Exploring the Impact of Negative Sampling on Patent Citation Recommendation

Rima Dessi, Hidir Aras, Mehwish Alam

Exploring the Impact of Negative Sampling on Patent Citation Recommendation

Rima Dessi
FIZ Karlsruhe - Leibniz Institute
for Information Infrastructure
rima.dessi@fiz-karlsruhe.de

Hidir Aras
FIZ Karlsruhe - Leibniz Institute
for Information Infrastructure
hidir.aras@fiz-karlsruhe.de

Mehwish Alam
Telecom Paris
Institut Polytechnique de Paris
mehwish.alam@telecom-paris.fr

Abstract

Due to the increasing number of patents being published every day, patent citation recommendations have become one of the challenging tasks. Since patent citations may lead to legal and economic consequences, patent recommendations are even more challenging as compared to scientific article citations. One of the crucial components of the patent citation algorithm is negative sampling which is also a part of many other tasks such as text classification, knowledge graph completion, etc. This paper, particularly focuses on proposing a transformer-based ranking model for patent recommendations. It further experimentally compares the performance of patent recommendations based on various state-of-the-art negative sampling approaches to measure and compare the effectiveness of these approaches to aid future developments. These experiments are performed on a newly collected dataset of US patents from Google patents.

In: R. Krestel, H. Aras, A. Hanbury, L. Andersson, F. Piroi, D. Alderucci (eds.): Proceedings of the 4th Workshop on Patent Text Mining and Semantic Technologies, Taipei, Taiwan, 27-July-2023, published at http://ceur-ws.org

1 Introduction

Negative sampling is a crucial task for several applications such as recommender systems [CLY+22, OLL+13, FLL15], text classification [JWS+21, TZAS20, TZKS19], computer vision [PAHS12], etc. In order to train a machine learning model it is essential to have an accurately labeled dataset that includes sufficient positive and negative samples for each class. However, in many applications such as recommender systems obtaining negative samples is quite a challenging task. In fact, it is easy to collect positive samples for the patent citation recommendation system by considering patents' actual citations, however, generating negative samples (i.e., potential citations that are irrelevant to the given patent) is much harder [OLL+13]. In this paper, we focus on the impact of negative sampling in the context of patent citation recommendation and its role in improving the performance of citation recommendation systems.

Patent citation recommendation [FLL15, CLY+22, OLL+13] is quite challenging due to the ever-increasing number of available patents, as well as their complex structure, and the usage of domain-specific vocabulary. Manually, finding potentially relevant citations from a massive amount of patents is time-consuming and expensive. Therefore, efficient and effective tools for automatically recommending citations for patents have become indispensable. In contrast to the paper citations, patent citations carry economic and legal significance [OLL+13]. In other words, missing prior relevant patents can have critical outcomes for patent applicants. Furthermore, the number of citations that the patent receives can determine the

business value of the patent. Therefore, identifying the right prior art patents to be cited is quite a significant task for both the patent applicant and the examiner.

Recently, several patent citation recommendation systems have been proposed [FLL15, CLY+22, OLL+13]. Most of the approaches are based on two steps, i.e., retrieval and ranking. While the *retrieval phase* aims to find the most relevant citation candidates, the *ranking phase* focuses on ranking the most relevant potential citations from the candidate list with respect to a score. The ranking function is often trained by utilizing a large amount of labeled data which includes both negative and positive samples.

Several techniques [HDD+21, YDZ+22] have been proposed to generate negative samples from a dataset that contains positive samples as well as unlabeled samples. Negative sampling aims to find the best technique to select the most representative negative instances from a given dataset. In the context of patent citation recommendation systems, the positive samples are the patents' actual citations, and each unlabeled sample could belong either to the positive class or the negative class based on the content of the given patent. The type and proportion of negative samples play an important role in the performance of such systems. In other words, it is essential for the performance of the ranking model to be trained on representative samples from each class which helps the model to distinguish between the positive and negative samples. Although several negative sampling approaches have been proposed for the recommender systems [HDD+21, YDZ+22], none of the mentioned approaches specifically have been applied to the patent domain. They seem to work well with item recommendation systems, however, it is important to note that the user-item relation differs from the patent-citation relation. In other words, each citation actually is a patent, so patents and citations can be modeled in the same way to find relevancy. However, users and items should be represented differently. For instance, to model a user there exist different types of features such as age, country, gender, purchase history, etc. Yet patents are mostly modeled based on their textual content, e.g., title, abstract, and claims.

