Technological Forecasting & Social Change 125 (2017) 236–244

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
Available online 15 August 2017
0040-1625/

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

data. Moreover, the US Patent database has a richer source of citations compared to other patent databases such as European patent database; On average, US patents cite over three times as many patent references and non-patent references compared to European patents (Michel and Bettels, 2001). Furthermore US patent database is usually updated as new technologies emerge thus introducing new classes and/or subclasses to accommodate the emerged technology. This makes it possible to utilize data labeled by patenting officers to forecast emergence of new technologies. The USPTO website maintains a list[1] of classes and the time at which they were established. Creation of a new class indicates emergency of a new technology unique enough from existing technologies to merit a class of its own. Furthermore, to enable us deal with a relatively manageable scope, we only dealt with utility patents granted between 1979 and 2010 downloaded from Reed Tech website.[2] Reed Tech entered into an agreement with USPTO to make patent and trademark bulk data available to anyone free of charge.

3.2. Pre-processing the data

We downloaded all the required patents and indexed them for faster access. We extracted features from each patent which are likely to discriminate between emerging technologies and seasoned technologies. This was the most crucial part of our research since accuracy of predictions made greatly depends on relevance of these features to emerging technologies. Earlier studies attempted to define features likely to be possessed by emerging technologies and choice of the features we discuss below was guided by research findings of these studies. For each patent, we extracted or derived the following features.

3.2.1. Number of claims

Break through patents have been found to have a higher renewal rate and higher number of claims (Moore, 2004). Given a very important new invention, inventors usually tend to come up with several claims so as to safeguard their interest as much as they can.

3.2.2. Number of citations

Inventors and examiners cite other patents to show dependency of the citing technology to the cited technology (Newman, 2010) as well as to pre-empt any later claims that the cited technology invalidates patentability of the citing technology. For the latter reason, the inventor usually presents a strong argument for the invention to merit its own patent. In both cases, a high number of citations are likely to mean a greater dependency or relationship between the citing technology and a big array of other technologies. For these reasons, patents that represent key turning points in technology are likely to have more prior art references than those that present just incremental advances to existing technologies. As such, a patent likely to be a turning point in a given industry is likely to be distinguished from other patents by the number of citations it receives.

3.2.3. Number of citations made to non-patent literature

Many new inventions are implementations of scientific research findings. Economically significant inventions are usually a result of considerable investment in scientific research mostly by government or large firms; as such, new inventions usually have a strong connection to non-patent literature. This feature is closely related to the Originality and Science indices used by (Breitzman and Thomas, 2015) to score patent clusters that are likely to contain new inventions.

3.2.4. Technology Cycle Time (TCT)

TCT is the mean age (in years) of patents cited on the front page of a patent. It is a measure of how fast technology in a given technological area changes; a small value of TCT indicates fast changing of technological generations within an area. This measure is particularly important if normalized across different technological areas since studies (Park et al., 2006) have shown TCT to differ from area to area while also varying from time to time within the same technological area (Kayal and Waters, 1999). Within a fast growing technological field, we expect patents containing emerging technologies to have relatively smaller TCT values compared to those merely incrementing on existing technologies. TCT is computed as shown in Eq. (1);

$$TCT_i = median_j\{|\ T_i - T_j\ |\} \tag{1}$$

where T_i is the application date of patent i and patents i and j are connected.

3.2.5. Patent class

A patent classification comprises of a main class and a subclass. The features we have discussed and those we are yet to discuss vary significantly across patent main classes, for example a TCT value of 10 years is normal for an emerging technology in the field of natural sciences but too high in fast evolving fields such as electronics and computing. We include patent class among the features to harmonize differences among other features across patent classes.

