Forecasting Patent Growth By Combining Time-Series Signals Using Covariance Patterns

Manajit Chakraborty
Università della Svizzera italiana (USI)
Switzerland
manajit.chakraborty@usi.ch

Seyed Ali Bahrainian
Università della Svizzera italiana (USI)
Switzerland
seyed.ali.bahreinian@usi.ch

Fabio Crestani
Università della Svizzera italiana (USI)
Switzerland
fabio.crestani@usi.ch

ABSTRACT

Bibliometrics has been employed previously with patents for technological forecasting. The primary challenge that technological forecasting faces is early-stage identification of technologies with the potential to have a significant impact on the socio-economic landscape. Bibliographic measures such as citations, are a good indicator of technological growth. With this intuition, we carry out an exploratory study using various time-series models and topic modeling over patent content to predict the growth or decline of various bibliographic measures for topics in the near future. Intuitively, in order to effectively uncover these citation trends shortly after the patents are issued, we need to look beyond raw citation counts and take into account both the geographical and temporal information. We posit that, instead of using only citation counts for time-series prediction, judicious use of signals from topics generated from documents belonging to various geographical locations can help improve the performance. We carry out experiments on a large collection of patents and present some insightful results and observations.

CCS CONCEPTS

• **Information systems** → *Data analytics.*

KEYWORDS

Correlation Analysis, Patent Citations, Prediction, Time series, Topic Modeling, LDA, Citation Growth.

1 INTRODUCTION

A patent is a contract between the inventor or assignee and the state, granting a limited period of time to the inventor to exploit the invention. The reasons for patenting could be myriad, ranging from the elementary need for exclusive rights to a technology or invention to building a positive image of an enterprise. Patents are pivotal for technological innovation in the context where they apply. They can be used to generate revenues, encourage synergistic partnerships, or to create a market advantage and be the basis for technological development.

Patent citations, namely references to prior patent documents and the state-of-the-art included therein, and their frequency are also often used as indicators for the technological and commercial value of a patent [20]. Citations are also used to identify "key" patents, which often varies depending on the nature of the technology. In pharmaceutical technologies, *e.g.*, a patent on one important substance can be determined as a key patent. However, for more involved and complex technologies, such as those used in renewable energy, patents are usually built upon existing technologies. In such cases, it can sometimes be difficult to identify a clear-cut key patent.

Technological forecasting has already been endorsed as an integral element to stay ahead of the curve for corporations and governments [7]. Previous studies, like the one by Acs *et al.* [1], suggested that patents provide a fairly reliable measure of innovative activity. Citation analysis and especially bibliometrics [3] has been used on citation graphs to identify similar works or to calculate the impact factor of journals, researchers *etc.* Predicting citation counts for patents is non-trivial and also less useful because citation counts in patents do not change as rapidly as in scholarly papers or other web articles . On the other hand, the change in citation counts, which refers to the rise or decline of the patents of certain categories could provide us a quantitative as well as a qualitative overview of patent landscape. It will also indicate which topics, and in turn which technological classes, are supposed to get traction in the upcoming years. Discovering topics from patents and analyzing their evolution over time is beneficial for making important decisions by research institutes, corporations, funding agencies, governments and any other organization involved in production or promotion of intellectual properties. For example, research funding organizations can adjust their granting policies based on insights produced by predictive models in order to favor topics that are trending and gaining increasing attention rather than those that are losing momentum and interest.

There are several factors which determine how innovation evolves in a particular geographical location over a period of time which includes political, social, environmental and judicial policies among others. While it is nearly impossible to chart all the factors and measure its impact on innovation, investigating how innovation grows irrespective of such influences is still important. While topic models have been used to forecast emerging technologies from the vantage point of technological classes [28], they have not been used in tandem with citations or to chart the citation growth of technology classes. In this paper, we assume that technology classes are represented by a group of patents which can further be delineated by topics drawn from them. We hypothesize that, if we can intelligently leverage the information from both citations and full-text of patents through topic modeling with additional inputs from the geographical regions from where certain topics emerge, it should aid us in predicting the growth of citations in the next time slice with increased accuracy. In light of this, our contributions are two-fold:

https://hbr.org/1967/03/technological-forecasting
https://ec.europa.eu/invest-in-research/pdf/download_en/final_report_hcp.pdf

- To provide a time-series representation that exploits and infuses the signals arising from patent (text) content, geographical region and accompanying bibliographic measures.
- To improve citation growth prediction using the infused signal with various time-series models.

In order to achieve our first goal, we propose a correlation analysis scheme, which we term as CORrelation Analysis using COvariance (CORACO), to harness the *best of both worlds* from topics derived using LDA [5] from patents and bibliographic measure namely citation counts. The second goal is realized by employing regression based time-series models on these infused correlated components. Our experiments are performed on a large open-source patent collection, MAREC. The efficiency of our approach is corroborated by the significantly improved prediction performance over three different baseline models.

