PLOS ONE

RESEARCH ARTICLE

Automating the search for a patent's prior art with a full text similarity search

Lea Helmers[1☯], Franziska Horn[1☯*], Franziska Biegler[2], Tim Oppermann[2], Klaus-Robert Müller[1,3,4*]

1 Machine Learning Group, Technische Universität Berlin, Berlin, Germany, **2** Pfenning, Meinig & Partner mbB, Berlin, Germany, **3** Department of Brain and Cognitive Engineering, Korea University, Anam-dong, Seongbuk-gu, Seoul 02841, Korea, **4** Max-Planck-Institut für Informatik, Saarbrücken, Germany

☯ These authors contributed equally to this work.
* franziska.horn@campus.tu-berlin.de (FH); klaus-robert.mueller@tu-berlin.de (KRM)

Citation: Helmers L, Horn F, Biegler F, Oppermann T, Müller K-R (2019) Automating the search for a patent's prior art with a full text similarity search. PLoS ONE 14(3): e0212103. https://doi.org/10.1371/journal.pone.0212103

Editor: Bridget McInnes, Virginia Commonwealth University, UNITED STATES

Received: February 9, 2018

Accepted: January 28, 2019

Published: March 4, 2019

Data Availability Statement: The source code and datasets generated during the current study are available online: https://github.com/helmersl/patent_similarity_search.

Funding: This work was supported by the Federal Ministry of Education and Research (BMBF) for the Berlin Big Data Center BBDC (01IS14013A) and Berlin Center for Machine Learning BZML (01IS18037I), as well as the Institute of Information & Communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (No. 2017-0-00451). The

Abstract

More than ever, technical inventions are the symbol of our society's advance. Patents guarantee their creators protection against infringement. For an invention being patentable, its novelty and inventiveness have to be assessed. Therefore, a search for published work that describes similar inventions to a given patent application needs to be performed. Currently, this so-called search for prior art is executed with semi-automatically composed keyword queries, which is not only time consuming, but also prone to errors. In particular, errors may systematically arise by the fact that different keywords for the same technical concepts may exist across disciplines. In this paper, a novel approach is proposed, where the full text of a given patent application is compared to existing patents using machine learning and natural language processing techniques to automatically detect inventions that are similar to the one described in the submitted document. Various state-of-the-art approaches for feature extraction and document comparison are evaluated. In addition to that, the quality of the current search process is assessed based on ratings of a domain expert. The evaluation results show that our automated approach, besides accelerating the search process, also improves the search results for prior art with respect to their quality.

Introduction

A patent is the exclusive right to manufacture, use, or sell an invention and is granted by the government's patent offices [1]. For a patent to be granted, it is indispensable that the described invention is not known or easily inferred from the so-called prior art, where prior art includes any written or oral publication available before the filing date of the submission. Therefore, for each application that is submitted, the responsible patent office performs a search for related work to check if the subject matter described in the submission is inventive enough to be patentable [1]. Before handing in the application to the patent office, the inventors will usually consult a patent attorney, who represents them in obtaining the patent. In order to assess the chances of the patent being granted, the patent attorney often also performs a search for prior art.

funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript. Pfenning, Meinig & Partner mbB provided support in the form of salaries for authors FB and TO, but did not have any additional role in the study design, data collection and analysis, decision to publish, or preparation of the manuscript. The specific roles of these authors are articulated in the 'author contributions' section.

Competing interests: LH, FH, and KRM declare no competing interests. FB and TO are affiliated with Pfenning, Meinig & Partner mbB, Berlin. This does not alter our adherence to PLOS ONE policies on sharing data and materials.

When searching for prior art, patent officers and patent attorneys are currently mainly relying on simple keyword searches such as those implemented by the Espacenet tool from the *European Patent Office*, the TotalPatent software developed by LexisNexis, or the PatSnap patent search, all of which provide very limited *semantic* search options. These search engines often fail to return relevant documents and due to constraints regarding the length of the entered search text, it is usually not possible to consider a patent application's entire text for the search, but merely query the database for specific keywords.

Current search approaches for prior art therefore require a significant amount of manual work and time, as given a patent application, the patent officer or attorney has to manually formulate a search query by combining words that should match documents describing similar inventions [2]. Furthermore, these queries often have to be adapted several times to optimize the output of the search [3, 4]. A main problem here is that regular keyword searches do not inherently take into account synonyms or more abstract terms related to the given query words. This means, if for an important term in the patent application a synonym, such as *wire* instead of *cable*, or a more specialized term, such as *needle* instead of *sharp object*, has been used in an existing document of prior art, a keyword search might fail to reveal this relation unless the alternative term was explicitly included in the search query. This is relevant as it is quite common in patent texts to use very abstract and general terms for describing an invention in order to maximize the protective scope [5, 6]. A line of research [7–11] has focused on automatically expanding the manually composed queries, e.g., to take into account synonyms collected in a thesaurus [9, 12] or include keywords occurring in related patent documents [13–15]. Yet, with iteratively augmented queries—be it by manual or automatic extension of the query—the search for prior art remains a very time consuming process.

Furthermore, a keyword-based search for prior art, even if done with most professional care, will often produce suboptimal results (as we will see e.g. later in this paper and in Section *D2* in S1 File). With possibly imperfect queries, it must be assumed that relevant documents are missed in the search, leading to *false negatives* (FN). On the other hand, query words can also appear in texts that, nonetheless, have quite different topics, which means the search will additionally yield many *false positives* (FP). When searching for prior art for a patent application, the consequences of false positives and false negatives are quite different. While false positives cause additional work for the patent examiner, who has to exclude the irrelevant documents from the report, false negatives may lead to an erroneous grant of a patent, which can have profound legal and financial implications for both the owner of said patent as well as competitors [16].

An approach to automate the search for prior art

To overcome some of these disadvantageous aspects of current keyword-based search approaches, it is necessary to decrease the manual work and time required for conducting the search itself, while increasing the quality of the search results by avoiding irrelevant patents from being returned, as well as automatically accounting for synonyms to reduce false negatives. This can be achieved by comparing the patent application with existing publications based on their *entire texts* rather than just searching for specific keywords. By considering the entire texts of the documents, much more information, including the context of keywords used within the respective documents, is taken into account. For humans it is of course infeasible to read the whole text of each possibly relevant document. Instead, state-of-the-art text processing techniques can be used for this task.