3.2.6. Cited Technology Similarity Index (CTSI)

This index is a measure of the similarity between the technological area of the citing patent and the technological area of cited patents. Novel technology patents tend to cite many other patents outside their own core technology. As such, earlier studies (Érdi et al., 2013) have used measures similar to CTSI to forecast emerging technologies by measuring the impact a patent has had on other technological areas outside its own technology. However, unlike these researchers who used forward citations received by a patent from patents in other technological areas, we use backwards citations to measure the similarity between the technological area of the citing patent and the technological area of the cited patents. We chose backwards citations because they are available as soon as the patent is published thus enabling us to make meaningful predictions on time. The US Patent Classification System (USPC) assigns patents into Classes and Subclasses. As of this writing, there are over 450 Classes and over 5600 Subclasses in the USPC. The Class reflects a major technology area, for example Electrical computers and digital processing systems: memory (Class 711) in which Hierarchical memories (Subclass 117) is one of the Subclasses. The Subclass defines finer details of the more general Class. A single Patent has both a Class and a Subclass assignment thus two patents may share a Class but with different Subclasses. In this article, we define a Classification as a combination of the Patent Class and Subclass. We compute CTSI of a patent basing on its classification and classifications of patents it cites. First, we define Classification Similarity between two Classifications i and j (CS_{ij}) as

$$CS_{ij} = \begin{cases} 0 \text{ if } i \text{ and } j \text{ differ in Class and Subclass} \\ 0.5 \text{ if } i \text{ and } j \text{ differ in Sublass but not Class} \\ 1 \text{ if } i \text{ and } j \text{ share the same Class and Subclass} \end{cases} \tag{2}$$

We then define Patent Classification Similarity between two Patents p and q (PCS_{pq}) basing on similarity of their Classifications. It should be noted that in the USPC, a Patent may belong to one or more Classifications.

$$PCS_{pq} = \frac{\sum_{i=1}^{N_p}\sum_{j=1}^{N_q} CS_{ij}}{N_p N_q} \tag{3}$$

where N_p and N_q are the numbers of Classifications of Patents p and q respectively.

Finally, we compute CTSI of a Patent, p, basing on the average similarity between its Classifications and Classifications of Patents it cites

[1] http://www.uspto.gov/page/classes-us-patent-classification-system-dates-established.

[2] http://patents.reedtech.com/pgrbft.php.

as in Eq. (4)

$$CTSI(p) = \frac{\sum_{q=1}^{N} PCS_{pq}}{N} \tag{4}$$

where N is the total number of patents cited by p (out-degree of p) and q is the qth patent cited by p.

3.2.7. Cited patents Assignee Similarity Index (CASI)

When a company makes a landmark new invention, it tends to apply for many patents around this invention (Albert et al., 1991) as a way of guarding against any intrusion from its competitors. As such, emerging technologies cite several other patents that belong to the same company thus landmark inventions are usually characterized by a large number of self-citations within assignees. However, usefulness of CASI is limited when dealing with small startup companies that may have very few important patents. Our derived feature *CASI* measures the similarity between assignees of given patent, X, and assignees of patents cited by X. A US patent may have one or more assignees. We compute Assignee Similarity (*AS*) between two patents, p and q using Eq. (5):

$$AS_{pq} = \frac{\sum_{i=1}^{N_p} \sum_{j=1}^{N_q} ASG_i ASG_j}{N_p N_q} \tag{5}$$

where N_p *and* N_q are the numbers of Assignees of patents p and q respectively, ASG_i is Assignee number i of patent p, ASG_j is Assignee number j of patent q, and

$$ASG_i ASG_j = \begin{cases} 1 \; if \; ASG_i \; and \; ASG_j \; are \; the \; same \\ 0 \; otherwise \end{cases} \tag{6}$$

Since number of assignees vary across patents, we normalize the value of AS_{pq} in by dividing by N_p and N_q. This is way, the index will not be unfair to patents with a small number of assignees.

We use *AS* to compute *CASI* of a patent, q, using as in Eq. (7):

$$CASI(q) = \frac{\sum_{q=1}^{N} AS_{pq}}{N} \tag{7}$$

where N is the total number of patents cited by p and q is a member of patents cited by p.

