

Previous literature links patents and innovation together and attempts to study the variation of patent innovative value with novelty. To measure patent novelty, it uses citations as a proxy (He and Luo, 2017). This method however has drawbacks, particularly the failure to account for non-patent citations (e.g., research published in a journal), which may have different qualities that are discarded when using this method. We develop a new approach to measure patent novelty and its effect on value by proposing a text-based method. We rely on patent text to determine patent novelty with respect to the corpus. This allows us to have an alternative way to evaluate patent novelty while not relying on citations as a sole measure, which mitigates its weaknesses.

2 LITERATURE REVIEW

As patent databases are getting bigger and as the need for better landscaping techniques is getting more relevant due to the complex interactions between different sub-fields of patents, patent landscaping techniques evolved with data science and adopted much of its procedures. Automated patent landscaping processes allow for joint human and machine efforts to leverage human domain knowledge and machine characteristics generating high quality patent landscapes with reduced effort (Abood and Feltenberger, 2018).

When developing a patentable product, a product designer is faced with the question of how much novel or different the product or invention should be in order to have the most value. Previous research has suggested the existence of a ‘sweet-spot’ when it comes to patent novelty when compared to the existing corpus (He and Luo, 2017). Contrary to intuition, having a higher novelty would not translate into better value due to the additional risks introduced and thus more exposure to variability (Fleming, 2001). Should the product be of too low or too high novelty, the value decreases on average and thus would not be optimal to target. A product designer should target an in-between novelty level where historically, patents are most likely to have most value (He and Luo, 2017).

Patents are considered to be a representative of inventive activities which hold a certain value, and can be seen as a reliable measure of innovation (Di Guardo and Harrigan, 2016). Patents, however, vary in value which is the measure of innovative qualities of the patent. Forward citation are considered a proxy to measure patent value or impact (Albert et al., 1991; Fischer and Leidinger, 2014), and are considered a direct sign of invention size (Lee et al., 2007). It is also argued that forward citations (for

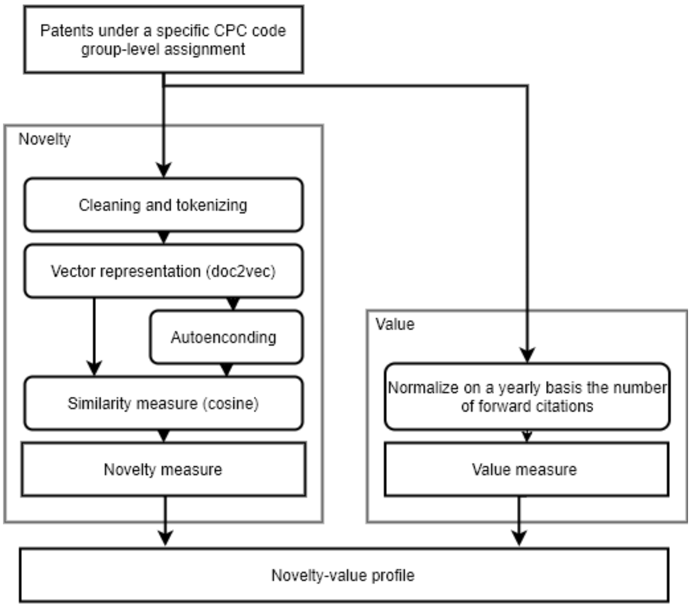
https://doi.org/10.1017/pds.2023.261 Published online by Cambridge University Press

a patent) is highly correlated with its reported economic value (Harhoff et al., 1999; Park and Park, 2006).

Finally, a survey paper describing the state-of-the-art in intellectual property (IP) analysis explains different ways to analyse patent data (Aristodemou and Tietze, 2018). It discusses the current approaches in many different domains such as artificial intelligence and machine learning. Autoencoders in a patent context are not mentioned in the paper, which signifies that the application of autoencoders for analyzing IP data is a novel and would further contribute to the field.

3 METHODOLOGY

We discuss an approach to harness patent data in order to measure patent innovative value relative to its novelty under a specific area in the corpus. In this case, we are interested in having a homogeneous corpus of patents due to the fact the text-based novelty measures require it. These patents should fall under one area in order to have a common ground to measure similarity with. Figure 1 shows the various steps of the proposed methodology.