This paper describes a novel approach to automate the search for prior art with *natural language processing* (NLP) and *machine learning* (ML) techniques, such as neural network

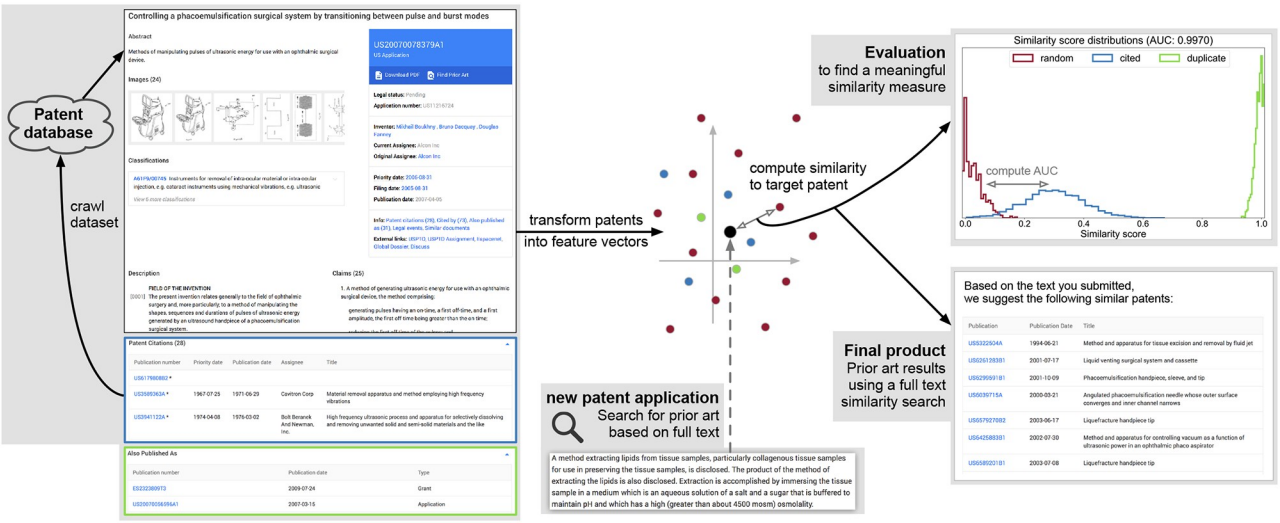

Fig 1. Illustration of the presented novel approach to the search for a patent's prior art. First, a dataset of patent applications is obtained from a patent database using a few manually selected seed patents and recursively including the patent applications they cite. Then, the patent texts are transformed into feature vectors and the similarity between two documents is computed based on said feature vectors. Finally, patents that are considered as very similar to a new target patent application are returned as possible prior art. An appropriate similarity measure for this process should assign high similarity scores to related patents (e.g. where one patent was cited in the search report of the other) and low scores to unrelated (randomly paired) patents. We compare different similarity measures by quantifying the overlap between the respective similarity score distributions of pairs of related documents and randomly paired patents using the AUC score.

https://doi.org/10.1371/journal.pone.0212103.g001

language models, in order to make it more efficient and accurate. The essence of this idea is illustrated in Fig 1. We first obtain a dataset of related patents from a patent database by using a few manually selected seed patents and then recursively adding the patents or patent applications that are cited by the documents already included in the dataset. The patent texts are then transformed into numerical feature vectors, based on which the similarity between two documents can be computed. We evaluate different similarity measures by comparing the prior art suggested by our automated approach to those documents that were originally cited in a patent's search report and, in a second step, to documents considered relevant prior art for this patent by a patent attorney. By analyzing and comparing different approaches for computing full text similarities between patent documents, we aim to identify a similarity measure based on which it is possible to automatically and reliably select relevant prior art given, e.g., the draft of a new patent application.

The remainder of the paper is structured as follows: After briefly reviewing existing strategies for prior art search as well as machine learning methods for full text similarity search and its applications, we discuss our approach for computing the similarities between the patents using different feature extraction methods. These methods are then evaluated on an example dataset of patents including their citations, as well as a second dataset where relevant patents were identified by a patent attorney. Furthermore, based on this manually annotated dataset, we also assess the quality of the original citation process itself. A discussion of the relevance of the obtained results and a brief outlook conclude this manuscript.

Related work

Most research concerned with facilitating and improving the search for a patent's prior art has focused on automatically composing and extending the search queries. For example, a manually formulated query can be improved by automatically including synonyms for the keywords

using a thesaurus [5, 9, 12, 17, 18]. A potential drawback of such an approach, however, is that the thesaurus itself has to be manually curated and extended [19]. Another line of research focuses on pseudo-relevance feedback, where, given an initial search, the first *k* search results are used to identify additional keywords that can be used to extend the original query [3, 14, 20]. Similarly, past queries [21] or meta data such as citations can be used to augment the search query [13, 15, 22]. A recent study has also examined the possibility of using the *word2-vec* language model [23–25] to automatically identify relevant words in the search results that can be used to extend the query [26].

Approaches for automatically adapting and extending queries still require the patent examiner to manually formulate the initial search query. To make this step obsolete, heuristics can be used to automatically extract keywords from a given patent application [27–29] or a *bag-of-words* (BOW) approach can be used to transform the entire text of a patent into a list of words that can then be used to search for its prior art [30–32]. Often times, partial patent applications, such as an extended abstract, may already suffice to conduct the search [31]. The search results can also be further refined with a graph-based ranking model [33] or by using the patents' categories to filter the results [34]. Different prior art search approaches have previously been discussed and benchmarked within the CLEF-IP project, see e.g. [35] and [36].

In our approach, detailed in the following sections, we also alleviate the required work and time needed to manually compose a search query by simply operating on the patent application's *entire* text. However, instead of only searching the database for relevant keywords extracted from this text, we transform the texts of all other documents into numerical feature representations as well, which allow us to compute the full text similarities between the patent application and its possible prior art.

Calculating the similarity between texts is at the heart of a wide range of information retrieval tasks, such as search engine development, question answering, document clustering, or corpus visualization. Approaches for computing text similarities can be divided into similarity measures relying on word similarities and those based on document feature vectors [37].

To compute the similarity between two texts using individual word similarities, the words in both texts first have to be aligned by creating word pairs based on semantic similarity and then these similarity scores are combined to yield a similarity measure for the whole text. Corley and Mihalcea [38] propose a text similarity measure, where the most similar word pairs in two texts are determined based on semantic word similarity measures as implemented in the WordNet similarity package [39]. The similarity score of two texts is then computed as the weighted and normalized sum of the single word pairs' similarity scores. This approach can be further refined using greedy pairing [40]. Recently, instead of using WordNet relations to obtain word similarities, the similarity between semantically meaningful word embeddings, such as those created by the *word2vec* language model [23], was used. Kusner et al. [41] defined the word mover's distance for computing the similarity between two sentences as the minimum distance the individual word embeddings have to move to match those of the other sentence. While similarity measures based on the semantic similarities of individual words are advantageous when comparing short texts, finding an optimal word pairing for longer texts is computationally very expensive and therefore these similarity measures are less practical in our setting, where the full texts of whole documents have to be compared.

To compute the similarity between longer documents, these can be transformed into numerical feature vectors, which serve as input to a similarity function. Rieck and Laskov [42] give a comprehensive overview of similarity measures for sequential data, some of which are widely used in information retrieval applications. Achananuparp et al. [43] test some of these similarity measures for comparing sentences on three corpora, using accuracy, precision, recall, and rejection as metrics to evaluate how many of the retrieved documents are relevant

in relation to the number of relevant documents missed. Huang [44] use several of these similarity measures to perform text clustering on *tf-idf* vectors. Interested in how well similarity measures reproduce human similarity ratings, Lee et al. [45] create a text similarity corpus based on all possible pairs of 50 different documents rated by 83 students. They test different feature extraction methods in combination with four of the similarity measures described in Rieck and Laskov [42] and calculate the correlation of the human ratings with the resulting scoring. They conclude that using the cosine similarity, high precision can be achieved, while recall is still not satisfying.