3.3. Processing patent cluster vectors

We use vectors of patent clusters to train our prediction model. From patent documents downloaded as explained in Section 3.1, we created a Lucene[3] index and used it to create a feature vector for each of the patents where each dimension is one of the features introduced in Section 3.2. We shall refer to this vector as the Patent Feature Vector (PFV) in the subsequent discussion. Our ultimate goal is to be able to train a model capable of labeling patent clusters in a given subject year as either constituting an Emerging Technology (ET) or Non Emerging Technology (NET). So, it is logical at this point to cluster all the Patents in each year using the PFV since elements of the PFV are capable of discriminating between Patents that comprise of emerging technologies and those that do not. Using historical data of technologies that emerged in past years, we can label the clusters and use the labeled clusters to train our model. The next decision to make is choice of the number of clusters for each year, since this number is not available at hand. Progress in algorithms that automatically determine the number of clusters is still slow and for most studies this number is obtained heuristically through trials. Considering the large size of our datasets and the need for a fast and efficient clustering algorithm, we started by using DbScan algorithm (Ester et al., 1996) to obtain a rough number of expected clusters then used K-means with K slightly greater than the number given by DbScan to obtain the final clusters. As we later explain, using a value of K slightly greater than the theoretical value does not affect our results since any splinter clusters would finally be re-linked together.

The result of the clustering process is a set of patent clusters for each year in the period under the study: some of the clusters will constitute patents comprising of emerging technologies while the majority will constitute patents comprising non-emerging technologies.

3.4. Labeling PFV clusters

The next step, which is our major contribution, is to label each of the patent clusters as either ET or NET by using a series of algorithms. Emerging technologies must show much coherence and persistence over some period of time so as to qualify as emerging. It is important for a technology to gain a relative state of stability before it can be labeled emerging since many new technologies show some features of ET but fail to gain stability leading to their exit before truly emerging. Creation of a new class, for a given technology, in established classification systems demonstrates persistence and has been used (Rotolo et al., 2015) to measure coherence. As such, we label a cluster of patents as ET if it shows evidence of possessing features that make it ripe to give rise to a new technological class in the near future. For a given subject year T, we label a cluster as an ET cluster if it is linked to a cluster of patents whose main class was established class established in year $T+1$. Intuitively, this means that the ET cluster in year T possessed features characteristic of a cluster likely to give birth to a new class/technology the following year. The cluster-labeling algorithm is shown below.

Algorithm 1. Cluster-labeling algorithm.

```
T ← start year
while T < end year
    G ← Set of patent groups granted in year T + 1 and grouped by main class
    C ← Set of patents clusters granted in year T and clustered basing on PFV
    foreach cluster of patents c in C
        foreach group of patents g in G
            BCS_gc ← Bibliographic Coupling similarity between c and g      // See equation 8.
        end foreach
        link c to g if their value of BCS_gc is the highest
        if the main class of g to which c has been linked was established in
        year T + 1 label c as ET cluster else label c as NET cluster
    end foreach
    T ← T + 1
end while
```

The result of the above algorithm is a set of PFV clusters of year T labeled as either ET or NET based on whether it's linked to a classification that emerged in year $T+1$ or not. This perfectly serves our objective since at this point in time, in year T, we are able to determine which patent clusters are likely to give rise to a new technology in the near future, year $T+1$. However, up to this point there is still one unanswered question of how to compute the Bibliographic Coupling Similarity between two patent clusters, which we answer right now:

Two documents are bibliographically coupled if they cite the same document. It is closely related to co-citation where two documents are said to be co-cited if they are cited by the same document. However, unlike co-citation, a bibliographic coupling value can be computed as soon as the two documents are published. Both bibliographic coupling and co-citation indicate a relationship between the two documents and their value may be either Boolean, in which case true would indicate presence of bibliographic coupling, or a number in which case the value would indicate number of common documents cited by two documents. In this study, we extended the idea of bibliographic coupling to apply to patent clusters thus we compute the Bibliographic Coupling Similarity (BCS) of two patent clusters *X and Y* by counting the number of unique common citations of patents in clusters *X and Y*. However, merely counting the number of common citations would be unfair to clusters with few patents thus we divide this number by the total number of

[3] https://lucene.apache.org/.