Patents under a specific CPC code group-level assignment
Novelty
Cleaning and tokenizing
Vector representation (doc2vec)
Autoenconding
Similarity measure (cosine)
Novelty measure
Value
Normalize on a yearly basis the number of forward citations
Value measure
Novelty-value profile

Figure 1. The methodology flowchart

To have a set of patents representative of one area, we select patents in a group-level class according to the CPC code. Only patents older than a several number of years and thus known forward citations are selected because they have a known value measure. Then, we clean the data and convert it to its mathematical embedded form. Additionally, to detect the novelty of each patent relative to the group, we compare it to a special copy of itself thus finding a novelty measure. Finally, we find the innovative value profile of patents over novelty values.

3.1 Obtaining patents

Many sources offer patent data with varying degrees of features and limitations. We decided to work with a source that encompasses multiple patent registries in a readable format that is easily fetchable, and thus we chose Google Cloud Platform (hereafter GCP) patent database, which is hosted by Google and allows for faster SQL queries due to the huge size of the data (1.8 TB). This offers a more approachable way to interact with patent data, as well as the data being updated frequently.

In order to get the patent corpus, we run a query to get patents exhaustively under a specific classification group as dictated by the CPC. This is because we want to get area-specific patents where we know beforehand that they belong to one area, and thus we can assess similarity later.

The fields shown in Table 1 were extracted to our offline storage. We will perform different operations on this data in the following sections.

https://doi.org/10.1017/pds.2023.261 Published online by Cambridge University Press

Table 1. The selected fields to be saved for further analysis

Value	Description
Ab.text	The patent abstract text localized, with mixed text and HTML text formatting
Publication_date	The publication date of the patent, displayed as YYYYMMDD
Publication_number	Publication number of the patent, as in the patent signified ‘number’
Ab.Language	The language of the patent to be used in filtering
country_code	The country code of the issuing party
c.code	The classification code (CPC) of the patent

3.2 Measure of patent novelty

3.2.1 Cleaning and tokenizing

While the data acquired from a data source would hold the information inside, fields themselves may not be clean enough in order to perform meaningful analysis. Even though the data is present in a readable format, cleaning and tidying the data is important before attempting to manipulate it. Our data is segmented data that can be parsed correctly, yet fields are not guaranteed to be error-free due to human mistakes and other factors. For example, if an author’s name is listed but spelled differently under two patents, it will in turn carry into and falsify the results. Thus, cleaning the patents should be done either manually, or using other tools. Mainly, we would be having an iterative process of tidying the data until we reduce irregularities to a minimum.

3.2.2 Vectorization

In order to analyze the data, we must transform it into a mathematical form as opposed to text. To transform the text into a feature representation, we used the doc2vec algorithm, a direct upgrade of word2vec, to learn constant-length representations from variable-length text as opposed to other frequency-based methods.

The word2vec method tries to solve one of the main problems of NLP, the loss of meaning of words after “one-hot” encoding (Mikolov et al., 2013). For example, if we encode “Detroit”, “Michigan”, and “Pizza” into labels they lose their meanings and their relation to each other. Clearly, “Detroit” is closer to “Michigan” as a term than to “Pizza”. Word2vec used two main methods to do this, one being using one word to predict the context, called Skip-gram, the other is to predict the target word knowing the context, Continuous bag-of-words (CBOW). For Skip-gram, which is considered slower, the objective function is given in Equation (1):

$$J(\theta) = -\frac{1}{V}\sum_{t=1}^{V}\sum_{-m\leq j\leq m}\log p(w_{t+j}|w_t) \tag{1}$$

Here, θ contains both the input vector representation of words and the output representation vectors, and m is the length of the window around text selected. For m = 1, a center word would have the word before and after it fed as input. The objective function is then minimized by stochastic gradient descent which is the loss to be minimized.

Expanding on this, we will attempt to use the doc2vec method to evaluate the data at hand. The doc2vec algorithm (Le and Mikolov, 2014) is an upgrade on the word2vec model that allows to use the whole patent as one vector entity.

https://doi.org/10.1017/pds.2023.261 Published online by Cambridge University Press

The Doc2vec method adds to the previous by adding a paragraph ID, which in our case would be each patent. This method has been proven to capture not only word but document similarity, which when applied to our context, patents, would mean capturing the details of each patent. It maps each patent to a vector that is then trained by a neural network to predict the context.