2 RELATED WORKS

It has already been established that statistical analysis of international patent records is a valuable tool for corporate technology analysis and planning. Patents provide a wealth of detailed information, comprehensive coverage of technologies and countries, a relatively standardized level of invention, and long time-series of data [21]. So, it essentially provides us with an indicator to measure technological growth, which in turn could be extrapolated to get a better understanding of the relation and mutual dependence of innovation and economics [16, 22]. One such study to analyze how quantitative R&D and technology indicators may be used to forecast company stock price performance was carried out by Patrick Thomas [26]. On the other hand, full-text analysis of patents using topic modeling has also yielded interesting insights for technological classification, clustering and prediction. With this in mind, the existing literature can be grouped into two broad themes in the context of our research problem:

2.1 Use of Citations for Technological Forecasting

One of the early studies to measure the technological impact based on patent citations was done by Karki [17]. He proposed a host of technological indicators based on citations among patents. Some studies, like the one by Albert et al. [2], have considered only citations counts as indicators of industrially important patents. Zhang *et al.*[30] proposed to weight 11 indicators of patent value using Shannon entropy, and selected forward citations as one of the most important indicators for technological value. The basic motivation for using citations received as an indicator of quality is that citations indicate some form of knowledge spillovers. As argued by Jaffe *et al.* [14], citations reflect the fact that either a new technology builds on an existing one, or that they serve a similar purpose.

2.2 Topic Models for Patent Analysis

Latent Dirichlet Allocation (LDA) is a generative topic model which finds latent topics in a text corpus, based on the assumption that authors generally write documents with respect to specific topics [5]. Using the LDA process, a document is represented as a mixture of topics that produce words with certain probabilities. Unlike latent semantic analysis [8], the topics coming from LDA are easier to interpret, because they are represented by combinations of words with contribution probabilities for each topic [29]. Besides, LDA is known as one of the best topic models when dealing with a large corpus and to interpret the identified latent topics [5].

LDA is known to outperform other dimension-reduction techniques when dealing with a large corpus and to interpret the identified latent dimensions [5]. Regarding patent-based analysis, studies have applied LDA to the technological trend identification of greenhouse gas reduction technology [18], knowledge organization system development [12], and firms' technological concentration trends on patent subjects [28]. LDA can identify sub-topics for a technology area composed of many patents, and represent each of the patents in an array of topic distributions. Kim *et al.* [19] use LDA for visualizing development paths among patents through sensitivity analyses based on semantic patent similarities and citations. Here, the authors use LDA to identify sub-topics of a given technology. Topic models have also been employed for patent classification [27] among other problems. Time-Series analysis has been previously used by Holger Ernst [9] to examine the relationship between patent applications and subsequent changes of company performance. We aim to harness the power of both the bibliographic measures and topic modeling to provide us with a more accurate prediction of citation growth using several time-series models.

3 METHODOLOGY

3.1 Topic Extraction

Our first step involved extraction of topics from the collection of patent documents that suitably indicate the latent themes of the collection. For this, we employed LDA with default parameters. Since the document collection is large and we needed to find the best representation of the data using LDA, we performed a optimum topic number estimation. To this end, similar to the method proposed by Griffiths and Steyvers [10], we performed a model selection process. This consists of keeping the LDA Dirichlet hyperparameters (commonly known as α and β) fixed and assigning several values to K (parameter for controlling the number of topics). We computed an LDA model for each assignment, and subsequently, we picked the model that satisfies:

$$\arg\min_K \log P(W|K)$$

where W indicates all the words in the corpus. We repeated this process for K from 100 to 600 in steps of 50 to find the optimal number of topics for all the time slices. We found the optimum value at $K = 500$ topics.

The next step involved measuring the *topic strength* $\mathcal{T}_j$ of each topic t_j per year y which can be defined as in Equation 1, where $|D_y|$ denotes the total number of documents in year y.

$$\mathcal{T}_{j,y} = \sum_{i=1}^{|D_y|} \frac{p(t_j|d_i)}{|D_y|} \tag{1}$$

The topic probabilities, $p(t_j|d_i)$, are produced by the LDA as scores for each document d_i along with the topics corresponding to d_i.