Table 1
Number of ET and NET and their associated attributes.

Year	# Clusters		# Classes established in subsequent year	Mean cluster size		Mean 5-year citation index		Mean probability of a patent in the cluster being key	
	ET	NET		ET	NET	ET	NET	ET	NET
1987	6	490	5	131	149	1.41	0.75	0.31	0.14
1988	6	492	4	120	142	1.45	0.75	0.32	0.12
1989	1	491	5	68	177	1.43	0.75	0.32	0.12
1990	10	495	8	140	168	1.38	0.79	0.31	0.12
1991	0	484	10	–	180	–	0.71	–	0.11
1992	12	493	9	150	185	1.53	0.76	0.34	0.12
1993	20	479	8	157	188	1.50	0.8	0.33	0.12
1994	5	476	2	143	196	1.51	0.76	0.36	0.12
1995	9	486	7	170	194	1.50	0.75	0.33	0.11
1996	19	480	7	173	213	1.47	0.73	0.33	0.14
1997	0	468	3	–	150	–	0.82	–	0.13
1998	19	493	7	260	281	1.41	0.78	0.31	0.13
1999	0	463	4	–	295	1.59	0.76	0.34	0.12
2000	0	486	0	–	304	1.50	0.76	0.33	0.12
2001	2	483	1	282	314	1.57	0.75	0.34	0.12
2002	4	475	3	314	330	1.58	0.7	0.34	0.11
2003	0	468	2	–	331	–	0.84	–	0.13
2004	0	482	2	–	318	–	0.82	–	0.14
2005	0	475	0	–	283	–	0.82	–	0.13
2006	0	473	1	–	341	–	0.81	–	0.10
2007	0	475	0	–	308	–	0.82	–	0.16
2008	1	478	1	235	305	1.65	0.87	0.35	0.13

unique citations in X and Y as in Eq. (8).

$$BCS_{XY} = \frac{n(C_x \cap C_y)}{n(C_x \cup C_y)} \tag{8}$$

where BCS_{XY} is the BCS value between clusters X *and* Y, C_x and C_y are sets of patents referenced by patents in clusters X *and* Y respectively.

It should be noted at this point that our PFV clusters comprise of features of individual patents. However, we are more interested in characteristic features of a PFV cluster as group rather than individual patents. As such, we derive features pertaining to the whole PFV cluster from constituent patents and later use these group features to train our model. For a given PFV cluster, we compute the group feature value by taking the arithmetic mean of values for the constituent patents. For example, if a PFV cluster comprises of 10 patent documents, we compute the overall CASI for the group as the average CASI of all the 10 constituent patents. Values for other numeric features are derived similarly. There is only one non-numeric feature, which is the patent class, for which we take the cluster value to be majority (mode) class of the cluster. In case of a tie where two or more classes constitute the majority, we move down to the subclass level and assign to the cluster the class of the mode classification. As noted earlier, a classification is composed of a class and a subclass, so we look for the majority classification within the cluster and assign to the cluster the class of this majority classification. The result of this step is a set of patent-cluster vectors each labeled as either ET or NET. We use these patent-cluster vectors to train and evaluate our model. In the next section, we briefly explain the steps taken to evaluate our model.