In our case, we would be using the distributed memory (DM) method, which is similar to the CBOW. This method introduces a new document-level token in addition to the words. However, the vectors are not summed but concatenated. The modified objective is to predict the target word knowing the concatenated word and document. The parameters of the classifier and the word vectors are not needed, and backpropagation is used to tune the paragraph vectors.

It is important to say that although the model is fairly high in dimensions, a certain loss is expected to happen since some semantic value is lost in the conversion coupled with the limited size of the vector and other factors which may contribute to it.

3.2.3 Autoencoding

In this section we use autoencoding to identify the novelty of each patent. An autoencoder learns a compressed representation of an input data of a fixed vector size, then tries to construct the initial data from that. It consists of 3 parts, the encoder which reduces the input into an encoded vector, the bottleneck which is the most compressed vector the data passes through, then the decoder which constructs the data from the previous step to try to emulate the input, with a varying degree of loss in the process. An autoencoder neural network tries to compress the input data through its bottleneck layer, then tries to rebuild it for the output. When passing an input that is expected, the autoencoder can reconstruct it with considerably better results than when passing outliers. For example if we train the autoencoder on documents related to a certain topic C, then try to predict 2 documents one from the same topic and one an outlier, the outlier predicted form will look dissimilar to its initial form, while the common document will have a fairly closer predicted copy.

An autoencoder was used in order to reduce and compress features in the input, then construct back a copy that when compared to the input, allows us to find a certain deviation. For each patent to have a correctly predicted vector, we only included chronologically older patents in the training data of the autoencoder for each patent we are trying to predict. This means the corpus training data includes patents up to the patent of interest, then the patent of interest is predicted using the autoencoder. This process is run on a random set of 100 patents to have enough data points to plot.

By having a predicted output from the autoencoder, and as the autoencoder removes the noise when encoding to a fewer features layer, unique patents would lose their content when being encoded, then the decoded output would lack the content of the input. This means when comparing each unique patent with its predicted copy, the two will not look alike. Alternatively, when comparing a patent that has no unique content compared to the corpus, its copy would keep most of the features intact if slightly different and would be relatively similar to its original form.

3.2.4 Novelty measure

Since the patent texts are represented as vectors extracted from textual data, the use of distance metric to measure similarity cannot be square distance or Euclidean as text data needs normalization to correctly get similar entries. The cosine similarity measure works with vector representations of text (Huang, 2008), and we use cosine similarity to find the similarity between the patent and the corpus. Equation (2) determines the similarity between the initial patent and its autoencoder version.

$$sim(P_i, P_j) = \frac{P_i P_j}{||P_i|| \; ||P_j||} \tag{2}$$

Our novelty measure is the complement of the similarity measure. Since unique patents would reconstruct poorly, the output will drastically change, which means it will tend to have a lower similarity or higher novelty when compared with the initial vectors. Likewise, common patents would

https://doi.org/10.1017/pds.2023.261 Published online by Cambridge University Press

have a lower novelty or better similarity value. Thus, a patent with a low similarity score is considered more novel compared to the corpus.

3.2.5 Value measure

To get the patent value, we use the forward citations as a measure (Albert et al., 1991; Fischer and Leidinger, 2014). The value of a patent or an invention is its significance in various social and economic aspects. The value of the patent is strongly correlated with the number of forward citations it gets. We normalize the citations on a yearly basis to account for variation in average citations per patent changes.

After getting both the novelty and value measures, we can construct the plot that shows the change of novelty and how it affects value. Previous research tries to find the relation between patent novelty and value using a citation-based novelty approach (He and Luo, 2017). Diverging from that, our novelty measure, which is based on text, attempts to rectify some of the limitations of the previous method by offering another point of view into novelty. We attempt to identify the novelty and value profile in order to identify any meaningful relations between these variables. Citation-based measures had a 'sweet-spot' of novelty where value was maximized. We try to verify if local patent corpus under a specific classification level feature the same characteristics of a certain 'sweet spot' where patents have the best value.