http://www.ifs.tuwien.ac.at/imp/marec.shtml

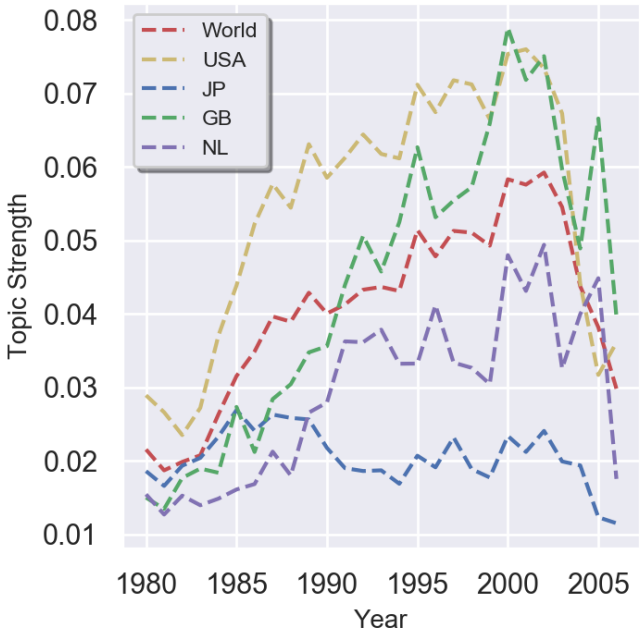

Figure 1: Topic Strength Distribution

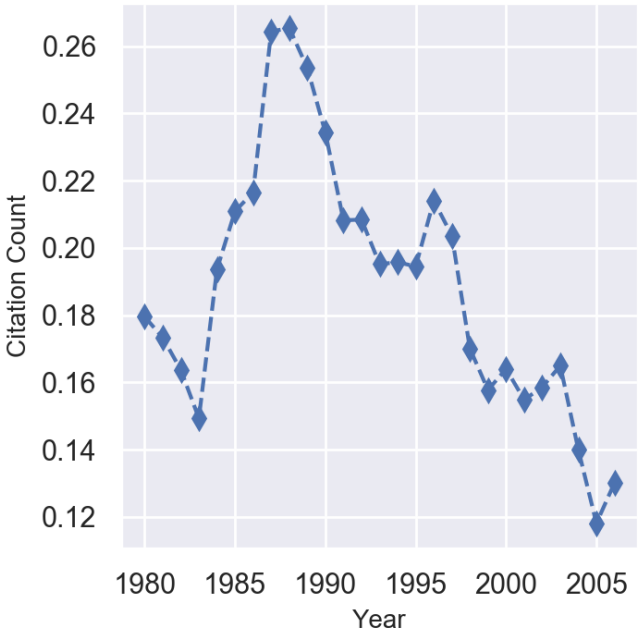

Figure 2: Citation Count Distribution

Additionally, we also compute topic strength for each of the six continents, since Antarctica has no patents, (listed in Section 4.2), $\mathcal{T}_j^{continent}$, and 127 countries, $\mathcal{T}_j^{country}$, in the dataset for each of the years 1980 through 2006. The topic strength distribution helps us gauge the change in the importance of a topic over a given period. This measure is better than topic frequency since it not only considers topic count but also accounts for the potential contribution of a certain topic t_j to some document d_i. For instance, the topic:

- ***Topic***: 2 (Mobile Radio Station)
- ***Words***: 0.236*"station" + 0.158*"mobile" + 0.088*"radio" + 0.052*"stations" + 0.014*"uplink" + 0.013*"downlink" + 0.011*"access" + 0.010*"quality" + 0.010*"cellular" + 0.008*"traffic"

extracted from our collection, which corresponds to "mobile radio stations" depict the underlying topic strength distribution as in Figure 1. From this figure, we can observe how the topic distribution changes with time depending on the geographical region (in this case, countries). When observed across topics, it also gives a qualitative overview of which countries are more invested in which topics.

3.2 Correlated Time-Series Analysis

While topic strength analysis may provide us with some clues regarding how the topics and the corresponding patents related to the topics are changing over a certain period of time; citations provide us with a more tangible resource which helps us measure the change in interest towards certain patents and by association, certain topics. Hence, our next step was to observe the changes in citation patterns for topics over the 27 year period. The normalized citation count $C_{j,y}$ for topic t_j for year y is determined according to Equation 2:

$$C_{j,y} = \sum_{t_j \rightarrow d_i \in D_y} \frac{c_{d_i,y}}{|Dy|} \tag{2}$$

where $c_{d_i,y}$ denotes the number of citations received by document d_i in year y. For the same topic example, "mobile radio station" in the previous section, the corresponding citation count distribution is presented in Figure 2.

Now, that we have two different distributions or signals from text of patents *i.e.* topic strength and a bibliometric measure *i.e.* citations; we would like to judiciously combine them in such a way that it maximizes the accuracy of prediction of citation growth (or decline). Our objective thus translates to quantitatively capturing the correlation between these two signals.

CORACO: In this paper, we propose a method for finding correlation components between two such independent distributions that maximize the commonality of two signals. We call this method as CORrelation Analysis with COvariance (CORACO). Thus, our problem can be redefined as a problem of finding two sets of basis vectors, one for $\mathbf{a}$ and the other for $\mathbf{b}$, such that the projections of the variables onto the covariance matrix of the two signals would be maximized. Here, for simplicity let us assume, $\mathbf{a}$ and $\mathbf{b}$ are placeholders for $\mathcal{T}_k$ and C_k for any given topic t_k, respectively. Let us assume the linear combinations $a = \mathbf{a}^T \hat{\mathbf{w}}_a$ and $b = \mathbf{b}^T \hat{\mathbf{w}}_b$ of the two variables $\mathbf{a}$ and $\mathbf{b}$ respectively, where $\hat{\mathbf{w}}_a$ and $\hat{\mathbf{w}}_b$ are canonical weights. We want to consider the case where only one pair of basis vectors are required corresponding to the largest correlation

component. This indicates that the function to be maximized is:

$$\begin{aligned}\rho &= \frac{E[ab]}{\sqrt{E[a^2]E[b^2]}} \\ &= \frac{E[\mathbf{a}^T\hat{\mathbf{w}}_a.\mathbf{b}^T\hat{\mathbf{w}}_a]}{\sqrt{E[\hat{\mathbf{w}}_a^T\mathbf{a}.\mathbf{a}^T\hat{\mathbf{w}}_a]E[\hat{\mathbf{w}}_b^T\mathbf{b}.\mathbf{b}^T\hat{\mathbf{w}}_b]}} \\ &= \frac{\mathbf{w}_a^T\mathbf{C}_{ab}\mathbf{w}_b}{\mathbf{w}_a^T\mathbf{C}_{aa}\mathbf{w}_a.\mathbf{w}_b^T\mathbf{C}_{bb}\mathbf{w}_b}\end{aligned} \tag{3}$$