In this paper, we explore the impact of negative sampling on the ranking of patent citation recommendations. To this end, we investigate three different sampling techniques namely, random, nearest-neighbor, and the Cooperative Patent Classification (CPC) code-based. After sampling, we train a transformer-based ranking model separately for each dataset and compare the results. Additionally, we analyze the impact of different feature combinations (e.g., abstract, claim, title) as well as the effect of varying negative sample proportions on the performance of the ranking system.

Overall, the main contributions of the paper are as follows:

- Generating training data for patent citation ranking systems using various negative sampling techniques and different proportions of negative samples.
- Demonstration of the impact of the negative samples on the performance of a transformer-based ranking model.
- We release 4 different datasets[1] which can be exploited for the patent citation recommendation task.

2 Related Work

This study aims to explore the impact of negative sampling on patent citation recommendation systems, hence this section presents prior related studies on Patent Citation Recommendation and Negative Sampling Techniques.

Patent Citation Recommendation Recent works [FLL15, CLY+22, OLL+13] employ machine learning approaches for patent citation recommendation. The proposed citation recommendation frameworks consist of 2 main phases namely, retrieval (i.e., candidate generation) and ranking. The first stage of [OLL+13] is based on textual similarity to generate the candidate list, and for the second step, RankSVM is utilized to rank the generated candidates. The most recent study [CLY+22] utilizes cosine similarity for the candidate generation phase, whereas for the ranking phase a deep neural network model is proposed. Moreover, [FLL15] presents a patent citation recommendation system for patent examiners who are usually responsible for the prior art search and assessing the patentability of patent applications. To this end, the proposed model exploits, textual content, and bibliographic information of the patents as well as the citations assigned by the patent applicant.

The aforementioned studies show that there is a large room for improvement in the recommendation results. In this paper, we focus on exploring the impact of the negative sampling strategy on patent citation recommendation.

Negative Sampling Techniques Despite the importance of negative sampling for recommender systems, literature on this topic is quite limited. [YDZ+22] proposes a negative sampling method for graph-based user-item recommendation systems. The model is sophisticated and cannot be easily applied to the other

[1] https://doi.org/10.5281/zenodo.7870197

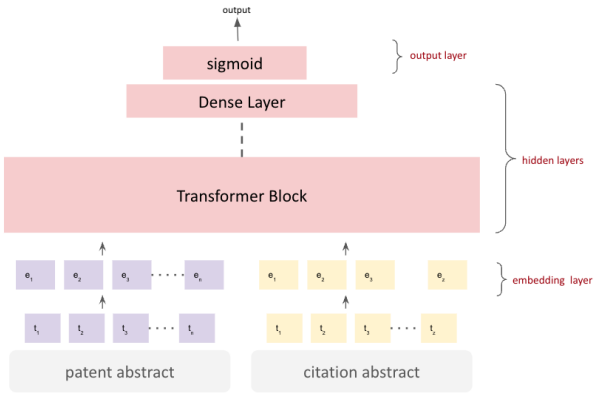

Figure 1: The general architectural overview of the ranking model.

recommendation systems, e.g., citation recommendation due to the nature of the data. The model divides the items into three regions based on the distance to the positive items. The experiments suggest that selecting negative samples from the intermediate level (i.e., items that are not too far from the positive samples) provides better performance than the items that are very close or too far from the positive samples. [HDD+21] presents a negative sampling model which is specifically designed for graph neural networks for collaborative filtering. The model utilizes a user-item graph to generate the negative samples.

