A Machine Learning Approach to Improving Nanotechnology Patent Citation Generation

Abstract

Understanding citation networks in patents - and assisting inventors and examiners in identifying patents that should be cited - is an important area in which natural language processing techniques can dramatically improve results. We seek to improve patent citation generation by combining natural language process approaches with machine learning techniques to better identify relevant patents. We use a variety of natural language processing techniques (particularly topic modeling, language models and information retrieval) as feature inputs to a machine learning model which improves on previous attempts at citation generation which have generally used these techniques to generate a score directly. We focus on patents in the nanotechnology space in order to reduce the complexity of our data.

Introduction

A broad and growing field of literature explores citation networks among both academic articles and patents in order to enhance our academic understanding of the spread of innovative ideas and for a variety of commercial applications. Academically, understanding patent citation networks helps researchers explore the diffusion of knowledge, identify influential innovations and discover under-patented areas which might provide fertile ground for future innovation. And for inventors, an ever growing body of patents – which is expanding fast enough that experts now have difficulty knowing even a very narrow subfield exhaustively - makes better technologies to explore existing patents and identify the correct references increasingly vital. Given the proliferation and complexity of patents, from the 1950's[1] [a] onwards, various improvements to keyword search have been proposed – and implemented in various proprietary systems – to enhance both inventors' and examiners' ability to find relevant patents. In this context, our goal is to develop an improved citation-generation engine for patents in the nanotechnology area[2] by applying machine learning techniques to a rich set of natural-language-processing generated features.

However, despite the practical and academic interest in the area, there are several key difficulties that make patent citation generation challenging and offer opportunities for improvement[3]. First, the size of data makes modeling the relationship between a given patent and all other patents extremely challenging. Second, the language in patents tends to be less standardized than in academic article which means that text modelling approaches may not describe the relation of the cited patent to the topic. This removes an important contextual

---

[1] M. Bailey, B. Lanham, and J. Leibowitz. Mechanized searching in the U.S. Patent Office. Journal of the Patent Office Society, 35(7):566{587, 1953.

[2] We thought we should limit our research to nanotechnology related patents in order to focus our efforts on applying Natural Language Processing techniques rather than big data management.

[3] L. Azzopardi, W. Vanderbauwhede, and H. Joho. Search system requirements of patent analysts. In SIGIR'10 , pages 775- 776. ACM, 2010

element that could help understand citation networks.

Background

Existing research into patent citation generation consists of three main aspects (1) semantic modelling (especially Subject-Action-Object or SAO based modeling), (2) topic modeling, language modeling and other standard NLP-based approaches and (3) proprietary (and vaguely documented) approaches. Focusing on the first two of these for obvious reasons, let us first consider semantic modeling.

Semantic modeling based approaches are certainly trendy in the patent-citation area and with many seeking to develop similarity and novelty measures based on SAO[4]. This technique seeks to model the key concepts and technological advancements covered in the patent by identify subjects, actions and objects (an extension of part-of-speech tagging) . However, this is clearly a somewhat limited approach and focuses very narrowly on one approach to similarity.

More traditional natural language-based approaches seek to determine patent similarity - and likelihood of citation - based on topic models or language models or a combination of the two[5]. We chose to engage with the more traditional NLP based approaches over the semantic modeling for greater flexibility and in order to engage a wider variety of the course materials. These models generally follow a series of similar steps.

First, due to the computational intensity of many topic and language modelling approaches a smaller subsample of candidate patents is created using keyword queries[6] or discriminative term buckets (which consider the citation networks within and between topic "buckets")[7]. Then, within the pool of candidate patents each patent that may be cited by a given query patent is scored based on the chosen model - often a Latent Dirichlet Allocation, a simple language model or a combination thereof[8] . These scores are used to rank the candidate patents by likelihood of citation, but generally do not perform particularly well, with mean average precisions in the <20% range[9]. One key weakness of these models is that they do not effectively and systematically combine the different scoring mechanism into the most accurate final score. So we seek to build on this literature by using the various natural language-based features identified in the literature along with structured data relating to the inventor and assignee as inputs to a machine learning model for training with checks whether the given patent was actually cited as the true label input to our model.

---

[4]eg Yoon, Janghyeok, and Kwangsoo Kim. "Identifying rapidly evolving technological trends for R&D planning using SAO-based semantic patent networks." Scientometrics 88.1 (2011): 213-228. ; A new instrument for technology monitoring: novelty in patents measured by semantic patent analysis , Jan M. Gerken and Martin G. Moehrle , 2012

[5] See Abbas, Assad, Limin Zhang, and Samee U. Khan. "A literature review on the state-of-the-art in patent analysis." World Patent Information 37 (2014): 3-13.