3.5. Evaluation

We evaluate the method by testing its ability to predict ET clusters before the USPTO defines new technology classes from these clusters. This method of model evaluation has previously been used before by other researchers (Érdi et al., 2013) and gives a good indication as to whether the model has achieved its objective. Though percentage classification accuracy has been criticized to be a flawed measure of performance especially with skewed datasets, it still gives a good measure of performance when augmented with other measures such as Precision and Recall. Since our major interest in predicting emerging technology is to retrieve as many ET clusters as possible (high Recall) with an averagely high precision, we also use Recall, Precision and F-measure to evaluate our method. Furthermore, F-Measure provides a combined effect of both Recall and Precision and is commonly known as a weighted harmonic mean of Precision and Recall. It is computed as in Eq. (9):

$$F = 2^{*}\frac{precision^{*}recall}{precision + recall} \tag{9}$$

4. Results and discussion

The ultimate objective of the cluster-labeling algorithm is to label clusters of patents as either *ET* or *NET*. As discussed before, a cluster in year T is labeled "*ET*" if it is linked to a class of technology that emerged in year T + 1. Ability of our algorithm to cluster similar technologies has potential value particularly if the cluster-labeling algorithm is able to accurately discriminate clusters that comprise high impact technologies from others. As such, our concern is to verify whether clusters labeled as ET indeed comprise of key patents likely to cause disruptions in their respective areas. We make this verification in two ways: (1) By comparing the impact patents in ET and NET clusters make on patents issued in the near future. We would expect patents in ET clusters to have a higher impact on patents issued in the near future than patents in NET clusters. (2) By performing text analysis on ET clusters to determine if there is a similarity between technologies embedded in the clusters and technologies that emerged in the same year. Table 1 shows the number of ET and NET clusters in each of the years considered in this study, together with other attributes relevant to validation of the proposed approach.

To assess the impact of a given cluster of patents on the near future, Breitzman and Thomas (2015) introduced a very interesting metric they called mean citation index. Citation index of a patent is derived using the number of forward citations a patent receives from patents issued in a given period of time after its issuance. Since patents usually take at least two years to start attracting citations, 5 years is a fair period to compute the citation index and was thus used in this study. To avoid the effect of age and variation of citation counts across technological areas, we derived citation index by dividing the number of citations received by a patent by the mean number of citations received by all patents issued in the same year and belonging to the same main class. For

example, to compute the citation index of a patent belonging to class 348 (Television class) issued in 2005, we divide the number of citations received by the patent between 2005 and 2010 inclusive with the mean number of citations received in the same period by all patents issued 2005 and belonging to class 348. The expected citation index of an individual patent is therefore one; a citation index greater than one implies the patent is cited more frequently than expected for example a value of 1.8 would imply the patent has been cited 80% more frequently than expected while 0.6 would mean 40% less frequently cited than expected. In Table 1 we show the mean citation index for patents in ET and NET clusters for a period of 5 years.

Table 1 exhibits both expected and abnormal trends worth explaining. We observe a high positive correlation between the number of ET clusters detected in a subject year T and new classes established in the USPC in the subsequent year. Since our method traces the evolution of an emerging technology starting from the time the technology is mature enough to be recognized by the USPC, we expect the number of ET clusters detected to be nearly equal to the number of new classes that will evolve in the subsequent year. However, some abnormalities exist: There are some years, for example 1989, where the number of ET clusters is much less than the number of classes established in the subsequent year. Explanation for this abnormality is two-fold: (1) several of the new classes in the subsequent year could have evolved from a single ET cluster thus a one-to-many mapping from a set of ET clusters to several sets of new classes in the subsequent year, and (2) failure of our cluster-labeling algorithm to detect some ET clusters owing to a limited number of patents belonging to classes established in the subsequent year. For example in 2006, our method hardly detected any ET cluster whereas there was a new class (506) established in 2007. However, a deeper look into class 506 in the dataset we downloaded reveals that although this class was established in 2007, no single patent belonged to this class in the same year. On the other hand there are some cases, for example in 1993, where the number of ET clusters by far outnumbers the number of classes established in the subsequent year. This is a typical phenomenon of technological convergence where different technologies fuse and evolve into a single new technology thus giving rise to a many-to-one mapping from several sets of ET clusters to a single new class set in the subsequent year.