The studies discussed in this section are mostly focused on items and users, however, our study focuses on patents. The patents pose the following challenges as compared to previously discussed systems: (1) often, patent-citation data is more sparse in comparison to user-item interaction data. Therefore, it is quite challenging to find the most relevant and similar patents. (2) Patents have a unique structure that consists of textual data (e.g., title, abstract, claim, description) as well as metadata (e.g., CPC and IPC code, family information, etc.).

3 Patent Citation Ranking Model

Citation recommendation (CR) systems assist patent applicants, examiners, etc. to find relevant patents that can be cited for patents under consideration. Similar to general recommendation systems, CR systems consist in general of 2 main steps namely, retrieval and ranking. In the retrieval phase, various techniques are used to identify a candidate list of citations that are potentially relevant to the given patent. In the second phase, the selected candidates are ranked with the ranking system often by applying different machine learning methods. The scores are usually $P(y|X)$, the probability of y given X such that y is a potential citation and the X is a patent. In order to compute such probability in the context of patent citation ranking systems both contextual features of the citation and the patent are exploited.

In this paper, we narrow our focus to explore the impact of different negative sampling techniques and proportions on the performance of the patent citation ranking model.

To this end, we design a transformer-based ranking model which is capable of ranking relevant as well as irrelevant citations based on a given patent accurately. Figure 1 illustrates the ranking model, i.e., the deep neural network model that has been designed for this study. It consists of a transformer block which is integrated as a layer, followed by a pooling layer, a dense layer, and a final sigmoid layer. The model takes as an input textual parts of patents and potential citations, such as abstracts, claims, and titles. Then the output of the model is $P(y = 1|X)$, where Y is a binary class label (either 1 or 0). The input of the model is 2 pieces of text both from a patent and its potential citation (e.g., title, abstract, claim, etc.), and the output is the relevancy score of the citation to the given patent. Figure 1 illustrates an example of input patent and its potential citation, first, the abstracts are tokenized and the embeddings of the tokens are utilized as an input to the transformer block. The embeddings are randomly initialized.

4 Experimental Results

In this section first, we present the negative sampling methods that we proposed. Second, the datasets that have been generated by applying the selected sampling techniques. Finally, we illustrate the obtained experimental results by exploiting the proposed ranking model which was trained on the generated datasets.

4.1 Negative Sampling Methods

In this study, we investigate three different negative sampling methods to assess the performance of the citation ranking model (see Section 3) as well as demonstrate the significance of these samples on the performance.

Following the exploited techniques are explained:

- **Random Sampling:** In this method, the negative samples are selected randomly. The recommendation datasets consist of positive samples as well as unlabeled samples. The negative samples are randomly selected from the unlabeled samples for each patent.

- **Nearest Neighbor Sampling:** First, all the patents and their citations are embedded into common vector space by exploiting the Sentence

Transformers with BERT for Patents[2] which has been trained by Google on over 100M patents. In order to obtain the embedding representation of patents and citations the abstracts have been exploited. In the second step, to find the nearest neighbor for each patent in the vector space, Faiss[3], a library for efficient similarity search of dense vectors is used.

- **CPC code-base Sampling:** The Cooperative Patent Classification (CPC[4]) is a system that is utilized to classify patents based on their technical features. The classification system consists of 9 main sections A-H and Y. Each main section consists of classes and subclasses. For generating negative samples, given a patent, we select the negative samples from the unlabeled examples of a given dataset by ensuring that the selected instances have the identical CPC subclass code as the patent.

It should be noted that the Nearest Neighbor Sampling and CPC code-based Sampling techniques aim to enable the model to distinguish between relevant and irrelevant citations from semantically similar as well as within the same technical field, respectively.