The maximum correlation component can thus be defined as the maximum value that ρ can assume with respect to $\mathbf{w}_x$ and $\mathbf{w}_y$. The subsequent correlation components are uncorrelated for different solutions, *i.e.*:

$$\begin{cases} E[a_i a_j] = E[\mathbf{w}_{ai}^T\mathbf{a}.\mathbf{a}^T\mathbf{w}_{aj}] = \mathbf{w}_{ai}^T\mathbf{C}_{aa}\mathbf{w}_{aj} = 0 \\ E[b_i b_j] = E[\mathbf{w}_{bi}^T\mathbf{b}.\mathbf{b}^T\mathbf{b}_{bj}] = \mathbf{w}_{bi}^T\mathbf{C}_{bb}\mathbf{w}_{bj} = 0 \text{ for } i \neq j. \\ E[a_i b_j] = E[\mathbf{w}_{ai}^T\mathbf{a}.\mathbf{b}^T\mathbf{w}_{bj}] = \mathbf{w}_{ai}^T\mathbf{C}_{ab}\mathbf{w}_{bj} = 0 \end{cases} \tag{4}$$

The projections onto $\mathbf{w}_a$ and $\mathbf{w}_b$, *i.e.* a and b, describe the underlying "latent" variables. Now, we know that for any two random variables $\mathbf{m}$ and $\mathbf{n}$ with zero mean, the total covariance matrix can be represented as:

$$\mathbf{C} = \begin{bmatrix} \mathbf{C}_{mm} & \mathbf{C}_{mn} \\ \mathbf{C}_{nm} & \mathbf{C}_{nn} \end{bmatrix} = E\left[\begin{pmatrix} \mathbf{m} \\ \mathbf{n} \end{pmatrix} \quad \begin{pmatrix} \mathbf{m} \\ \mathbf{n} \end{pmatrix}^T \right] \tag{5}$$

is a square matrix where $\mathbf{C}_{mm}$ and $\mathbf{C}_{nn}$ are the intra-set covariance matrices of m and n respectively and $\mathbf{C}_{mn} = \mathbf{C}_{nm}^T$ is the inter-set covariance matrix.

Thus, the correlation components between $\mathbf{a}$ and $\mathbf{b}$ can be found by solving the eigenvalue equations:

$$\begin{cases} \mathbf{C}_{aa}^{-1}\mathbf{C}_{ab}\mathbf{C}_{bb}^{-1}\mathbf{C}_{ba} = \rho^2\hat{\mathbf{w}}_a \\ \mathbf{C}_{bb}^{-1}\mathbf{C}_{bx}\mathbf{C}_{aa}^{-1}\mathbf{C}_{ab} = \rho^2\hat{\mathbf{w}}_b \end{cases} \tag{6}$$

where the eigenvalues ρ^2 are the squared correlations and the eigenvectors $\mathbf{w}_a$ and $\mathbf{w}_b$ are the normalized correlation basis vectors. The number of non-zero solutions to these equations are limited to the smallest dimensionality of $\mathbf{a}$ and $\mathbf{b}$.

Country	No. of docs.
JP	159,433
US	148,434
GB	23,869
NL	21,767
IT	15,795
SE	6,208
DE	4,799
CH	3,990
CA	3,740
KR	3,634
..	..
World	567,547

Table 2: Distribution of patents for top-10 countries

In our case, the random variables $\mathbf{a}$ and $\mathbf{b}$ correspond to vectors $\mathcal{T}_k$ and C_k for any given topic t_k. The length of both these vectors, $\mathcal{T}_k$ and C_k, is 27 since we are observing the values for 27 years (1980-2006). Now, while C_k is fixed, the $\mathcal{T}_k$ can vary depending on the geographical region. North America, Asia and Europe has the largest share of patents, as depicted in Table 1. Also, from Table 2 we observe the distribution of patents by country. We chose to focus on the top-3 countries for our analysis, since they contribute a large share (58.5%) of the total patents produced in the world. Hence, we need to compute the following four sets of CORACO components:

$$\begin{aligned} X_{world}, Y_{world} &= \text{CORACO}(\mathcal{T}_k^{world}, C_k) \\ X_{\text{JP}}, Y_{\text{JP}} &= \text{CORACO}(\mathcal{T}_k^{\text{JP}}, C_k) \\ X_{\text{US}}, Y_{\text{US}} &= \text{CORACO}(\mathcal{T}_k^{\text{US}}, C_k) \\ X_{\text{GB}}, Y_{\text{GB}} &= \text{CORACO}(\mathcal{T}_k^{\text{GB}}, C_k) \end{aligned} \tag{7}$$