Full text similarity measures have previously been used to improve search results for MEDLINE articles, where a two step approach using the cosine similarity measure between *tf-idf* vectors in combination with a sentence alignment algorithm yielded superior results compared to the boolean search strategy used by PubMed [46]. The Science Concierge [47] computes the similarities between papers' abstracts to provide content based recommendations, however it still requires an initial keyword search to retrieve articles of interest. The PubVis web application by Horn [48], developed for visually exploring scientific corpora, also provides recommendations for similar articles given a submitted abstract by measuring overlapping terms in the document feature vectors. While full text similarity search approaches have shown potential in domains such as scientific literature, only few studies have explored this approach for the much harder task of retrieving prior art for a new patent application [49], where much less overlap between text documents is to be expected due to the usage of very abstract and general terms when describing new inventions. Specifically, document representations created using recently developed neural network language models such as *word2vec* [23, 24, 50] or *doc2vec* [51] were not yet evaluated on patent documents.

Methods

In order to study our hypothesis that the search for prior art can be improved by automatically determining, for a given patent application, the most similar documents contained in the database based on their full texts, we need to evaluate multiple approaches for comparing the patents' full texts and computing similarities between the documents. To do this, we test multiple approaches for creating numerical feature representations from the documents' raw texts, which can then be used as input to a similarity function to compute the documents' similarity.

All raw documents first have to be preprocessed by lower casing and removing non-alphanumeric characters. The simplest way of transforming texts into numerical vectors is to create high dimensional but sparse *bag-of-words* (BOW) vectors with *tf-idf* features [52]. These BOW representations can also be reduced to their most expressive dimensions using dimensionality reduction methods such as *latent semantic analysis* (LSA) [49, 53] or *kernel principal component analysis* (KPCA) [54–57]. Alternatively, the neural network language models (NNLM) [58]*word2vec* [23, 24] (combined with BOW vectors) or *doc2vec* [51] can be used to transform the documents into feature vectors. All these feature representations are described in detail in Section *A1* in S1 File.

Using any of these feature representations, the pairwise similarity between two documents' feature vectors $\mathbf{x}_i$ and $\mathbf{x}_j$ can be calculated using the cosine similarity:

$$\mathrm{sim}(\mathbf{x}_i, \mathbf{x}_j) = \frac{\mathbf{x}_i^\top \mathbf{x}_j}{\|\mathbf{x}_i\| \|\mathbf{x}_j\|},$$

which is 1 for documents that are (almost) identical, and 0 (in the case of non-negative BOW feature vectors) or below 0 for unrelated documents [44, 59, 60]. Other possible similarity functions for comparing sequential data [42, 61] are discussed in Section *A2* in S1 File.

Data

Our experiments are conducted on two datasets, created using a multi-step process as briefly outlined here and further discussed in Section *B* in S1 File. For ease of notation, we use the term patent when really referring to either a granted patent or a patent application.

We first obtained a patent corpus containing more than 100,000 patent documents from the *Cooperative Patent Classification scheme* (CPC) category A61 (*medical or veterinary science and hygiene*), published between 2000 and 2015. From these documents, our first dataset was compiled, starting with the roughly 2,500 patents in the corpus published in 2015, which we will refer to as "target patents" in the remaining text. Each of the target patents cites on average 17.5 (standard deviation: ± 28.4) other patents in our corpus (i.e. published after 2000), which we also include in the dataset. Additionally, we randomly selected another 1,000 patents from the corpus, which were not cited by any of the selected target patents. This results in altogether 28,381 documents, which contain on average 13,530 (± 18,750) words. From these documents, the first dataset was then created by pairing up the patents and assigning each patent pair a corresponding label: Each target patent is paired up with a) all the patents it cites, these patent pairs are assigned the label 'cited', and b) the 1,000 patents not cited by any of the target patents, these patent pairs are labelled 'random'. This first dataset consists of 2,470,736 patent pairs with a 'cited/random' labelling.

The second dataset is created by obtaining additional, more consistent human labels from a patent attorney for a small subset of the first dataset. These labels should show which of the cited patents are truly relevant to the target patent and whether important prior art is missing from the search reports. For ten of the target patents, we selected their respective cited patents as well as several random patents that either obtained a relatively high, medium, or low similarity score as computed with the cosine similarity on *tf-idf* BOW features. These 450 patent pairs were then manually assigned 'relevant/irrelevant' labels and constitute our second dataset.

Evaluation

A pair of patents should have a high similarity score if the two texts address a similar or almost identical subject matter, and a low score if they are unrelated. Furthermore, if two patent documents address a similar subject matter, then one document of said pair should have been cited in the search report of the other. To evaluate the similarity computation with different feature representations, the task of finding similar patents can be modelled as a binary classification problem, where the samples correspond to pairs of patents. A patent pair is given a positive label, if one of the patents was cited by the other, and a negative label otherwise. We can then compute similarity scores for all pairs of patents and select a threshold for the score where we say all patent pairs with a similarity score higher than this threshold are relevant for each other while similarity scores below the threshold indicate the patents in this pair are unrelated. With a meaningful similarity measure, it should be possible to choose a threshold such that most patent pairs associated with a positive label have a similarity score above the threshold and the pairs with negative labels score below the threshold, i.e., the two similarity score distributions should be well separated. For a given threshold, we can compute the *true positive rate* (TPR), also called *recall*, and the *false positive rate* (FPR) of the similarity measure. By plotting the TPR against the FPR for different decision thresholds, we obtain the graph of the *receiver operating characteristic* (ROC) curve, where the *area under the* ROC *curve* (AUC) conveniently translates the performance of the similarity measure into a number between 0.5 (similarity scores assigned to patent pairs with a 'cited' relationship and randomly paired patents are in the same range) and 1 (semantically related patents receive consistently higher similarity

scores than unrelated patent pairs). Further details on this performance measure can be found in Section *C* in S1 File.

While the AUC is a very useful measure to select a similarity function based on which relevant and irrelevant patents can be reliably separated, the exact score also depends on characteristics of the dataset and may therefore seem overly optimistic [62]. Especially in our first dataset, many of the randomly selected patents contain little overlap with the target patents and can therefore be easily identified as irrelevant. With only a small fraction of the random pairs receiving a medium or high similarity score, this means that for most threshold values the FPR will be very low, resulting in larger AUC values. To give a further perspective on the performance of the compared similarity measures, we therefore additionally report the *average precision* (AP) score for the final results. For a specific threshold, precision is defined as the number of TP relative to the number of all returned documents, i.e., TP+FP. As we rank the patent pairs based on their similarity score, precision and recall can again be plotted against each other for n different thresholds and the area under this curve can be computed as the weighted average of precision (P) and recall (R) for all n threshold values [63]:

$$AP = \sum_n (R_n - R_{n-1}) P_n.$$

Results

The aim of our study is to identify a robust approach for computing the full text similarity between two patents. To this end, in the following we evaluate different document feature representations and similarity functions by assessing how well the computed similarity scores are aligned with the labels of our two datasets, i.e., whether a high similarity score is assigned to pairs that are labelled as *cited* (*relevant*) and low similarity scores to *random* (*irrelevant*) pairs. Furthermore, we examine the discrepancies between patents cited in a patent application's search report and truly relevant prior art.