We also note that the number of ET clusters seems too small compared to the number of technologies that possibly emerge each year. This is expected since our method uses new classes established in the USPC to find traces of emerging technologies from years prior to the date of establishment of the class. A class in the USPC encompasses several subclasses of technology implying that a single ET cluster could actually contain several technologies that emerged at that time. This is further supported by our results which reveal that an average ET cluster comprises of about 180 patents related by class with, possibly, several new technologies encapsulated in subclasses.

From Table 1, we also note a general steady increase in the mean size of both ET and NET clusters from 1987 to 2008, the most recent year examined in this study. This implies that the rate of innovation has increased in the recent years and this resonates well with the increase in number of patents granted since 1987. Since a large ET cluster is likely to contain a higher number of new technologies than a smaller one, this further suggests that the rate of emergency of new technologies has progressively increased since 1987. Generally, an NET cluster size is much bigger than that of an ET cluster for the same subject year and this echoes findings of Breitzman and Thomas (2015), Érdi et al. (2013), and Rotolo et al. (2015) that most of the patents are just incremental improvements to existing technology with radical new innovations taking a minority share. Data from Table 1 reveals that the average size of an ET cluster is 180 patents whereas that of an NET cluster is 243 patents.

The mean 5-year citation index in the right half of Table 1 reveals that patents in ET clusters generally have a higher citation index than their NET counterparts in each year. As explained before, this implies that patents in ET clusters impact more strongly on the direction taken by technologies in the subsequent five years than their NET counterparts do. This is an interesting and important result since it validates our claim that the cluster-labeling algorithm accurately labels the clusters. It is important to note here that the high mean citation values for ET are not as a result of ET clusters being favored by their small size since citation index is computed at the level of an individual patent rather than cluster level. The results reveal, for example, that in 2008, an ET patent is on average cited 65% more frequently than expected while its NET counter is 13% less frequently than expected cited. The same trend occurs in all the years considered in this study.

Although the mean citation index has a high potential of discerning between an ET and an NET as well as showing relevance of the cluster to technologies in subsequent years, it also has potential flaws since it may fail to detect clusters that are missing characteristics of a true ET cluster. We believe a true ET cluster must possess at least a few key patents. We define a key patent as a patent with an extra ordinary high citation index. Within any technological area, patents that define turning points are usually central and, as such, receive a significant number of citations. Basing on mean citation index, we may falsely label a cluster as ET if it has a high enough proportion of patents with citation index slightly above expected but missing key patents. To define a key patent, therefore, we must set a threshold for citation index above which the patent is regarded key. For this study, we empirically set the threshold to 2.5 basing on mean citation index data from Table 1. For each cluster in a given subject year, we computed the probability of a patent in the cluster being key, by dividing the number of patents with citation index greater than or equal to the threshold by the total number of patents in the cluster; the results are in shown in the rightmost part of Table 1. Table 1 clearly shows that a higher mean citation index does not necessarily indicate a higher presence of key patents in the cluster, for example: the mean citation index of patents in ET clusters in 1992 (1.53) is higher than the corresponding citation index in 1994 (1.51). However, the probability of a 1994 ET cluster patent being key (0.36) is higher than the corresponding probability in 1992 (0.34). This implies that although 1992 ET clusters have a higher mean citation index than corresponding 1994 clusters, the average percentage of key patents per cluster, with citation index greater than or equal to 2.5, in 1994 (36%) is greater than the corresponding value in 1992 (34%). From Table 1, it is also evident that ET patents have relatively higher probabilities of being key compared to NET patents. Although the probabilities seem somewhat low, this is expected because key patents are usually very rare yet the threshold we set was very stringent.