[6] R. Krestel, P. Smyth, Recommending patents based on latent topics. Proc. 7th ACM Conf. Recomm. Syst. - RecSys '13, 395–398 (2013)

[7] Yu, Xiao, et al. "Citation prediction in heterogeneous bibliographic networks." Proceedings of the 2012 SIAM International Conference on Data Mining. Society for Industrial and Applied Mathematics, 2012

[8] R. Krestel, P. Smyth, Recommending patents based on latent topics. Proc. 7th ACM Conf. Recomm. Syst. - RecSys '13, 395–398 (2013)

[9] ibid (same as reference 8)

Methods

Our approach consists of four key steps: (1) sampling nanotechnology related patents, (2) selecting a candidate pool of patents for each patent in our test set, (3) using a variety of natural language processing techniques to generate features, (4) training a machine learning model with the goal of correctly identifying patents that are cited by that particular patent.

For step 1[10], we followed the sampling strategy used by Arora et al to identify journal articles related to nanotechnology, which involved a combination of keyword searches for nanotechnology related terms and the exclusion of similar but unrelated terms (those relating solely to nano-measures or nano-animals)[11]. This search strategy was applied across the four main text fields available in patent data (abstract, the brief summary, detailed description and the claims). This ensures that patents that mention any of these nanotechnology associated terms are captured so we have more comprehensive coverage of potentially relevant areas. This search strategy gave us 456,345 patents over the period 2001-2017, with substantially more patents in recent years.

For step 2, we sought to narrow the search space for candidate patents in order to be able to calculate cosine and Jaccard similarities between patents. Because each of these methods calculates the similarities between all pairs of patents (with O(N2) complexity for the number of patents), calculating these for the full set of patents was unrealistically slow. Therefore, we used a topic modelling approach to divide the nanotechnology related patents into 100 sub-topic areas. Prior to performing topic modelling, pre-processing steps including filtering out of stop words and snowball stemming were performed to improve the accuracy of topics. A sample of the words that characterize each topic is shown in the results section.

Next, in step 3, we consider each sub-topic areas independently and and generate features for our machine learning model by calculating the similarity between all pairs of patents in our subtopic area. Language models combined with topic models have demonstrated improved performance over either approach individually – likely because language models capture terms associate with specific sub-domains while topic models capture more general similarities – so we include similarity scores based on both approaches. The term document in the following equations refer to the tf-idf vectors of the combined title and abstract field for a patent. For each pair of patents, we use the following three similarity measures:(i) cosine similarity, (ii) Jaccard distance and (iii) second order cosine similarity[12].

(i) Cosine Similarity is the cosine of the angle between two nonzero vectors in the inner product space.This provides a way to normalize the tf-idf scores for lengths of the documents

[10] The script 'identifying_nano_patents.py' contains the code for this step.

[11] Arora, Sanjay K., et al. "Capturing new developments in an emerging technology: an updated search strategy for identifying nanotechnology research outputs." Scientometrics 95.1 (2013): 351-370.

[12] Ahlgren, Per , et al. "Document–document similarity approaches and science mapping: Experimental comparison of five approaches"Journal of Informetrics Volume 3, Issue 1, January 2009, Pages 49–63

$$sim1(d_i, d_j) = \frac{\sum_{m=1}^{k} w_{m,i} \times w_{m,j}}{\sqrt{\sum_{m=1}^{k} (w_{m,i})^2} \times \sqrt{\sum_{m=1}^{k} (w_{m,j})^2}}$$

where,
sim1(d_i,d_j) is the cosine similarity between document d_i and d_j
w is the m^{th} component of the vector

(ii) Jaccard similarity is defined as the size of the intersection divided by the size of the union of two finite sets /documents

(iii) Second degree cosine similarity is the cosine similarity (as defined in (i)) of the cosine similarity between the query document and a pair of candidate documents.

$$sim2(d_i, d_j) = \frac{\sum_{m=1}^{N} sim1(d_m, d_i) \times sim1(d_m, d_j)}{\sqrt{\sum_{m=1}^{N} (sim1(d_m, d_i))^2} \times \sqrt{\sum_{m=1}^{N} (sim1(d_m, d_j))^2}}$$

where,
sim2(d_i,d_j) is the second degree cosine similarity between document d_i and d_j
N is the total number of documents

Finally, in step 4, these different similarity measures are provided as features to our logistic regression model. The simple logistic regression classification model allows for future expansion of our project to include structured data from the patents (like kind, classification, inventor or assignee information) or additional similarity measure approaches. This is the key contribution of our paper – combining the commonly-used natural language processing approaches to scoring similarity with a machine learning approach to improve our mean average precision. Our trained model is then scored based on its performance on our held out test data.