In Figure 3, we present the infused signals as provided by CORACO for the World for the same topic as in Section 3.1. By *World* we mean the complete set of patent documents in the collection. The CORACO signals are new projected signals onto the covariance of the original signals. We can observe that CORACO successfully combines the distribution of topic strength of the World with its corresponding Citation Count distribution such that it minimizes points in the distribution where the covariance is large (*e.g.* between 1990-1995 in Figure 3). Essentially, it tries to bring both the signals closer on a singular scale to achieve maximum points of similarity. These reinforced signals are then given to time-series models as input for prediction of citation growth. As results in Section 5 will show, using CORACO components instead of raw citation counts indeed improves the performance thus validating our hypothesis.

3.3 Citation Growth Prediction

The final step is concerned with prediction based on the CORACO components. We argue using CORACO components as input to the time-series models instead of using the raw distribution of citation counts for topics could significantly enhance the performance. This stems from the fact that CORACO components are representations for the correlation between two distributions which should be able to model the commonalities better. With this in mind, we choose to employ three different time-series models:

1. Linear Regression or Autoregression (AR)
2. Moving Average (MA)
3. Simple Exponential Smoothing (SES)

Due to lack of apparent trends or seasonality attributes in the observed variables, we could not use other models such as the Autoregression Moving Average (ARMA), the Autoregressive Integrated Moving Average (ARIMA) or the Seasonal Autoregressive Integrated Moving-Average (SARIMA). It is imperative to mention that our objective is to observe and predict the direction of change in number of citations (increase or decrease) *i.e.* polarity (ΔC_k) of the citations for the next time window. As stated earlier, we consider one year as a time window. Thus, we are not concerned with predicting the actual *number* of citations that a topic is supposed to gain in the next time window. This is because the number of citations a patent receives can vary on several exogenous factors such as

https://www.itl.nist.gov/div898/handbook/pmc/section4/pmc4.htm

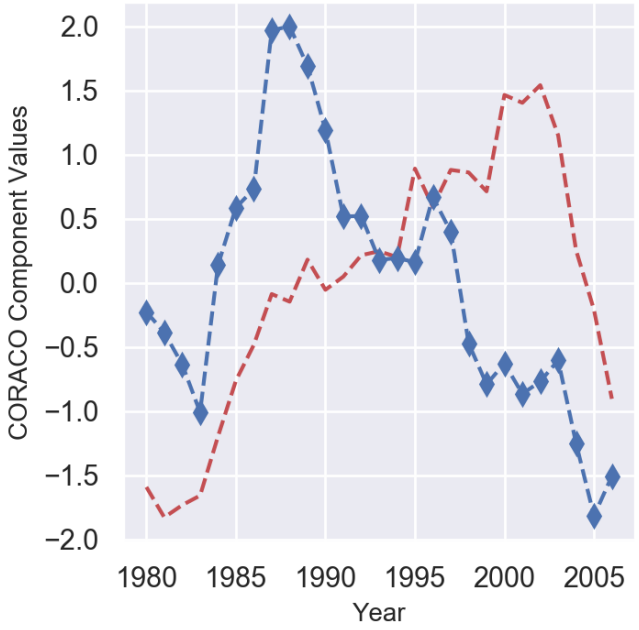

Figure 3: CORACO components for $\mathcal{T}^{world}$, C^{world}

Continent	No. of docs.
North America (NA)	220,153
Europe (EU)	162,845
Australia (OC)	4,129
South America (SA)	664
Africa (AF)	890
Asia (AS)	178,288
Antarctica (AN)	0
World	567,547

Table 1: Distribution of patents by Continents

niche popularity of the technological area, a continuation of similar research or product by a company or group of companies, a country or region's sudden interest in a particular technological class *etc.* The trend of numbers of citations for any patent and thereby any topic is non-decreasing, since citations are accumulative. Therefore, our goal is to predict whether the number of citations for a particular topic is going to increase or decrease in the next time window compared to the last time slice. It then gives us an indication of how popular a topic (and by extrapolation a technological class) is going to be in the near future and provides funding agencies, corporations and governments to adjust their funding strategies accordingly in areas which show a strong upward growth in the next few years. The results and further analysis of the prediction performance are discussed in Section 5.

4 EXPERIMENTAL SETUP

4.1 Dataset

For this study, we used the European Patent (EP) collection from the MAtrixware REsearch Collection (MAREC). MAREC is a static collection of patent applications and granted patents in a unified file format normalized from EP, WO, US, and JP sources, spanning a range from July 1976 to June 2008. The collection contains documents in several languages, the majority being English, German and French, and about half of the documents include full text. In MAREC, the documents from different countries and sources are normalized to a common XML format with a uniform patent numbering scheme and citation format. The standardized fields include dates, countries, languages, references, person names, and companies as well as rich subject classifications. It is a comparable corpus, where many documents are available in similar versions in other languages. In particular, we considered only English language patents of the EP sub-collection. We had to discard a few documents with one or more missing fields such as classification codes, patent citations, applicant country *etc.* The final dataset amounted to 567,547 documents. The citation network built out of this reduced dataset consisted of 646,537 citations. Admittedly, the citation network is very sparse, which conforms to the norm that patents are not as frequently cited as academic publications [6].