Using full text similarity to identify cited patents

The similarities between the patents in each pair contained in the cited/random dataset are computed using the different feature extraction methods together with the cosine similarity and the obtained similarity scores are then evaluated by computing the AUC with respect to the pairs' labels (Table 1). The similarity scores are computed using either the full texts of the patents to create the feature vectors, or only parts of the documents, such as the patents' abstracts or their claims, to identify which sections are most relevant for this task [31, 64].

Table 1. Evaluation results on the cited/random dataset.

Features	patent section: AUC		
	full text	***abstract***	***claims***
Bag-of-words	**0.9560**	0.8620	0.8656
LSA	0.9361	0.8579	0.8561
KPCA	0.9207	0.8377	0.8250
BOW + word2vec	0.9410	0.8618	0.8525
doc2vec	0.9314	**0.8919**	**0.8898**

AUC values when computing the cosine similarity with BOW, LSA, KPCA, *word2vec*, and *doc2vec* features constructed from different patent sections of the cited/random dataset.

https://doi.org/10.1371/journal.pone.0212103.t001

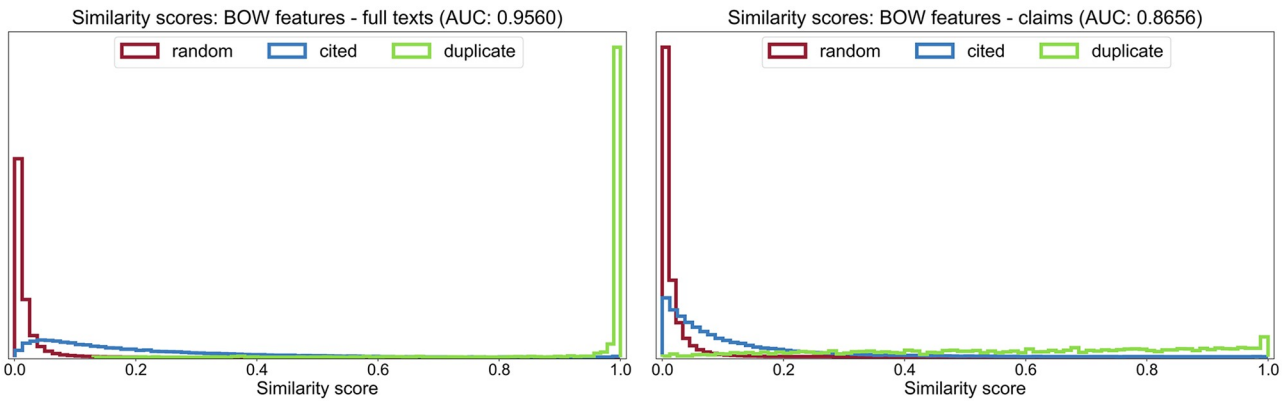

Fig 2. Distributions of cosine similarity scores. Similarity scores for the patent pairs are computed using BOW feature vectors generated either from full texts *(left)* or only the claims sections *(right)*. Scale on the *y*-axis is irrelevant and was therefore omitted.

https://doi.org/10.1371/journal.pone.0212103.g002

Additionally, the results on this dataset using BOW feature vectors together with other similarity measures can be found in Section *D1* in S1 File.

The BOW features outperform the tested dimensionality reduction methods LSA and KPCA as well as the NNLM *word2vec* and *doc2vec* when comparing the patents' full texts (Table 1). Yet, with AUC values greater than 0.9, all methods succeed in identifying cited patents by assigning the patents found in a target patent's search report a higher similarity score than those that they were paired up with randomly. When only certain patent sections are taken into account, the NNLMs perform as good (*word2vec*) or even better (*doc2vec*) than the BOW vectors, and LSA performs well on the claims section as well. The comparably good performance, especially of *doc2vec*, on individual sections is probably due to the fact that these feature representations are more meaningful when computed for shorter texts, whereas when combining the embedding vectors of too many individual words, the resulting document representation can be rather noisy.

When looking more closely at the score distributions obtained with BOW features on the patents' full texts as well as their claims sections (Fig 2), it can be seen that when only using the claims sections, the scores of the duplicate patent pairs, instead of being clustered near 1, range nearly uniformly between 0 and 1. This can be explained by divisional applications and the fact that during the different stages of a submission process, most of the time only the claims section is revised (usually by weakening the claims), such that several versions of a patent application will essentially differ from each other only in their claims whereas abstract and description remain largely unchanged [31, 32].

Identifying truly relevant patents

The search for prior art for a given patent application is in general conducted by a single person using mainly keyword searches, which might result in false positives as well as false negatives. Furthermore, as different patent applications are handled by different patent examiners, it is difficult to obtain a consistently labelled dataset. A more reliably labelled dataset would therefore be desirable to properly evaluate our automatic search approach. In the previous section, we showed that by computing the cosine similarity between feature vectors created from full patent texts we can identify patents that occur in the search report of a target patent. However, the question remains, whether these results translate to a real setting and if it is possible to find patents previously overlooked or prevent the citation of actually irrelevant patents.

Table 2. Confusion matrix for the dataset subsample.

	cited	random
relevant	65	18
irrelevant	86	281

The original cited/random labelling is compared to the more accurate relevant/irrelevant labels.

https://doi.org/10.1371/journal.pone.0212103.t002

To get an estimate of how many of the cited, as well as the patents identified through our automated approach, are truly relevant for a given target patent, we asked a patent attorney to label a small subsample of the first dataset. As the patent attorney labelled these patents very carefully, her decisions merit a high confidence and we therefore consider them as the ground truth when her ratings are in conflict with the citation labels.

Using this second, more reliably labelled dataset, we first assess the amount of (dis)agreement between the cited/random labelling, based on the search reports, and the relevant/irrelevant labelling, obtained from the patent attorney. We then evaluate the similarity scores computed for this second dataset to see whether our automated approach is indeed capable of identifying the truly relevant prior art for a new patent application.

Comparing the current citation process to the additional human labels. To see if documents found in the search for prior art conducted by the patent office generally coincide with the documents considered relevant by our patent attorney, the confusion matrix as well as the correlation between the two human labellings is analysed. Please keep in mind that, in general, patent examiners can only assess the relevance of prior art that was actually found by the keyword driven search.

Taking the relevant/irrelevant labelling as the ground truth, the confusion matrix (Table 2) shows that 86 FP and 18 FN are produced by the patent examiner, which results in a recall of 0.78 and a precision score of 0.43. The large number of false positives can, in part, be explained by applicants being required by the USPTO to file so-called Information Disclosure Statements (IDS) including, according to the applicant, related background art [65]. The documents cited in an IDS are then included in the list of citations by the examiner, thus resulting in very long citations lists.

To get a better understanding of the relationship between the cosine similarity computed using BOW feature vectors and the relevant/irrelevant as well as the cited/random labelling, we calculate their pairwise correlations using Spearman's ρ (Table 3). The highest correlation score of 0.652 is reached between the relevant/irrelevant labelling and the cosine similarity, whereas Spearman's ρ for the cosine similarity and the cited/random labels is much lower (0.501).