4.2 Generated Datasets

In order to apply different negative sampling methods (see Section 4.1) first we randomly collected around 250,000 US patents from Google Patents[5]. Each patent has roughly on average 27 citations. The positive samples are constructed by pairing patents with their actual citations. Since, this paper explores the impact of negative sampling techniques as well as the proportion of negative samples on the performance of the patent citation ranking model, 2 different datasets have been generated. In the first dataset, the focus is on investigating the different negative sampling techniques whereas, in the second dataset, the focus is on examining the impact of different proportions of negative samples.

By applying the above techniques we generated three different datasets which are utilized to investigate the impact of negative sampling techniques. The number of generated negative samples is equal to the number of existing positive samples in the dataset to ensure a balanced dataset. Due to the computational difficulties, we selected 1 million samples from each generated dataset. In order to compare the performance of the ranking model on different negative sampling techniques we trained three distinct ranking models by utilizing the generated datasets.

Table 1: Comparison of Performance for Different Negative Sampling Techniques

Sampling Method	Accuracy
Random	0.887
nearest-neighbor	0.71
CPC subclass	0.70

Further datasets have been generated to explore the effect of negative sample proportions. In other words, for each positive pair, a varying number of negative samples i.e., 2, 3, and 5 are generated randomly. Similarly, for each dataset, three distinct ranking models are trained.

4.3 Evaluation of Patent Citation Ranking Model with the Generated Datasets

In order to assess the performance of the ranking model three different sets of experiments have been conducted. In each experiment, the transformer-based ranking model (see Section 3) has been trained and evaluated based on a given dataset. As mentioned before, the datasets consist of positive and negative samples, where each positive sample is the actual citation of corresponding patents and the negative samples are the generated ones that are the irrelevant citations of corresponding patents.

In the first and second sets of experiments (see Table 1 and 2), the model takes the abstract of a patent and a potential citation as input and computes the probability score which is used as a ranking system for the given pair. The threshold of the ranking model is set to 0.5. The potential citation is considered to be relevant if the score is above the threshold, otherwise, it is considered to be irrelevant. In the third set of experiments (see Table 3), the same ranking system has been applied with different features. In other words, abstract, claim, and title of patents and citations have been utilized distinctly as input to the ranking model, to explore the impact of individual features on identifying the relevant and irrelevant citations.

Table 1 illustrates the performance of the ranking model on datasets that have been created by the different sampling techniques, namely, random, nearest-neighbor and CPC subclass-based. The random sampling approach which is the most straightforward one provides the best performance with 0.887 accuracy. The reason that more diverse samples have been created with random sampling is that this enables the model to distinguish between relevant and irrelevant citations. According to Table 1 results, it can be concluded that cited patents are semantically similar as well as share the same technical content.

[2]https://huggingface.co/anferico/bert-for-patents

[3]https://faiss.ai/

[4]https://www.epo.org/searching-for-patents/helpful-resources/first-time-here/classification/cpc.html

[5]https://pypi.org/project/google-patent-scraper/

Table 2: Comparison of Performance for Different Negative Sampling Proportion

Negative Sample Proportion	Accuracy
0.67 (2 neg. samples for each pos.)	0.888
0.75 (3 neg. samples for each pos.)	0.891
0.83 (5 neg. samples for each pos.)	0.911

Table 3: Comparison of Performance for Different Feature Combinations

Feature Combination	Accuracy
Abstract	0.887
Claim	0.868
Title	0.504

Table 2 presents experimental results of the ranking model on datasets which contain different proportions of randomly selected negative samples. According to the results presented in this table as the number of negative samples increases, the accuracy also increases. Conventionally, when training a machine-learning model it is a common practice to have a balanced dataset that consists of roughly, an equal number of positive and negative samples. However, depending on the problem and the domain, an imbalanced dataset could yield higher accuracy than a balanced dataset. For instance, for image classification, the experimental result of [PAHS12] shows that the imbalanced dataset enhances the performance of the ranking algorithm. Similarly in our experiments, the best performance (see Table 2) has been achieved with the imbalanced dataset. The reason here can be attributed to the model's ability to distinguish positive samples from negative samples by being trained mostly with negative samples. Further, the results also show that patents cite relevant patents and often there are no missing citations.