4.2 Preprocessing

The patent collection has mainly two types of documents: (1) Type A (A1, A2 ...): European patent application files. (2) Type B (B1, B2 ...): European patent specification files. Of these, we used A1 and B1 documents since they are the most informative ones and contain the textual content of the patent. Now, it is imperative to mention that in the collection, not all A1 documents have a corresponding B1 document. A1 documents contains *Abstract* along with other bibliographic information including citations while B1 documents contain bibliographic information with *Description* and *Claims* of the patent. Given this, we had to confine our collection to only patents that had both A1 and B1 counterparts. In case any one of them was missing, we did not include them in our collection. This step reduced our initial collection of English-language patents from 837,715 to 567,547. Post this step, we extracted all relevant bibliographic information (such as application date, grant date, classification codes, applicant name, applicant country *etc.*) including citations in a separate file.

We combined the title, abstract, description and claims for each patent into a single document which we refer to as 'full-text' of the patent in our paper. It should be noted that in this paper, we have

used the terms *full-text of patents* and *documents* interchangeably. Additional preprocessing steps include stopword removal (using an extended stopword list of over 800 words combining NLTK stopwords and other open source libraries), expunging unintelligible and non-alphabetic terms and tokenization. The cleaned documents were then segregated based on sectors of technology (A-H) they belong to and their countries and continents of origin (Table 1). The continent-wise distribution of the patent documents is presented in Table 1. The documents are also arranged by their year of patent registration date. So, each time slice is considered as a *year*. Based on this distribution, we chose to discard the documents from the years 1978, 1979, 2007 and 2008, since there are very few documents in these years. Retaining these documents tends to skew the topic distribution negatively. The total number of these discarded documents amount to less than 1% of the whole collection. So, essentially our patent dataset consists of patents from the year 1980 through 2006.

4.3 Tools

For LDA, we used the open source tool Gensim [24], with default settings. The time-series models were employed from the *statsmodels* package [25] built in Python. Other Python packages used include matplotlib [13], SciPy [15], scikit-learn [23], pyconvert-country *etc.*. For experiments, we used a Linux based server with Intel(R) Xeon(R) CPU @ 2.70GHz with 32 cores and 256 GB memory.

5 RESULTS AND ANALYSIS

Baseline. For the baseline, we consider the actual citation count distribution, C_k, as input to the three time-series models. So, this method basically tries to predict the change in citation counts based on the historical data giving us four baseline models. We feed the citation counts for each of the 500 topics as a vector for years 1980-2005 as training data. The output from the time-series models are then compared with the true labels of change of citation counts for the last time slice (2006). The true labels indicate whether the citation count has actually risen or fallen from previous time step. The labels are predetermined and can be one of the three types:

- *UP*: indicating that next time slice will receive more citations than the current time slice.
- *DOWN*: indicating that next time slice will receive less citations than the current time slice.
- *UC*: indicating that next time slice will receive as many citations as the current time slice.

The label 'UC' (unchanged) is a typical case and occurs only for 38 topics.

Metric. The metric for calculating the performance of CORACO based prediction is the ratio of correctly predicted labels against all true labels for 500 topics.

$$\text{Accuracy}_{\text{model}} = \frac{|\text{Correctly Predicted labels for } \Delta C_k^{n+1}|_{\text{model}}}{|\text{True label for } \Delta C_k^{n+1}|}$$

The citation counts are integers, while the topic strengths are described by decimal numbers, and also their scales are different. So, we had to perform min-max normalization on citation counts before computing the CORACO components for each case. The comparative performance of the four models using four different geographical regions (World, Japan, USA and Great Britain) are presented in Tables 3 and 4. In Table 3, all the 500 topics are considered, while in Table 4, only topics which do not have 'UC' (unchanged) as their labels are considered. This is because, for any prediction algorithm, it is difficult to accurately predict the last element in the series for the next time step. Even if the predicted and actual values are close, by our accuracy metric, it would be considered as either one of 'UP' or 'DOWN'. We can clearly observe that by eliminating 'UC' labeled topics, the performance of all models including baseline improves by a small margin. Also, the number of such topics, 38, is quite low (7.6% of the total number of topics). From this table we can observe that in the best case, with Simple Exponential Smoothing, we achieve 78.3% better performance in prediction when compared to all four baseline models. Among all the CORACOs, $\text{CORACO}_{\text{GB}}$ is the worst performer with Autoregression model and still it records a 39.9% improvement. In terms of CORACO components, the topic strength distribution of the world is shown to provide synergistic improvement to the prediction of polarity of the citations. While, among the three countries United States of America, has the biggest influence in improving the predictions even though it is not the largest country in terms of patent production in our dataset. Comparing among time-series models, Simple Exponential Smoothing seems to provide the biggest gain when CORACO signals are provided as input but fails poorly for the baseline input. The improvements achieved by the CORACO models are statistically significant and hence the corresponding results have been marked with an asterisk in the tables.

