Technological Forecasting & Social Change xxx (xxxx) xxx–xxx

Contents lists available at ScienceDirect

Technological Forecasting & Social Change

journal homepage: www.elsevier.com/locate/techfore

Forecasting emerging technologies: A supervised learning approach through patent analysis

Moses Ntanda Kyebambe[a], Ge Cheng[b,*], Yunqing Huang[a], Chunhui He[a], Zhenyu Zhang[a]

[a] Department of Mathematics and Computational Science, Xiangtan University, Xiangtan, Hunan, China
[b] College of Information Engineering, Xiangtan University, Xiangtan City, Hunan Province, China

ARTICLE INFO

Keywords:
Technology forecasting
Industrial technology roadmap
R & D planning
Patent analysis
Citation analysis

ABSTRACT

Both private and public enterprises have great interest in prior knowledge of emerging technologies to enable them make strategic investments. Technology forecasting offers a relevant opportunity in this direction and is currently a hot upcoming area of research. However, accurate forecasting of emerging technologies is still problematic mainly due to absence labeled historical data to use in training of learners. Previous studies have approached the technological forecasting problem through unsupervised learning methods and, as such, are missing out on potential benefits of supervised learning approaches such as full automation. In this study, we propose a novel algorithm to automatically label data and then use the labeled data to train learners to forecast emerging technologies. As a case study, we used patent citation data provided by the United States Patent and Trademark Office to test and evaluate the proposed algorithm. The algorithm uses advanced patent citation techniques to derive useful predictors from patent citation data with a result of forecasting new technologies at least a year before they emerge. Our evaluation reveals that our proposed algorithm can retrieve as high as 70% of emerging technologies in a given year with high precision.

1. Introduction

Due to a very fast pace at which technology is evolving, enterprises are faced with a hard decision of the best and most suitable technology to invest in. In this paper, we propose an algorithm for predicting emerging technologies to support enterprises to make data driven decisions over which technologies to invest in. The model is designed to detect signals of a technology likely to cause a significant disruption in an industry at least a year before the technology fully emerges. This way, the model has a potential of reducing the risk of being late to adopt a technology by an enterprise. Automatic forecasting of technologies remains a difficult task largely due to scarcity of labeled data to train reliable classifiers; traditional approaches have relied on unsupervised learning methods. Patent databases offer a huge source of technological inventions data that many researchers have exploited with unsupervised learning methods to forecast technologies mostly relying on citations. In the context of patent studies, a citation is reference to previous work (also known as prior art) that is relevant to the current patent application. For a specific granted patent, the patents it cites are known as backward citations while future patents that cite it are known as forward citations. All methods that base on forward citations to forecast technological trends suffer one major limitation of a large time lag between the date a patent is published to the date it begins attracting citations. In this study, rather than relying on forward citations which take long to build, we use backward citations to derive several features most capable of discriminating a high-impact patent of technology likely to disrupt business in a given industry from patents of just incremental technology. Furthermore, we propose an algorithm for labeling emerging technology patent clusters based on new classes progressively established in the United States Patent Classification (USPC) system overtime. Besides the USPC, the proposed method is extensible to make use of other data sources such as blogs, conferences and social networks in labeling emerging technology patent clusters.

This study is part of a growing number of studies (Érdi et al., 2013; Fleming et al., 2006; Karvonen and Kässi, 2013; Sorenson et al., 2006) that have employed patent citation analysis in predictive analytics, particularly technological trends.

2. Literature review

Use of citations in analytics dates far back in the 1970s with Garfield's (Garfield, 1979) extensive article on citation index theory and its application to patent literature analysis, scientific journal analysis and many other areas. The Science Citation Index is indeed still widely

* Corresponding author.
E-mail addresses: mntanda@cis.mak.ac.ug (M.N. Kyebambe), chengge@xtu.edu.cn (G. Cheng), huangyq@xtu.edu.cn (Y. Huang), zhenyuzhang@smail.xtu.edu.cn (Z. Zhang).

http://dx.doi.org/10.1016/j.techfore.2017.08.002
Received 22 November 2016; Received in revised form 15 July 2017; Accepted 1 August 2017

used in valuation of scientific literature. Citations do not only occur in research publications but also in patents though for a different purpose. However, in both cases a citation indicates some relationship between the citing document and the cited document. Within patent literature, citations have recently been used to analyze technological evolution (Wong and Wang, 2015) as well as forecasting new technologies (Breitzman and Thomas, 2015). New technologies are believed to be a blend of different components of preceding technologies thus studies seeking to analyze evolution of technologies as well as forecasting emerging technologies usually make use of patent citations to link different generations of technology. Many earlier studies used forward patent citations (citations a patent receives from later patents) in one way or another to create patent citation networks for purposes of technological road mapping and forecasting (Albert et al., 1991; Érdi et al., 2013; Fleming, 2001; Seung-wook et al., 2014).

In Érdi et al. (2013), the researchers developed a model in which emergence of new technologies was detected by emergence of new clusters within a patent citation network. They constructed a patent citation network where each node in the network is a patent vector constructed by calculating the sum of citations received by the patent from patents in 36 selected technological areas. By taking patent citation graph snapshots at different time series, they were able to detect emergence of a new technological area well before the USPTO identified it and later established a class for it. However, critics (Rotolo et al., 2015) of their method argue that science and technology are fast-evolving such that subsequent annual networks are likely to have a very high percentage of suitably emerging clusters. In addition, their method is based on forward citations yet patents take a considerable amount of time before beginning to attract citations.