When plotting the cosine similarity and the relevant/irrelevant labelling against each other for individual patents (e.g. Fig 3), in most cases, the scorings agree on whether a patent is relevant or not for the target patent. Yet it is worthwhile to inspect some of the outliers to get a

Table 3. Correlations between labels and similarity scores on the dataset subsample.

	cited/random	relevant/irr.
cosine (BOW)	0.501	0.652
relevant/irr.	0.592	—

Spearman's ρ for the cosine similarity calculated with BOW feature vectors and the relevant/irrelevant and cited/random labelling.

https://doi.org/10.1371/journal.pone.0212103.t003

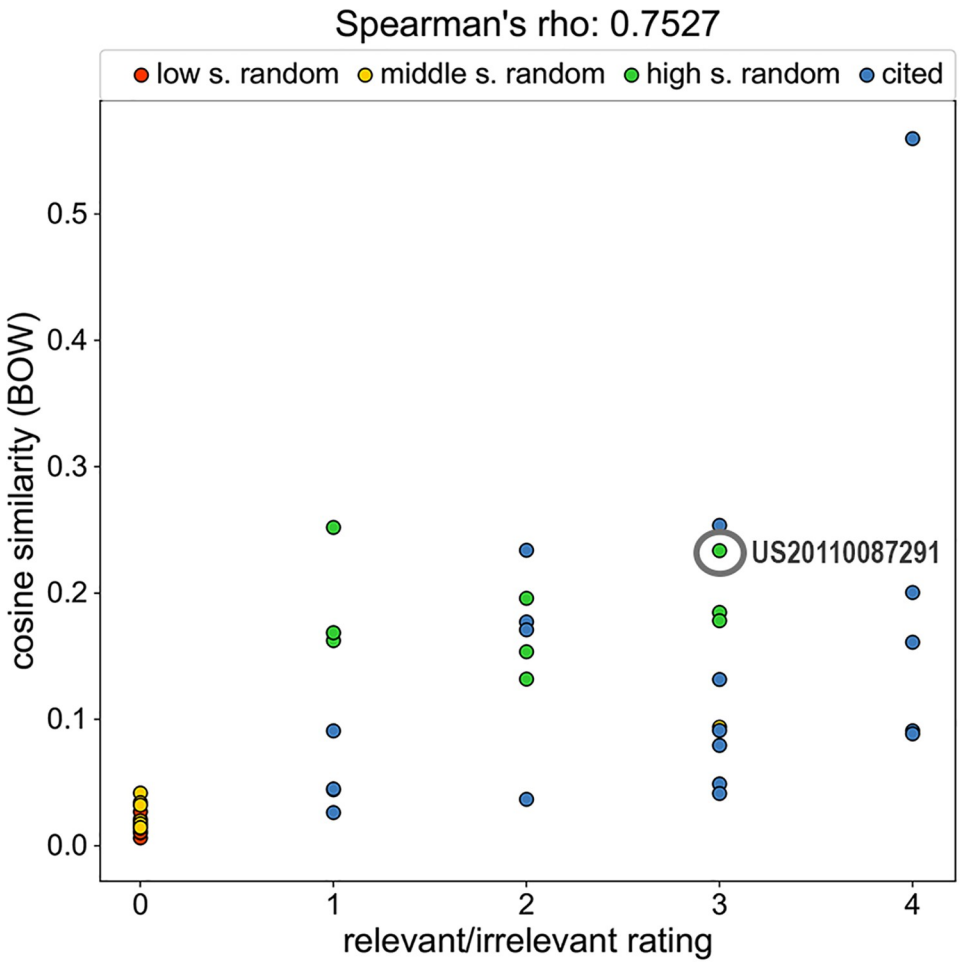

Fig 3. Score correlation for the patent with ID US20150018885. A false negative (ID US20110087291) caught by the cosine similarity is circled in gray.

https://doi.org/10.1371/journal.pone.0212103.g003

better understanding of the process. In Section *D2* in S1 File we discuss two false positives, one produced by our approach and one found in a patent's search report. More problematic, however, are false negatives, i.e., prior art that was missed when filing the application. For the target patent with ID US20150018885 our automated approach would have discovered a relevant patent, which was missed by the search performed by the patent examiner (Fig 3). The patent with ID US20110087291 must be considered as relevant for the target patent, because both describe rigid bars that are aimed at connecting vertebrae for stabilization purposes with two anchors that are screwed into the bones. While in the target patent, the term *bone anchoring member* is used, the same part of the device in patent US20110087291 is called *connecting member*, which is a more abstract term. Moreover, instead of talking about a *connecting bar*, as it is done in the target patent, the term *elongate fusion member* is used in the other patent application.

Using full text similarity to identify relevant patents. In order to systematically assess how close the similarity score ranking can get to the one of the patent attorney (relevant/irrelevant) compared to the one of the patent office examiners (cited/random), the experiments performed on the first dataset with respect to the cited/random labelling were again conducted on this dataset subsample. For the analysis, it is important to bear in mind that this dataset is

Table 4. Summary of evaluation results.

Features	AUC			AP		
	subsample		*full*	*subsample*		*full*
	relevant	cited	cited	relevant	cited	cited
Bag-of-words	0.8118	0.8063	**0.9560**	0.5274	0.7095	**0.4705**
LSA	0.7798	0.7075	0.9361	0.4787	0.5921	0.3257
KPCA	0.7441	0.6740	0.9207	0.4721	0.5832	0.2996
BOW + word2vec	**0.8408**	**0.8544**	0.9410	**0.5443**	**0.7354**	0.4019
doc2vec	0.7658	0.8138	0.9314	0.4749	0.6829	0.3121

AUC and average precision (AP) scores for the different feature extraction methods on the dataset subsample with cited/random and relevant/irrelevant labelling, as well as the full dataset.

https://doi.org/10.1371/journal.pone.0212103.t004

different from the one used in the previous experiments, as it only consists of the 450 patent pairs scored by the patent attorney. For each of the feature extraction methods, it was assessed how well the cosine similarity could distinguish between the relevant and irrelevant as well as the cited and random patent pairs of this smaller dataset.

The AUC and AP values achieved with the different feature representations on both labellings as well as, for comparison, on the original dataset, are reported in Table 4. On this dataset subsample, the AUC w.r.t. the cited/random labelling is much lower than in the previous experiment on the larger dataset (0.806 compared to 0.956 for BOW features), which can be in part explained by the varying number of easily identifiable negative samples and their impact on the FPR: The full cited/random dataset contains many more low-scored random patents than the relevant/irrelevant subsample, where we included an equal amount of low- and high-scored random patents for each of the ten target patents. Yet, for most feature representations, the performance is better for the relevant/irrelevant than for the cited/random labelling of the dataset subsample, and the best results on the relevant/irrelevant labelling are achieved using the combination of BOW vectors and *word2vec* embeddings as feature vectors.

Discussion

The search for prior art for a given patent application is currently based on a manually conducted keyword search, which is not only time consuming but also prone to mistakes yielding both false positives and, more problematically, false negatives. In this paper, an approach for automating the search for prior art was developed, where a patent application's *full* text is automatically compared to the patents contained in a database, yielding a similarity score based on which the patents can be ranked from most similar to least similar. The patents whose similarity scores exceed a certain threshold can then be suggested as prior art.