Prediction Error Comparison

While, our proposed models perform better than baselines, in terms of citation growth performance, it is interesting to also compare the error in prediction of citation count values. In Table 5, we present the Mean Absolute Error (MAE) of the baseline models as well as our proposed models. It must be noted that since the CORACO components have a different scale compared to baseline models, we applied min-max normalization on all prediction error vectors (for 500 topics). From the table, we can clearly observe that the error produced by our CORACO approaches are lower than that of baselines which operate on actual citation counts. In general, $\textbf{CORACO}_{\text{world}}$ gives the best performance similar to citation growth prediction. So, we can positively conclude that not only does our proposed models perform better with respect to citation growth prediction accuracy but even the citation count prediction errors are lower than all three baseline models.

6 CONCLUSIONS AND FUTURE WORK

Patent analysis delivers comprehensive competitive intelligence about innovators, relevant technologies, and help estimate the value of patents owned by competitor companies and governments. Patent Citations and Topic Models have been employed separately in existing literature towards forecasting of technological growth. In this paper, we proposed a novel approach that leveraged both patent citation counts and topic importance with geographical relevance to improve the prediction of patent citation growth in

https://www.nltk.org/
https://pypi.org/project/pycountry-convert/

Model \ Input	Baselines	CORACO$_{\text{world}}$	CORACO$_{\text{US}}$	CORACO$_{\text{JP}}$	CORACO$_{\text{GB}}$
AR	0.108	0.452*	0.446*	**0.454***	0.428
MA	0.086	**0.396***	0.374*	0.390*	0.386*
SES	0.048	**0.470***	0.416*	0.438*	0.434*

Table 3: Accuracy comparison of models with all labels. Best performances are marked in bold. Statistically significant results are marked with an asterisk (*)

Model \ Input	Baselines	CORACO$_{\text{world}}$	CORACO$_{\text{JP}}$	CORACO$_{\text{US}}$	CORACO$_{\text{GB}}$
AR	0.139	0.492*	0.476*	**0.499***	0.463*
MA	0.193	**0.428***	0.414*	0.402*	0.417
SES	0.148	**0.605***	0.550*	0.574*	0.570*

Table 4: Accuracy comparison of models without UC labels. Best performances are marked in bold. Statistically significant results are marked with an asterisk (*)

Model \ Input	Baselines	CORACO$_{\text{world}}$	CORACO$_{\text{JP}}$	CORACO$_{\text{US}}$	CORACO$_{\text{GB}}$
AR	0.0371	**0.0315**	0.0321	0.0329	0.0324
MA	0.0416	0.0363	0.0358	0.0359	**0.0357**
SES	0.0383	**0.0320**	0.0330	0.0321	0.0334

Table 5: Prediction Error (MAE) comparison of models. Best performances are marked in bold.

the next year. To this end, we proposed a covariance based correlated time-series method that maximizes the similarity of two distributions. For prediction, we employed three time-series models and compared our approach against three baseline models by also providing a comparative overview of the geographical region's influence on the prediction. Our results substantiate our hypothesis that correlated time-series model modifies the signal in such a way that is superior to all baseline models using the original time-series vectors.

As part of our future work, we would like to study the impact of our proposed approach on other complex time-series models such as LSTM networks [11]. We will investigate ways to extend our model to higher dimensions such that we could find representations of multiple signals. We would also like to employ dynamic topic models for topic elicitation such as the one proposed by Bahrainian *et al.* [4] to account for topic evolution over time. Lastly, we will apply our model to other time-series problems other than patent analysis.

ACKNOWLEDGEMENTS

We thank the anonymous reviewers for their valuable comments. This work was partially supported by *The Global Structure for Knowledge Networks* project grant under the SNSF National Research Programme 75 "Big Data" (NRP 75).

REFERENCES

[1] Zoltan J Acs, Luc Anselin, and Attila Varga. 2002. Patents and innovation counts as measures of regional production of new knowledge. *Research Policy* 31, 7 (2002), 1069 – 1085.

[2] M.B. Albert, D. Avery, F. Narin, and P. McAllister. 1991. Direct validation of citation counts as indicators of industrially important patents. *Research Policy* 20, 3 (1991), 251 – 259.

[3] Leonidas Aristodemou and Frank Tietze. 2018. Citations as a measure of technological impact: A review of forward citation-based measures. *World Patent Information* 53 (2018), 39 – 44.

[4] Seyed Ali Bahrainian, Ida Mele, and Fabio Crestani. 2018. Predicting Topics in Scholarly Papers. In *Advances in Information Retrieval - 40th European Conference on IR Research, ECIR 2018, Grenoble, France, March 26-29, 2018, Proceedings*. 16–28.

[5] D.M. Blei, A.Y. Ng, and M.I. Jordan. 2003. Latent Dirichlet allocation. *Journal of Machine Learning Research* 3, 4-5 (2003), 993–1022.