Although mean citation index coupled with probabilities of patents within clusters being key are sufficient to prove that ET clusters comprise of high impact patents, we took a further step to verify whether patents in ET clusters have linkages to technologies that emerged in subsequent years. We computed the Term Frequency – Inverse Document Frequency (TF-IDF) of each cluster and then extracted the top 100 terms. For each cluster, we created a vector of TF-IDF values of the extracted terms. Furthermore, we examined new technologies that emerged during the years in the period of study and created a new-technology vector whose elements were the top-100 keywords that define that technology. We then computed the cosine similarity between each ET cluster in each year and each of the new-technology vectors in the subsequent years; if any of the cosine similarity value exceeded a subjectively set threshold of 0.8, we concluded that the cluster under examination truly comprises of high impact patents. In this research, we examined a random sample of 5 ET clusters and Table 2 shows the cosine similarity value of each of the chosen clusters and the most similar new technology that emerged around the same time. From the table, we realize that 80% of clusters labeled as ET have over 0.8 cosine similarity with new technology that emerged around the same time. This adds validation that the cluster-labeling algorithm was accurate.

Table 2
Cosine similarity values between emerging clusters and technology emerged at the time.

ET cluster name	ET cluster year	Top similar emerging technology	Cosine similarity
Cluster336–1988	1988	World Wide Web	0.93
Cluster387–1990	1990	Digital TV	0.82
Cluster462–1992	1992	DVD	0.89
Cluster93–1994	1994	USB	0.72
Cluster303–2002	2002	Flat Screens & HDTV	0.85

Table 3
Classification results by classifier. A represents Emerging Technologies Category while B represents Non-Emerging Technologies.

	NB	ANN	LR	SVM	RF	RT
% Correct	59.0	54.2	57.0	70.6	56.6	60.0
Recall - A	0.53	0.59	0.43	0.74	0.59	0.51
Recall - B	0.64	0.50	0.69	0.68	0.55	0.67
Precision - A	0.54	0.50	0.54	0.70	0.52	0.57
Precision - B	0.62	0.59	0.59	0.72	0.61	0.62
F-Measure - A	0.53	0.54	0.48	0.72	0.55	0.54
F-Measure - B	0.63	0.54	0.64	0.70	0.58	0.64

We used five classification algorithms – Naïve Bayes (NB), Artificial Neural Networks (ANN), Logistic Regression (LR), Support Vector Machine (SVM), Random Forest (RF), and Random Tree (RT) – to evaluate the accuracy and precision of the proposed approach in retrieving emerging technologies. From Table 1, it is apparent that we are dealing with a biased dataset with NET clusters far outnumbering ET clusters. To counter the effect of the skewed dataset, we used stratified tenfold cross-validation to train and evaluate each of the classifiers in Table 3. The results in Table 3 show that there are tradeoffs to be made when considering which classifier to use.

From Table 3, we realize that the SVM algorithm outperformed all the other algorithms with an overall accuracy of 70.6%. In addition, SVM gave relatively higher Precision and Recall (0.70 and 0.74 respectively for category A) compared to other algorithms. These are promising results since they give us confidence that our method has ability to retrieve 74% of the Emerging technologies (category A) in a given year with 70% precision. Besides SVM, the Random Tree also performed averagely well followed by Naïve Bayes, and Logistic Regression respectively.

However, besides SVM which outperformed all the other classifiers, there is no other single classifier among those examined in this study that is better than all the others in all the metrics used in this study. For example, whereas the Random Tree has a higher overall prediction accuracy and higher recall for category B (NET) than Random Forest, the Random Forest outperforms it with a higher recall and higher F measure for category A (ET). A similar analysis reveals that each classification algorithm at least has a metric where it outperforms its counterpart besides SVM.