Results and Discussion

Our sampling strategy yields collection of patents that are clearly nano-tech related, a sample of titles includes:

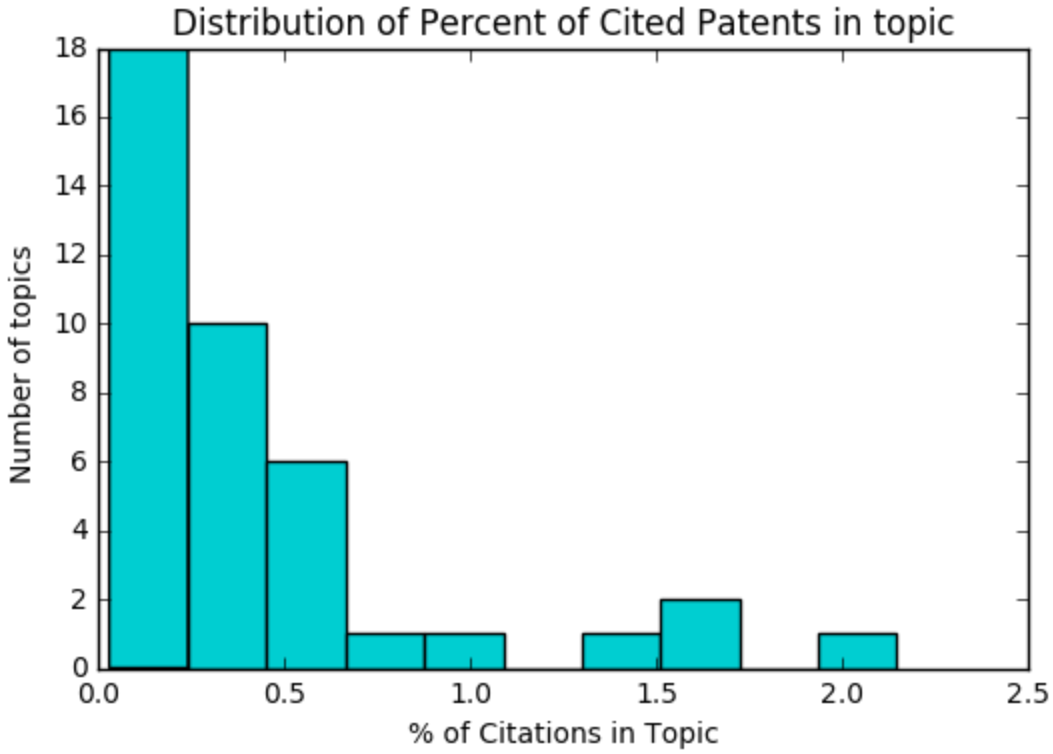
Table 1: Example Titles
Extension of a protein-protein interaction surface to inactive the function of a cellular protein
Liquid crystal alignment film, method of manufacturing the film, liquid crystal display using the film and method, and method of manufacturing the liquid crystal display
Transistor in a semiconductor device and method of manufacturing the same
Photocatalytic oxidation of organics using a porous titanium dioxide membrane and an efficient oxidant
Tactical thermal luminescence sensor for ground path contamination detection

It is important to consider to what extent these thematically related patents constitute a distinct field of research. While the strong literature that validates our keyword search strategy suggest that we are in fact successfully identifying a scientific area, it remains to be seen to what extent this area is separate from other scientific areas. To evaluate this, we determined what fractions of the patents cited by our nanotechnology patents were also in the nanotechnology field. We find that patents in this area cited 1, 436, 931 distinct patents only 8.6% of which were in the nanotechnology field themselves. This suggest that there is strong cross pollination with other areas of innovation and also puts an upper limit on the accuracy our model will achieve -- since we are only considering patents within this particular subject area, we will necessarily miss all the patents that in fact belong to a different topic area.

Considering the distinctions within the nanotechnology space itself, we find clear divisions within the nanotechnology field and we would not expect a patent focused on say, cellular proteins, to cite the same patents as one discussing transistors. Latent Dirichlet Allocation - which seeks to identify underlying unobserved groups (in this case topic areas), mixtures of which can generate the observations - was used to select a pool of different topic areas. The table below shows the top four words associated with a sample of the topics, and illustrates that this modelling approach is finding distinct logically consistent groupings.