Several feature extraction methods for transforming documents into numerical vectors were evaluated on a dataset consisting of several thousand patent documents. In a first step, the evaluation was performed with respect to the distinction between cited and random patents, where cited patents are those included in the given target patent's search report and random patents are randomly selected patent documents that were not cited by any of the target patents. We showed that by computing the cosine similarity between feature vectors created from full patent texts, we can reliably identify patents that occur in the search report of a target patent. The best distinction between these cited and random patents on the full corpus could be achieved when computing the cosine similarity using the well-established *tf-idf* BOW features, which is conceptually the method most closely related to a regular keyword search.

To examine the discrepancies between the computed similarity scores and cited/random labels, we obtained additional and more reliable labels from a patent attorney to identify truly relevant patents. As illustrated by Tables 3 and 4, the automatically calculated similarities between patents are closer to the patent attorney's relevancy scoring than to the cited/random labellings obtained from the search report. The comparison of different feature representations on the smaller dataset not only showed that the same feature extraction method reaches different AUCs for the two labellings, but also that the feature extraction method that best distinguishes between cited and random patents on the full corpus (BOW) was outperformed on the relevant/irrelevant dataset by the combination of *tf-idf* BOW feature vectors with *word2vec* embeddings. This again indicates that the keyword search is missing patents that use synonyms or more general and abstract terms, which can be identified using the semantically meaningful representations learned by a NNLM. Therefore, with our automated similarity search, we are able to identify the truly relevant documents for a given patent application.

Most importantly, we gave an example where the cosine similarity caught a relevant patent originally missed by the patent examiner (Fig 3). As discussed at the beginning of this paper, missing a relevant prior art document in the search is a serious issue, as this might lead to an erroneous grant of a patent with profound legal and financial implications for both the applicant as well as competitors.

Consequently, our findings show that the search for prior art for a given patent application, and thereby the citation process, can be greatly enhanced by a precursory similarity scoring of the patents based on their full texts. With our NLP based approach we would not only greatly accelerate the search process, but, as shown in our empirical analysis, our method could also improve the *quality* of the results by reducing the number of omitted yet relevant documents.

Given the so far unsatisfying precision (0.43) and recall (0.78) values of the standard citation process compared to the relevancy labellings provided by our patent attorney, in the future it is clearly desirable to focus on improving the separation of relevant and irrelevant instead of cited and random patents. Our results on the small relevant/irrelevant dataset, while very encouraging, should only be considered as a first indicative step; clearly the creation of a larger dataset, reliably labelled by several experts, will be an essential next step for any further evaluation.

While we have demonstrated that our search approach is capable of identifying FP and FN w.r.t. the documents cited in a patent's original search report, it is not clear whether this original search for prior art was always conducted using any of the more sophisticated IR approaches discussed in the related works section at the beginning of the paper, i.e., going beyond a basic manual keyword search. Therefore, a future step in the evaluation of our search approach would be to benchmark our methods against these existing IR techniques specifically developed for the prior art search, for example, using the CLEF-IP datasets [35, 36].

Furthermore, the methods discussed within this paper should also be applied to documents from other CPC classes to assess the quality of the automatically generated search results in domains other than medical or veterinary science and hygiene. Additionally considering the (sub)categories of the patents as features when conducting the search for prior art also seems like a promising step to further enhance the search results [34, 66].

It should also be evaluated how well these results translate to patents filed in other countries [67, 68], especially if these patents were automatically translated using machine translation methods [69, 70]. Here it may also be important to take a closer look at similarity search results obtained by using only the texts from single patent sections. As related work has shown [31, 64], an extended abstract and description may often suffice to find prior art. This can speed up the patent filing process, as all relevant prior art can already be identified early in the patent application process, thereby reducing the number of duplicate submissions with only revised

(i.e. weakened) claims. However, as patents filed in different countries have different structures, these results might not directly translate to, e.g., patents filed with the European Patent Office.

It might also be of interest to compare other NNLM based feature representations for this task, e.g., by combining the *word2vec* embeddings with a convolutional neural network [71, 72]. To better adapt a similarity search approach to patents from other domains, it could also be advantageous to additionally take into account image based similarities computed from the sketches supplied in the patent documents [2, 10].

An important challenge to solve furthermore is how an exhaustive comparison of a given patent application to all the millions of documents contained in a real world patent database could be performed efficiently. Promising approaches for speeding up the similarity search for all pairs in a set [73] should be explored for this task in future work.

The search for a patent's prior art is a particularly difficult problem, as patent applications are purposefully written in a way that is to create little overlap with other patents, as only by distinguishing the invention from others, a patent application has a chance of being granted [6]. By showing that our automated full text similarity search approach successfully improves the search for a patent's prior art, consequently these methods are also promising candidates for enhancing other document searches, such as identifying relevant scientific literature.

Supporting information

S1 File. Pdf file with supporting information. *A* Methods: *A1* Feature representations of text documents, *A2* Functions for measuring similarity between text documents. *B* Data. *C* Evaluation. *D* Results: *D1* Identifying cited patents using different similarity functions with BOW features, *D2* Detailed examination of outliers in the citation process.
(PDF)

Author Contributions

Conceptualization: Franziska Biegler, Klaus-Robert Müller.

Data curation: Franziska Biegler, Tim Oppermann.

Formal analysis: Lea Helmers, Franziska Horn.

Funding acquisition: Klaus-Robert Müller.

Investigation: Franziska Horn.

Methodology: Lea Helmers, Franziska Horn.

Project administration: Klaus-Robert Müller.

Resources: Franziska Biegler, Klaus-Robert Müller.

Software: Lea Helmers, Franziska Horn.

Supervision: Klaus-Robert Müller.

Validation: Klaus-Robert Müller.

Visualization: Franziska Horn.

Writing – original draft: Lea Helmers, Franziska Horn.

Writing – review & editing: Lea Helmers, Franziska Horn, Franziska Biegler, Tim Oppermann, Klaus-Robert Müller.