Finally, Table 3 illustrates the accuracy of the ranking model on different feature combinations. Typically, claims of a patent give a clear definition of what the patent legally protects, and the abstract gives a brief summary of the technical content of patent documents. Claims are often long and hard to model as a feature of a transformer-based ranking model due to their complexity. Therefore, in order to use claims as a feature, we collected from each patent and citation their first independent claims[6] which present the fundamental features of the invention. In other words, a claim focuses on a single characteristic of the invention, whereas an abstract provides a brief summary of the information presented in the description, claims, and drawings. Therefore, the abstract carries more information in comparison to single claims. Titles are often short and do not carry sufficient semantic information alone to help the model distinguish between relevant and irrelevant.

Exploding all dependent and independent claims as input to the ranking model would probably increase the accuracy due to more contextual information. However, claims are often long text, therefore it requires special effort to be modeled efficiently and effectively. We leave this as our next future work.

Overall, based on the experiments it can be concluded that negative sampling techniques that are being employed and the negative sample proportion play a significant role in the patent recommendation system.

5 Conclusion and Future Work

This paper targets the problem of negative sampling approaches for the patent citation recommendation. More specifically, it proposes a transformers-based architecture for ranking citations for citation recommendation. The features used for this purpose include patent title, abstract, and claims. It further performs an experimental comparison of various negative sampling approaches for patent recommendations such as random negative sampling, negative sampling based on nearest neighbor as well as CPC class hierarchy. The experiments were conducted on newly generated datasets extracted from Google patents. The results suggest that random negative sampling performs the best in terms of accuracy. Moreover, the most effective features are the patent abstract and the claim. In future work, we plan to employ a retrieval model to generate a candidate list for each given patent and then apply the ranking model to the candidate list to present a complete patent citation recommendation system.

References

[CLY+22] Jaewoong Choi, Jiho Lee, Janghyeok Yoon, Sion Jang, Jaeyoung Kim, and Sungchul Choi. A two-stage deep learning-based system for patent citation recommendation. *Scientometrics*, 2022.

[FLL15] Tao-Yang Fu, Zhen Lei, and Wang-Chien Lee. Patent citation recommendation for examiners. In *ICDM*. IEEE Computer Society, 2015.

[HDD+21] Tinglin Huang, Yuxiao Dong, Ming Ding, et al. Mixgcf: An improved training method for graph neural network-based recommender systems. In *SIGKDD*, 2021.

[6] https://new.epo.org/en/legal/guidelines-epc/2023/f_i v_34.html

[JWS+21] Ting Jiang, Deqing Wang, Leilei Sun, et al. Lightxml: Transformer with dynamic negative sampling for high-performance extreme multi-label text classification. In *AAAI*, 2021.

[OLL+13] Sooyoung Oh, Zhen Lei, Wang-Chien Lee, Prasenjit Mitra, and John Yen. CV-PCR: a context-guided value-driven framework for patent citation recommendation. In *CIKM*, 2013.

[PAHS12] Florent Perronnin, Zeynep Akata, Zaid Harchaoui, and Cordelia Schmid. Towards good practice in large-scale learning for image classification. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, 2012.

[TZAS20] Rima Türker, Lei Zhang, Mehwish Alam, and Harald Sack. Weakly supervised short text categorization using world knowledge. In *ISWC*, 2020.

[TZKS19] Rima Türker, Lei Zhang, Maria Koutraki, and Harald Sack. Knowledge-based short text categorization using entity and category embedding. In *ESWC*, 2019.

[YDZ+22] Zhen Yang, Ming Ding, Xu Zou, Jie Tang, Bin Xu, Chang Zhou, and Hongxia Yang. Region or global a principle for negative sampling in graph-based recommendation. *IEEE Transactions on Knowledge and Data Engineering*, 2022.