3.2.6 Validation of novelty approach

Since our novelty measure method is unsupervised, it is hard to validate due to the lack of any labelling. We treat the problem as a supervised one to have some insight on how well it performs.
To get labels on novel or non-novel data in order to validate our autoencoder, we trained our doc2vec model on our data then generated 5000 artificial data points outside the bounding hypercube of the patent vectors in the feature space with a maximum offset of 5 in either direction. The artificial data is around 7% of the normal data size. For example, for a specific feature with a minimum of -2 and a maximum of 1, the artificial data can only be in the intervals [-7,-2) or (1,6] respectively. This data is then labelled as novel while the rest is labelled otherwise. This is the foundation behind our ground truth. We then split the data into training and testing data with an 80/20 split. We then train the autoencoder on the training data and predict the testing data output. The testing data predicted value is then compared with the input using our novelty measure. We assess the approach by using the ground truth we labeled earlier. We treat the problem as a classification problem, where one group of patents would be classified as novel and the rest would not be. This allows us to construct a receiver operating characteristic (ROC) curve to evaluate the method, as shown in Figure 2.

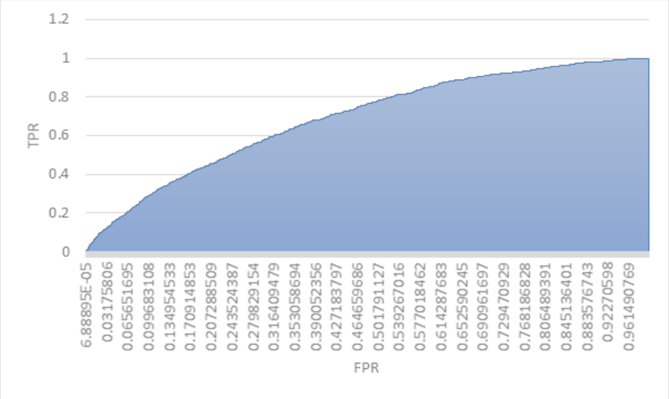


Figure 2. The ROC curve

An ROC curve allows us to determine the separability measure, or in other terms the ability to distinguish between categories. The curve explains how well our approach was able to identify the novelty in patents compared to our ground truth, by measuring the True Positive Rate (TPR), the ratio of identified true novel patents to the total novel patents, and its variation with the False Positive Rate (FPR), the ratio of identified falsely novel patents to the total common patents. The

https://doi.org/10.1017/pds.2023.261 Published online by Cambridge University Press

method was able to identify our truths of novel patents fairly accurately with an Area Under Curve (AUC) value of 0.7.

4 CASE STUDY

4.1 Obtaining and converting patents to vectors

We chose an arbitrary group of patents under the group-level classification. The reason would be that since we are aiming to identify similarity, our data has to take shape as reporting to a clearly defined description, which the group level in the CPC achieves. For example, the chosen patent set belongs mainly to prosthetics. Table 4 shows the chosen group level class code and description as shown in the CPC.

Table 2. The chosen CPC group code

A61F2/	Filters implantable into blood vessels; Prostheses, i.e. artificial substitutes or replacements for parts of the body; Appliances for connecting them with the body; Devices providing patency to, or preventing collapsing of, tubular structures of the body

The patents are then prepared and the text cleaned from noise, filtered, then tokenized. The list of patents totalling 72,313 is then fed into the doc2vec algorithm in order to get the document vector representation of the corpus.

4.2 Applying the text-based novelty measure

The vectors are then used to run several instances of the autoencoder where training data is used up to the patent being used to output a predicted version. The number of instances ran was 100. The cosine similarity of each patent's input vector and the predicted one is calculated, and the result was stored to be analysed.

4.3 The novelty-value profile

After getting the novelty measure, we get the value measure which is the forward citations per each patent in the corpus. We then plot novelty and value, where novelty on the x-axis is compared with the value measure on the y-axis as shown in Figure 3 (left side). We then fit a regression curve to the data to determine the average value for different novelty levels as shown in Figure 3 (right side).

Our results show that for a specific range of novelty for our local corpus (near 0.95), the average value measure reaches a crest which signifies better value for patents in this range. This shows a 'sweet spot' of novelty that should be achieved in order to have a better expected value measure, with the measure being text-based rather than citation-based (He and Luo, 2017).