References

1. Publication W. WIPO Intellectual Property Handbook Second Edition. vol. No. 489 (E). WIPO; 2004.
2. Alberts D, Yang CB, Fobare-DePonio D, Koubek K, Robins S, Rodgers M, et al. Introduction to Patent Searching. In: Lupu M, Mayer K, Kando N, Trippe AJ, editors. Current Challenges in Patent Information Retrieval. Berlin, Heidelberg: Springer Berlin Heidelberg; 2017. p. 3–45.
3. Golestan Far M, Sanner S, Bouadjenek MR, Ferraro G, Hawking D. On term selection techniques for patent prior art search. In: Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval. ACM; 2015. p. 803–806.
4. Tseng YH, Lin CJ, Lin YI. Text Mining Techniques for Patent Analysis. Inf Process Manage. 2007; 43(5):1216–1247. https://doi.org/10.1016/j.ipm.2006.11.011
5. Tannebaum W, Rauber A. PatNet: a lexical database for the patent domain. In: European Conference on Information Retrieval. Springer; 2015. p. 550–555.
6. Andersson L, Hanbury A, Rauber A. The Portability of Three Types of Text Mining Techniques into the Patent Text Genre. In: Lupu M, Mayer K, Kando N, Trippe AJ, editors. Current Challenges in Patent Information Retrieval. Berlin, Heidelberg: Springer Berlin Heidelberg; 2017. p. 241–280.
7. Kando N, Leong MK. Workshop on patent retrieval (SIGIR 2000 workshop report). In: SIGIR Forum. vol. 34; 2000. p. 28–30.
8. Alberts D, Yang CB, Fobare-DePonio D, Koubek K, Robins S, Rodgers M, et al. Introduction to patent searching. In: Current challenges in patent information retrieval. Springer; 2011. p. 3–43.
9. Lupu M, Piroi F, Stefanov V. An Introduction to Contemporary Search Technology. In: Lupu M, Mayer K, Kando N, Trippe AJ, editors. Current Challenges in Patent Information Retrieval. Berlin, Heidelberg: Springer Berlin Heidelberg; 2017. p. 47–73.
10. Lupu M, Hanbury A. Patent retrieval. Foundations and Trends® in Information Retrieval. 2013; 7(1):1–97. https://doi.org/10.1561/1500000027
11. Shalaby W, Zadrozny W. Patent Retrieval: A Literature Review. arXiv preprint arXiv:170100324. 2017;.
12. Magdy W, Jones GJ. A study on query expansion methods for patent retrieval. In: Proceedings of the 4th workshop on Patent information retrieval. ACM; 2011. p. 19–24.
13. Fujii A. Enhancing patent retrieval by citation analysis. In: Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval. ACM; 2007. p. 793–794.
14. Mahdabi P, Crestani F. Learning-based pseudo-relevance feedback for patent retrieval. In: Information Retrieval Facility Conference. Springer; 2012. p. 1–11.
15. Mahdabi P, Crestani F. The effect of citation analysis on query expansion for patent retrieval. Information retrieval. 2014; 17(5-6):412–429. https://doi.org/10.1007/s10791-013-9232-5
16. Trippe A, Ruthven I. Evaluating Real Patent Retrieval Effectiveness. In: Lupu M, Mayer K, Kando N, Trippe AJ, editors. Current Challenges in Patent Information Retrieval. Berlin, Heidelberg: Springer Berlin Heidelberg; 2017. p. 143–162.
17. Magdy W, Leveling J, Jones GJ. Exploring structured documents and query formulation techniques for patent retrieval. In: Workshop of the Cross-Language Evaluation Forum for European Languages. Springer; 2009. p. 410–417.
18. Wang F, Lin L, Yang S, Zhu X. A semantic query expansion-based patent retrieval approach. In: Fuzzy Systems and Knowledge Discovery (FSKD), 2013 10th International Conference on. IEEE; 2013. p. 572–577.
19. Zhang L. PatSearch: An Integrated Framework for Patent Document Retrieval. In: An Integrated Framework for Patent Analysis and Mining. FIU Electronic Theses and Dissertations; 2016.
20. Ganguly D, Leveling J, Magdy W, Jones GJ. Patent query reduction using pseudo relevance feedback. In: Proceedings of the 20th ACM international conference on Information and knowledge management. ACM; 2011. p. 1953–1956.
21. Tannebaum W, Rauber A. Using query logs of USPTO patent examiners for automatic query expansion in patent searching. Information retrieval. 2014; 17(5-6):452–470. https://doi.org/10.1007/s10791-014-9238-7
22. Mahdabi P, Crestani F. Query-driven mining of citation networks for patent citation retrieval and recommendation. In: Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management. ACM; 2014. p. 1659–1668.
23. Mikolov T, Chen K, Corrado G, Dean J. Efficient estimation of word representations in vector space. arXiv preprint arXiv: 13013781. 2013;.

24. Mikolov T, Sutskever I, Chen K, Corrado G, Dean J. Distributed Representations of Words and Phrases and their Compositionality. Advances in Neural Information Processing Systems. 2013; p. 3111–3119.
25. Mikolov T, Yih Wt, Zweig G. Linguistic Regularities in Continuous Space Word Representations. In: Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies; 2013. p. 746–751.
26. Singh J, Sharan A. Relevance Feedback-based Query Expansion Model using Ranks Combining and Word2Vec Approach. IETE Journal of Research. 2016; 62(5):591–604. https://doi.org/10.1080/03772063.2015.1136575
27. Mahdabi P, Keikha M, Gerani S, Landoni M, Crestani F. Building queries for prior-art search. In: Information Retrieval Facility Conference. Springer; 2011. p. 3–15.
28. Konishi K. Query Terms Extraction from Patent Document for Invalidity Search. In: Proceedings of NTCIR-5 Workshop Meeting, 2005-12; 2005.
29. Verma M, Varma V. Applying key phrase extraction to aid invalidity search. In: Proceedings of the 13th International Conference on Artificial Intelligence and Law. ACM; 2011. p. 249–255.
30. Verberne S, D'hondt E. Prior art retrieval using the claims section as a bag of words. In: Workshop of the Cross-Language Evaluation Forum for European Languages. Springer; 2009. p. 497–501.
31. Bouadjenek MR, Sanner S, Ferraro G. A study of query reformulation for patent prior art search with partial patent applications. In: Proceedings of the 15th International Conference on Artificial Intelligence and Law. ACM; 2015. p. 23–32.
32. Xue X, Croft WB. Transforming patents into prior-art queries. In: Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval. ACM; 2009. p. 808–809.
33. Mihalcea R, Tarau P. Textrank: Bringing order into text. In: Proceedings of the 2004 conference on empirical methods in natural language processing; 2004.
34. Verma M, Varma V. Exploring Keyphrase Extraction and IPC Classification Vectors for Prior Art Search. In: CLEF (Notebook Papers/Labs/Workshop); 2011.
35. Piroi F. CLEF-IP 2010: Prior art candidates search evaluation summary. Technical Report IRF Report 2010-00003, Information Retrieval Facility, Vienna; 2010.
36. Piroi F, Lupu M, Hanbury A. Overview of clef-ip 2013 lab. In: International Conference of the Cross-Language Evaluation Forum for European Languages. Springer; 2013. p. 232–249.
37. Gomaa WH, Fahmy AA. A survey of text similarity approaches. International Journal of Computer Applications. 2013; 68(13).
38. Corley C, Mihalcea R. Measuring the Semantic Similarity of Texts. In: Proceedings of the ACL Workshop on Empirical Modeling of Semantic Equivalence and Entailment. EMSEE'05. Stroudsburg, PA, USA: Association for Computational Linguistics; 2005. p. 13–18.
39. Patwardhan S, Banerjee S, Pedersen T. Using Measures of Semantic Relatedness for Word Sense Disambiguation. In: Proceedings of the 4th International Conference on Computational Linguistics and Intelligent Text Processing. CICLing'03. Berlin, Heidelberg: Springer-Verlag; 2003. p. 241–257.
40. Lintean MC, Rus V. Measuring Semantic Similarity in Short Texts through Greedy Pairing and Word Semantics. In: FLAIRS Conference; 2012. p. 244–249.
41. Kusner M, Sun Y, Kolkin N, Weinberger K. From word embeddings to document distances. In: International Conference on Machine Learning; 2015. p. 957–966.
42. Rieck K, Laskov P. Linear-Time Computation of Similarity Measures for Sequential Data. J Mach Learn Res. 2008; 9:23–48.
43. Achananuparp P, Hu X, Shen X. The Evaluation of Sentence Similarity Measures. In: Proceedings of the 10th International Conference on Data Warehousing and Knowledge Discovery. DaWaK'08. Berlin, Heidelberg: Springer-Verlag; 2008. p. 305–316.
44. Huang A. Similarity measures for text document clustering. In: Proceedings of the Sixth New Zealand Computer Science Research Student Conference (NZCSRSC2008); 2008. p. 49–56.
45. Lee MD, Pincombe B, Welsh M. An empirical evaluation of models of text document similarity. In: Proceedings of the Cognitive Science Society. vol. 27; 2005. p. 1254–1259.
46. Lewis J, Ossowski S, Hicks J, Errami M, Garner HR. Text similarity: an alternative way to search MEDLINE. Bioinformatics. 2006; 22(18):2298–2304. https://doi.org/10.1093/bioinformatics/btl388 PMID: 16926219
47. Achakulvisut T, Acuna DE, Ruangrong T, Kording K. Science Concierge: A fast content-based recommendation system for scientific publications. PLOS ONE. 2016; 11(7):e0158423. https://doi.org/10.1371/journal.pone.0158423 PMID: 27383424