Table 2: Word stems most strongly associated with given topics				
Topic 1	polyurethan	paper	polym	foam
Topic 37	cell	solar	heater	fuel
Topic 40	maiz	plant	produc	varieti
Topic 55	acid	amino	compris	sequen
Topic 75	carbon	diamond	layer	core

While these topics seem logically distinct and consistent - much like a human might classify these topics - it is ultimately clear that most the citations for each given patent do not fall within it particular topic area. In fact, the average topic has only .4% of its citations from within the same topic areal. This suggests that important improvements to our model would come from a different approach to partitioning the data.



Finally, turning to our complete model[13], we find that our regression model performs moderately well at predicting the likelihood of citation. Our model ends up with a binary accuracy of 98.6% (improvement from 55% accuracy in our earlier iterations) - meaning that we are very effectively predicting which patents will be cited (the vast majority will not) - and, considering the whole suggested citation list for a given patent, a mean average precision of 3%. This low MAP score reflects the issues discussed above with not all patents or all citations coming from the selected search space (both nanotechnology patents in general or each topic individually) , but our high accuracy suggests that the similarity measures are in fact strong predictors of whether a patent will be cited or not. The accuracy of our model compares extremely favorably with our baseline model - where patents were grouped randomly and then ranked by cosine similarity - which achieved an accuracy of only 10%. It also performs slightly better than a model with only cosine similarity, which achieved an accuracy of 94.3%.

Conclusions and Further Research

In conclusion, we find that our combined topic-modelling and document similarity approach yields reasonably accurate results. However, we find several key limitations of our

[13] The accuracy calculation is based off one individual topic due to time constraints on running large matrix multiplications within the second-order cosine similarity calculation. We can parallelize this process and use MRjob (as we did for the midterm submission)

approach. Perhaps most damagingly, we find that the nanotechnology space is not entirely independent - a large number of the citations from nanotechnology patents come from outside the space. This makes this method of dividing the search space less efficient.

Our research design was specifically implemented to allow several fruitful avenues of further research. First, our machine learning based prediction approach allows the integration of different similarity measures and also structured data. A natural next step would be to include structured data about a patent - the category, kind, inventor, company, number of citations and the like - as features in the model along with document similarity measures. This would potentially improve the predictive accuracy of the model. Second, a more accurate partition of the data is needed to ensure that all candidate patents are in the pool to be ranked. No matter how well our model works in theory, in practice the accuracy is limited by the completeness of the pool of data used. Therefore, alternative approaches like key-word based queries might be needed to most accurately build the candidate pool for a given patent.

References

Abbas, Assad, Limin Zhang, and Samee U. Khan. "A literature review on the state-of-the-art in patent analysis." World Patent Information 37 (2014): pages 3-13.

Arora, Sanjay K., et al. "Capturing new developments in an emerging technology: an updated search strategy for identifying nanotechnology research outputs." *Scientometrics* 95.1 (2013): pages 351-370.

L. Azzopardi, W. Vanderbauwhede, and H. Joho. Search system requirements of patent analysts. In SIGIR'10 , pages 775- 776. ACM, 2010

M. Bailey, B. Lanham, and J. Leibowitz. Mechanized searching in the U.S. Patent Office. Journal of the Patent Office Society, 1953,35(7): pages 566-587

Gerken, Jan M., and Martin G. Moehrle. "A new instrument for technology monitoring: novelty in patents measured by semantic patent analysis." *Scientometrics* 91.3 (2012): pages 645-670.

R. Krestel, P. Smyth, Recommending patents based on latent topics. *Proc. 7th ACM Conf. Recomm. Syst. - RecSys '13*, pages 395–398 (2013)

Yoon, Janghyeok, and Kwangsoo Kim. "Identifying rapidly evolving technological trends for R&D planning using SAO-based semantic patent networks." *Scientometrics* 88.1 (2011): pages 213-228.

Yu, Xiao, et al. "Citation prediction in heterogeneous bibliographic networks." Proceedings of the 2012 SIAM International Conference on Data Mining. Society for Industrial and Applied Mathematics, 2012

Ahlgren, Per , et al. "Document–document similarity approaches and science mapping: Experimental comparison of five approaches"Journal of Informetrics Volume 3, Issue 1, January 2009, pages 49–63