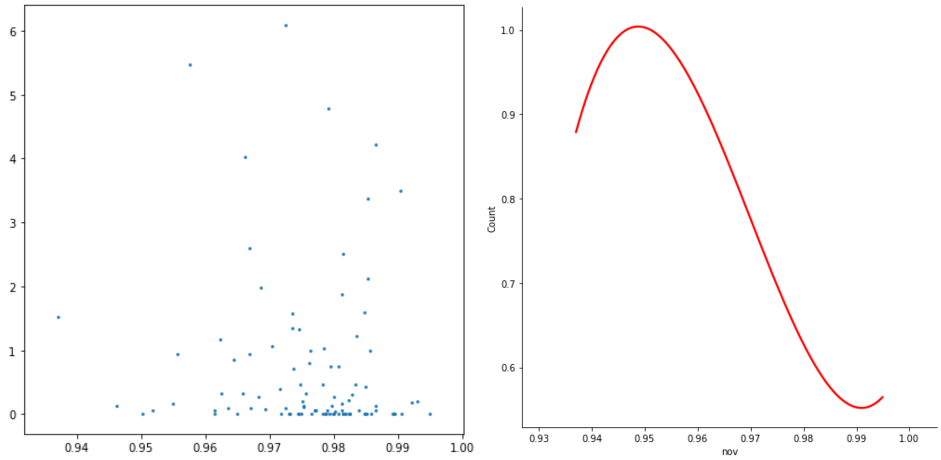


Figure 3. The scatter plot of the patent corpus value-novelty profile (left) and fitted curve (right)

https://doi.org/10.1017/pds.2023.261 Published online by Cambridge University Press

4.4 Forecasting a Patent's future value by using its abstract

We took a highly relevant case in the product design field that is a possible application to our method. A product developer would want to assess a patent or invention’s actual value before submitting it or developing a full patent application. As our novelty measure is text-based, a patentee could simply predict the value of a draft abstract to a certain degree by using the trained model on existing corpora. In our case, we predict patent novelty using our model which will then be used to predict the innovative value of the patent in use, which is built using corpus data. This allows the patentee to estimate patent value relative to the corpus.

To test the model, we create a synthesized abstract text by concatenating both texts from 2 abstracts present in the corpus shown in Table 3. This new abstract was then fed into the model which is trained on the corpus data. The model outputted a 0.96 novelty value. This in turn, when applied to the novelty and value relationship gave an expected innovative value of ratio 1, which is relatively high and falls in the sweet spot of the novelty measure.

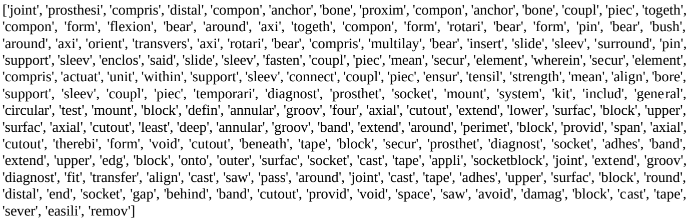
Table 3. The two abstracts

Patent number: US-2014148913-A1
Abstract
A joint prosthesis comprises a distal component for anchoring to a first bone a proximal component for anchoring to a second bone and a coupling piece that together with the first component forms a flexion bearing around a first axis and together with the second component forms a rotary bearing formed by the pin and the bearing bush around a second axis oriented transversely to the first axis The rotary bearing comprises a multilayer bearing insert having a sliding sleeve surrounding the pin and a support sleeve that encloses said sliding sleeve and is fastened to the coupling piece by means of a securing element wherein the securing element comprises an actuation unit within the support sleeve and can be connected to the coupling piece such as to ensure tensile strength by means of two aligned bores in the support sleeve and the coupling piece

Patent number: US-2010211188-A1
Abstract
A temporary diagnostic prosthetic socket mounting system and kit including a generally circular test mounting block defined by an annular groove and four axial cutouts extending from the lower surface of the block to the upper surface The axial cutouts are at least as deep as the annular groove A band extending around the perimeter of the block is provided which spans the axial cutouts thereby forming a void in the cutouts beneath the tape The block is secured to a prosthetic diagnostic socket by adhesive with the band extending up over the upper edge of the block and onto the outer surface of the socket Casting tape is applied over the socketblock joint extending into the groove After diagnostic fitting and transfer of the alignment a cast saw can be passed around the joint through the casting tape and into the adhesive between the upper surface of the block and the rounded distal end of the socket The gap behind the band in the cutouts provides a void space for the saw that avoids damage to the block and the casting tape is severed and easily removed