Co-citation analysis pioneered in the early 1970s (Small, 1973) is closely related to citation analysis and is geared towards producing co-citation networks. It assumes that documents cited together frequently cover closely related subject matter. It has been used in several predictive analytics studies (Blondel et al., 2008; Chen et al., 2010; Ittipanuvat et al., 2014; Lai and Wu, 2005; Shibata et al., 2008; Shibata et al., 2010).

However, studies based on forward citations suffer from one major limitation of a time lag between the publication of a document and the time it begins attracting citations. The cumulative advantage (De Solla Price, 1965), identical to the “rich get richer” aphorism, of old literature over recent literature has led to the recent insurgence of studies such as Breitzman and Thomas (2015) that seek to overcome the problem. In patent literature, a patent averagely takes at least two years to start attracting citations thus forward citations are not very efficient in predicting emerging technologies in real time. Although some studies (Valverde, 2014) have indicated that the cumulative advantage doesn't last forever, its effects in the short run render forward citations impractical for real time forecasting.

As an alternative, studies have started exploring the use of backward citations which are available as soon as a patent is published. Breitzman and Thomas (2015) identify a patent likely to contain an emerging technology by considering its linkage to prior “hot” patents through backward citations. Their method overcomes the citation time lag limitation suffered by methods based on citations received by a patent. However, as acknowledged by the authors themselves, the model highly depends on the now defunct National Institute of Standards and Technology's Advanced Technology Program (ATP) to identify emerging technology clusters. Moreover, the method fails to detect pioneer emerging technologies that have no linkages to underlying hot patents. Most related to our work here is the use of the current patents' linkage to previously published patents to forecast emerging technologies. However, the algorithm we use to link current patents to prior “hot” patents significantly differs from that used in their study (Breitzman and Thomas, 2015).

Bibliographic coupling, based on backward citations, measures the extent to which two documents cite the same set of documents (Kessler, 1963). It is somehow similar to co-citation since documents that most frequently cite the same other set of documents are likely to be related. In our study, we made use of bibliographic coupling to group related documents into clusters. Since prediction of emerging technologies is sensitive to time at which a prediction is done bibliographic coupling overcomes the time lag limitation suffered by other methods discussed before. Other alternatives to counter the cumulative advantage problem include using a citation index (Breitzman and Thomas, 2015), and application of textual analysis instead of citations to link patents (Smalheiser, 2001; Swanson, 1987; Tseng et al., 2007) among others. Other limitations of studies based on patent citations have been revealed such as some applicants strategically withholding citations to related prior art (Lampe, 2012) so as to avoid invalidation of their inventions and companies strategically refusing to seek patents for their technologies and rather conceal the technologies as trade secrets. Although earlier studies (Klavans and Boyack, 2015; Shibata et al., 2009) favored direct citation to bibliographic coupling in detecting emerging research front, their findings cannot be applied to this study because their evaluations were based on detecting research fronts as they emerge rather than forecasting. Since new technologies usually emerge from a blend of recent technologies, we believe that using bibliographic coupling which was found to be most accurate (Boyack and Klavans, 2010) for short window periods gives our method the best performance.

Besides citations, other ways have been explored such as using subject-action-object (SAO) structures (Park et al., 2013; Yoon and Kim, 2011), a combination of objects (companies, inventors, and technical content) (Tang et al., 2012) to construct patent networks. Fleming and Sorenson (2004) used patent citation data to explore the value of using science to guide innovation by tracking the number of patent citations to non-patent sources, and measures the difficulty of an invention by looking at how subclasses related to the patent were previously combined.

3. Materials and methods

For technology forecasting to be beneficial to enterprises, forecasts need to be made at least a year ahead to enable enterprises make informed adjustments in their budget allocations; our method aims at achieving this goal. We hypothesize that given historical data of emerged technologies, we can derive trends that allow us to forecast future technologies. We achieve this through a series of steps. Traces of emerging technologies are usually traceable from patent databases a few years prior to full emergence. New technologies are usually not confined in a single patent but rather a cluster of patents. Therefore at a given point back in time, we study features possessed by a cluster of patents that later gave birth to a new technology. Using these features, we train our model to forecast technologies before they emerge. The major steps of our methodology are: (1) Take a step back in time and identify technologies that emerged (2) Take a further step back and identify clusters of patents from which the technologies identified above emerged (3) Study features possessed by patent clusters identified in (2) above. (4) Build a model based on these features and use the model to forecast emerging technologies. A detailed discussion of how we performed the above steps follows below.

3.1. Datasets

Besides patent databases, there are several other sources of data for forecasting technologies, for example: conference proceedings (Furukawa et al., 2014), social networks analysis and so on. However, patent databases provide a cheaper source since they are freely publicly available and the documents are in a well formatted structure which makes processing relatively easier. Moreover, patent database are maintained and annotated by highly experienced domain experts. The US patent database is one of the earliest and most organized patent databases in the world and for this reason we chose it as our source of