[6] Setfano Breschi, G Tarasconi, C Catalini, L Novella, P Guatta, and H Johnson. 2006. Highly Cited Patents, Highly Cited Publications, and Research Networks. *CESPRIBOCCONI UNIVERSITY, http://ec. europa. eu/invest-in-research/pdf/download_en/final_report_hcp. pdf (Accessed: 01/07/2019)* (2006).

[7] Richard S. Campbell. 1983. Patent trends as a technological forecasting tool. *World Patent Information* 5, 3 (1983), 137 – 143.

[8] Scott Deerwester, Susan T. Dumais, George W. Furnas, Thomas K. Landauer, and Richard Harshman. 1990. Indexing by latent semantic analysis. *Journal of the American Society for Information Science* 41, 6 (1990), 391–407.

[9] Holger Ernst. 2001. Patent applications and subsequent changes of performance: evidence from time-series cross-section analyses on the firm level. *Research Policy* 30, 1 (2001), 143 – 157.

[10] Thomas L. Griffiths and Mark Steyvers. 2004. Finding scientific topics. *Proceedings of the National Academy of Sciences* 101, suppl 1 (2004), 5228–5235.

[11] Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long Short-Term Memory. *Neural Computation* 9, 8 (1997), 1735–1780.

[12] Zheng Hu and Xin-Yan Deng. 2014. Aerodynamic interaction between forewing and hindwing of a hovering dragonfly. *Acta Mechanica Sinica* 30, 6 (01 Dec 2014), 787–799.

[13] J. D. Hunter. 2007. Matplotlib: A 2D graphics environment. *Computing in Science & Engineering* 9, 3 (2007), 90–95.

[14] Adam B Jaffe, Manuel Trajtenberg, and Michael S Fogarty. 2000. *The meaning of patent citations: Report on the NBER/Case-Western Reserve survey of patentees*. Technical Report. National bureau of economic research.

[15] Eric Jones, Travis Oliphant, Pearu Peterson, et al. 2001–. SciPy: Open source scientific tools for Python. http://www.scipy.org/ [Online; accessed 01/07/2019].

[16] Junegak Joung and Kwangsoo Kim. 2017. Monitoring emerging technologies for technology planning using technical keyword based analysis from patent data.

Technological Forecasting and Social Change 114 (2017), 281 – 292.

[17] M.M.S. Karki. 1997. Patent citation analysis: A policy analysis tool. *World Patent Information* 19, 4 (1997), 269 – 272.

[18] Gabjo Kim, Sangsung Park, and Dongsik Jang. 2014. Technology Analysis from Patent Data Using Latent Dirichlet Allocation. In *Soft Computing in Big Data Processing*, Keon Myung Lee, Seung-Jong Park, and Jee-Hyong Lee (Eds.). Springer International Publishing, Cham, 71–80.

[19] Mujin Kim, Youngjin Park, and Janghyeok Yoon. 2016. Generating patent development maps for technology monitoring using semantic patent-topic analysis. *Computers & Industrial Engineering* 98 (2016), 289 – 299.

[20] Doug Lichtman and Mark A Lemley. 2007. Rethinking Patent Law's Presumption of Validity. *Stan. L. Rev.* 60 (2007), 45.

[21] Mary Ellen Mogee. 1991. Using Patent Data for Technology Analysis and Planning. *Research-Technology Management* 34, 4 (1991), 43–49.

[22] Eleonora Pantano, Constantinos-Vasilios Priporas, Stefano Sorace, and Gianpaolo Iazzolino. 2017. Does innovation-orientation lead to retail industry growth? Empirical evidence from patent analysis. *Journal of Retailing and Consumer Services* 34 (2017), 88 – 94.

[23] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research* 12 (2011), 2825–2830.

[24] Radim Řehůřek and Petr Sojka. 2010. Software Framework for Topic Modelling with Large Corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*. ELRA, Valletta, Malta, 45–50.

[25] Skipper Seabold and Josef Perktold. 2010. Statsmodels: Econometric and statistical modeling with python. In *9th Python in Science Conference*.

[26] Patrick Thomas. 2001. A relationship between technology indicators and stock market performance. *Scientometrics* 51, 1 (2001), 319–333.

[27] Subhashini Venugopalan and Varun Rai. 2015. Topic based classification and pattern identification in patents. *Technological Forecasting and Social Change* 94 (2015), 236 – 250.

[28] Bo Wang, Shengbo Liu, Kun Ding, Zeyuan Liu, and Jing Xu. 2014. Identifying technological topics and institution-topic distribution probability for patent competitive intelligence analysis: a case study in LTE technology. *Scientometrics* 101, 1 (01 Oct 2014), 685–704.

[29] Chong Wang and David M. Blei. 2011. Collaborative Topic Modeling for Recommending Scientific Articles. In *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (San Diego, California, USA) *(KDD '11)*. ACM, New York, NY, USA, 448–456.

[30] Yi Zhang, Yue Qian, Ying Huang, Ying Guo, Guangquan Zhang, and Jie Lu. 2017. An entropy-based indicator system for measuring the potential of patents in technological innovation: rejecting moderation. *Scientometrics* 111, 3 (01 Jun 2017), 1925–1946.