Another aspect to consider is the sensitivity of the model to changes in the abstract text. We ran a sensitivity analysis on the text to measure the change in the output when the input changes. To do this, we randomly removed ten terms from the synthesized text and ran it through the model. The inferred vector from the summation of the two documents is shown in Figure 4. The resulting value was 0.963, a small deviation from the initial value.

https://doi.org/10.1017/pds.2023.261 Published online by Cambridge University Press



```
['joint', 'prosthesi', 'compris', 'distal', 'compon', 'anchor', 'bone', 'proxim', 'compon', 'anchor', 'bone', 'coupl', 'piec', 'togeth',
'compon', 'form', 'flexion', 'bear', 'around', 'axi', 'togeth', 'compon', 'form', 'rotari', 'bear', 'form', 'pin', 'bear', 'bush',
'around', 'axi', 'orient', 'transvers', 'axi', 'rotari', 'bear', 'compris', 'multilay', 'bear', 'insert', 'slide', 'sleev', 'surround', 'pin',
'support', 'sleev', 'enclos', 'said', 'slide', 'sleev', 'fasten', 'coupl', 'piec', 'mean', 'secur', 'element', 'wherein', 'secur', 'element',
'compris', 'actuat', 'unit', 'within', 'support', 'sleev', 'connect', 'coupl', 'piec', 'ensur', 'tensil', 'strength', 'mean', 'align', 'bore',
'support', 'sleev', 'coupl', 'piec', 'temporari', 'diagnost', 'prosthet', 'socket', 'mount', 'system', 'kit', 'includ', 'general',
'circular', 'test', 'mount', 'block', 'defin', 'annular', 'groov', 'four', 'axial', 'cutout', 'extend', 'lower', 'surfac', 'block', 'upper',
'surfac', 'axial', 'cutout', 'least', 'deep', 'annular', 'groov', 'band', 'extend', 'around', 'perimet', 'block', 'provid', 'span', 'axial',
'cutout', 'therebi', 'form', 'void', 'cutout', 'beneath', 'tape', 'block', 'secur', 'prosthet', 'diagnost', 'socket', 'adhes', 'band',
'extend', 'upper', 'edg', 'block', 'onto', 'outer', 'surfac', 'socket', 'cast', 'tape', 'appli', 'socketblock', 'joint', 'extend', 'groov',
'diagnost', 'fit', 'transfer', 'align', 'cast', 'saw', 'pass', 'around', 'joint', 'cast', 'tape', 'adhes', 'upper', 'surfac', 'block', 'round',
'distal', 'end', 'socket', 'gap', 'behind', 'band', 'cutout', 'provid', 'void', 'space', 'saw', 'avoid', 'damag', 'block', 'cast', 'tape',
'sever', 'easili', 'remov']
```

Figure 4. The inferred vector from the summation of the two documents

5 CONCLUSION

The product development lifecycle exhibits many big data flows of internal or external sources and destinations. Until recently, means of analysing these data flows were severely limited due to performance and storage limits. With the advancement in technology, one can utilize these data flows to improve the product development process. In this paper, we select a relevant data source, patent data, which plays a significant role in the conceptual and detailed design phases of product development process.

In this paper, we proposed a model to extract text representations from a specific patent group under a defined classification code, then extract a text-based patent novelty measure with respect to other patents in a local corpus along with a value measure. We then studied the variation of patent value with novelty and identified a target range of novelty that the value of patents is the best at. The text-based novelty measure would complement other citation-based measures and would help product developers acquire a better idea on the novelty-value relation.

There are several decisions made when constructing our model, which could introduce limitations and drawbacks in the intended result in retrospect. These limitations are discussed with the reasoning behind each and its respective effects. We chose to run text mining on the abstracts rather than the claims due to the following reasons. First, while naturally claims text is the more descriptive part of the patent stating each claim with detail, abstract text is sufficiently inclusive of what the patent is about (Adams, 2010). Additionally, claims often contain complex elements and characters, which are a challenge to clean in an automated way, and may introduce further noise in the result which diminishes any perceived improvement. Finally, although claim text is more important when determining the patent grant and is the most edited and revised text, it also means that patent lawyers are more exposed to it and that would mean more specifically reworded text to match certain criteria.