48. Horn F. Interactive Exploration and Discovery of Scientific Publications with PubVis. arXiv preprint arXiv: 170608094. 2017;.
49. Moldovan A, Boţ RI, Wanka G. Latent semantic indexing for patent documents. International Journal of Applied Mathematics and Computer Science. 2005; 15:551–560.
50. Horn F. Context encoders as a simple but powerful extension of word2vec. In: Proceedings of the 2nd Workshop on Representation Learning for NLP. Association for Computational Linguistics; 2017. p. 10–14.
51. Le Q, Mikolov T. Distributed Representations of Sentences and Documents. In: Jebara T, Xing EP, editors. Proceedings of the 31st International Conference on Machine Learning (ICML-14). JMLR Workshop and Conference Proceedings; 2014. p. 1188–1196.
52. Manning CD, Raghavan P, Schütze H. Scoring, term weighting and the vector space model. In: Introduction to Information Retrieval. Cambridge University Press; 2008.
53. Landauer TK, Foltz PW, Laham D. An introduction to latent semantic analysis. Discourse processes. 1998; 25:259–284. https://doi.org/10.1080/01638539809545028
54. Schölkopf B, Smola A, Müller KR. Nonlinear Component Analysis as a Kernel Eigenvalue Problem. Neural Computation. 1998; 10(5):1299–1319. https://doi.org/10.1162/089976698300017467
55. Müller KR, Mika S, Rätsch G, Tsuda K, Schölkopf B. An introduction to kernel-based learning algorithms. IEEE Transactions on Neural Networks. 2001; 12(2):181–201. https://doi.org/10.1109/72.914517 PMID: 18244377
56. Schölkopf B, Smola AJ. Learning with kernels: support vector machines, regularization, optimization, and beyond. MIT press; 2002.
57. Schölkopf B, Smola AJ. A Short Introduction to Learning with Kernels. In: Mendelson S, Smola AJ, editors. Advanced Lectures on Machine Learning: Machine Learning Summer School 2002 Canberra, Australia, February 11–22, 2002 Revised Lectures. Berlin, Heidelberg: Springer Berlin Heidelberg; 2003. p. 41–64.
58. Bengio Y, Ducharme R, Vincent P, Janvin C. A Neural Probabilistic Language Model. J Mach Learn Res. 2003; 3:1137–1155.
59. Crocetti G. Textual Spatial Cosine Similarity. CoRR. 2015; abs/1505.03934.
60. Baeza-Yates RA, Ribeiro-Neto B. Modeling. In: Modern Information Retrieval. Boston, MA, USA: Addison-Wesley Longman Publishing Co., Inc.; 1999.
61. Pele O. Distance Functions: Theory, Algorithms and Applications. The Hebrew University of Jerusalem; 2011.
62. Saito T, Rehmsmeier M. The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. PloS one. 2015; 10(3):e0118432. https://doi.org/10.1371/journal.pone.0118432 PMID: 25738806
63. Zhu M. Recall, precision and average precision. Department of Statistics and Actuarial Science, University of Waterloo, Waterloo. 2004; 2:30.
64. D'hondt E, Verberne S. CLEF-IP 2010: Prior Art Retrieval using the different sections in patent documents. In: CLEF-IP 2010. Proceedings of the Conference on Multilingual and Multimodal Information Access Evaluation (CLEF 2010), CLEF-IP workshop. Padua, Italy: [sn]; 2010.
65. Information Disclosure Statements. In: Manual of Patent Examining Procedure of the United States Patent and Trademark Office 9th Edition. United States Patent and Trademark Office; 2018.
66. Magali M, Ferraro G, Shlomo G. Four patent classification problems in information management: A review of the literature and a determination of the four essential questions for future research. Information Research. 2016; 21(1):paper 705.
67. Piroi F, Hanbury A. Evaluating Information Retrieval Systems on European Patent Data: The CLEF-IP Campaign. In: Lupu M, Mayer K, Kando N, Trippe AJ, editors. Current Challenges in Patent Information Retrieval. Berlin, Heidelberg: Springer Berlin Heidelberg; 2017. p. 113–142.
68. Lupu M, Fujii A, Oard DW, Iwayama M, Kando N. Patent-Related Tasks at NTCIR. In: Lupu M, Mayer K, Kando N, Trippe AJ, editors. Current Challenges in Patent Information Retrieval. Berlin, Heidelberg: Springer Berlin Heidelberg; 2017. p. 77–111.
69. Tinsley J. Machine Translation and the Challenge of Patents. In: Lupu M, Mayer K, Kando N, Trippe AJ, editors. Current Challenges in Patent Information Retrieval. Berlin, Heidelberg: Springer Berlin Heidelberg; 2017. p. 409–431.
70. Diallo B, Lupu M. Future Patent Search. In: Lupu M, Mayer K, Kando N, Trippe AJ, editors. Current Challenges in Patent Information Retrieval. Berlin, Heidelberg: Springer Berlin Heidelberg; 2017. p. 433–455.

71. Arras L, Horn F, Montavon G, Müller KR, Samek W. Explaining Predictions of Non-Linear Classifiers in NLP. In: Proceedings of the 1st Workshop on Representation Learning for NLP. Berlin, Germany: Association for Computational Linguistics; 2016. p. 1–7.

72. Arras L, Horn F, Montavon G, Müller KR, Samek W. "What is relevant in a text document?": An interpretable machine learning approach. PloS one. 2017; 12(8):e0181142. https://doi.org/10.1371/journal.pone.0181142 PMID: 28800619

73. Bayardo RJ, Ma Y, Srikant R. Scaling up all pairs similarity search. In: Proceedings of the 16th international conference on World Wide Web. WWW'07. New York, NY, USA: ACM; 2007. p. 131–140.