REFERENCES

Abood, A., and Feltenberger, D. (2018), "Automated patent landscaping", *Artificial Intelligence and Law*, Vol. 26, No. 2, pp. 103–125. https://doi.org/10.1007/s10506-018-9222-4

Adams, S. (2010), "The text, the full text and nothing but the text: Part 1–Standards for creating textual information in patent documents and general search implications", *World Patent Information*, Vol. 32, No. 1, pp. 22–29. https://doi.org/10.1016/j.wpi.2009.06.001

Albert, M. B., Avery, D., Narin, F. and McAllister, P. (1991), "Direct validation of citation counts as indicators of industrially important patents", *Research Policy*, Vol. 20, No, 3, pp. 251–259. https://doi.org/10.1016/0048-7333(91)90055-u

Aristodemou, L., and Tietze, F. (2018), "The state-of-the-art on Intellectual Property Analytics (IPA): A literature review on artificial intelligence, machine learning and deep learning methods for analysing intellectual property (IP) data", *World Patent Information*, 55, 37–51. https://doi.org/10.1016/j.wpi.2018.07.002

Bertoni, A. (2020), "Data-driven design in concept development: systematic review and missed opportunities". In *Proceedings of the Design Society: DESIGN Conference*, Vol. 1, pp. 101-110, Cambridge University Press. https://doi.org/10.1017/dsd.2020.4

Brossard, M., Erntell, H., and Hepp, D. (2018), "Accelerating product development: The tools you need now", *McKinsey Quarterly*.

Di Guardo, M. C., and Harrigan, K. (2016), "Shaping the path to inventive activity: The role of past experience in R&D alliances", *The Journal of Technology Transfer*, Vol. 41, No. 2, pp. 250–269. https://doi.org/10.1007/s10961-015-9409-8

Fischer, T., and Leidinger, J. (2014), "Testing patent value indicators on directly observed patent value—An empirical analysis of Ocean Tomo patent auctions", *Research Policy*, Vol. 43, No. 3, pp. 519–529. https://doi.org/10.1016/j.respol.2013.07.013

Fleming, L. (2001), "Recombinant uncertainty in technological search", *Management Science*, Vol. 47, No. 1, pp. 117–132. https://doi.org/10.1287/mnsc.47.1.117.10671

Harhoff, D., Narin, F., Scherer, F. M., and Vopel, K. (1999), "Citation frequency and the value of patented inventions", *Review of Economics and Statistics*, Vol. 81, No. 3, pp. 511–515. https://doi.org/10.1162/003465399558265

He, Y., and Luo, J. (2017), "The novelty 'sweet spot' of invention", *Design Science*, Vol. 3, p.e21. https://doi.org/10.1017/dsj.2017.23

Huang, A. (2008), "Similarity measures for text document clustering", *Proceedings of the Sith New Zealand Computer Science Research Student Conference* (NZCSRSC2008), Christchurch, New Zealand, 4, 9–56.

Kuo, Y.-H., and Kusiak, A. (2018). "From data to big data in production research: The past and future trends", *International Journal of Production Research*, pp. 1–26. https://doi.org/10.1080/00207543.2018.1443230

Le, Q., and Mikolov, T. (2014), "Distributed representations of sentences and documents", *International Conference on Machine Learning*, pp. 1188–1196.

Lee, Y.-G., Lee, J.-D., Song, Y.-I., and Lee, S.-J. (2007), "An in-depth empirical analysis of patent citation counts using zero-inflated count data model: The case of KIST", *Scientometrics*, Vol. 70, No. 1, pp. 27–39. https://doi.org/10.1007/s11192-007-0102-z

Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013), "Efficient estimation of word representations in vector space", ArXiv Preprint ArXiv:1301.3781.

Park, G., and Park, Y. (2006), "On the measurement of patent stock as knowledge indicators," *Technological Forecasting and Social Change*, Vol. 73, No. 7, pp. 793–812. https://doi.org/10.1016/j.techfore.2005.09.006

Zhong, R. Y., Xu, X., Klotz, E., and Newman, S. T. (2017). Intelligent manufacturing in the context of industry 4.0: A review. *Engineering*, Vol. 3, No. 5, pp. 616–630. https://doi.org/10.1016/j.eng.2017.05.015

https://doi.org/10.1017/pds.2023.261 Published online by Cambridge University Press