Although somewhat basic, classification accuracy in Table 3 gives an overview of the overall performance of our method with complete interaction of all the features. The question of whether performance is a function of some key feature or interaction between features still remains. We would be interested in assessing the performance of the method when specific features are dropped and this interest cannot be served by existing feature selection and dimension reduction methods since most of them use some criteria, such as information gain, to evaluate the worth of a feature with respect to the class before finally ranking features by their worthiness. Such methods therefore assess the worth of a single feature in isolation rather than a subset of features. We performed exhaustive testing to evaluate the effect, on classification accuracy, of dropping a subset of features; this resulted into 127 test cases of each classification algorithm and a total of 762 test cases for all

Table 4
Classification accuracy when a subset of features is used.

SVM									ANN								
Feature subset used									Feature subset used								
Patent class	TCT	NumCitations	NonPatCitations	CASI	NumClaims	CTSI	Accuracy		Patent class	TCT	NumCitations	NonPatCitations	CASI	NumClaims	CTSI	Accuracy	
●	●	●	●	●	●	●	70.6		●	●	●	●	●	●	●	54.2	
●	●	●	●	●	●	●	65.6		●	●	●	●	●	●	●	51.6	
●	●	●	●	●	●	●	64.8		●	●	●	●	●	●	●	50.4	
●	●	●	●	●	●		63.2		●	●	●	●	●	●		48.3	
●	●	●		●	●	●	61.3		●	●	●		●	●	●	46.6	
●	●		●	●	●	●	58.7		●	●		●	●	●	●	43.8	
●	●	●	●	●	●	●	55.1		●	●	●	●	●	●	●	40.3	
●	●	●	●	●	●	●	54.6		●	●	●	●	●	●	●	39.8	
●	●		●	●	●	●	49.6		●	●		●	●	●	●	34.8	
●	●	●	●	●	●	●	23.4		●	●	●	●	●	●	●	20.8	
●	●	●	●				22.5		●	●	●	●				15.3	
●	●		●				22.4		●	●		●				12.6	
●	●	●	●	●			21.8		●	●	●	●	●			14.4	
●	●		●	●	●		19.2		●	●		●	●	●		9.5	
●	●	●	●	●	●	●	15.6		●	●	●	●	●	●	●	8.2	

the classification algorithms. All the algorithms used in this study showed almost similar effects on accuracy when subsets of the features were dropped from the feature set. For brevity, we present results for only SVM, since it was the best overall algorithm as per our data, and ANN because it paints the best generalized picture of the remaining algorithms. The top nine best performing feature subsets and the worst six are shown in Table 4 when the SVM and ANN algorithms were used to forecast ET clusters. For both algorithms, best performance occurs when a complete set of the features is used. This suggests that optimum performance of our method is achieved through interaction of all the features. However, there are some feature subsets that show relatively good performance. As expected, a single feature does not possess enough discriminating information to discern between an ET cluster and an NET cluster. As such, worst performance occurs when only a single feature is used. From the table, worst classification results are observed when CTSI is used with SVM and ANN achieving as low as 15.6% and 8.2% accuracy respectively. However, this does not seem to imply that CTSI is the least significant feature since dropping CTSI significantly degrades accuracy of the SVM by 7.4% (from 70.6 to 63.2) compared to just a 5% decrease when the NumCitations feature is dropped. Similarly, as seen from the right half of Table 4, accuracy of the ANN algorithm drops by 5.9% when the CTSI feature is discarded compared to a 2.6% decrease when NumCitations feature is discarded. It is also worth noting that accuracy of both algorithms reduces with decreasing subset size and, as such, the top ten best performing feature subset list is dominated by subsets with six features. There are two interesting cases, shown in bold, to note; the subset in which NumCitations and NumClaims features were dropped as well as that in which the NumCitations and NonPatCitations were dropped.

We would have expected dropping of two features to have more adverse effects on accuracy than dropping a single feature but this is not the case when we compare the effect of dropping NumCitations and NumClaims with Patent Class when using SVM; the decrease is 11.9% and 47.2%. This seems to suggest that Patent Class has a strong bearing on accuracy thus dropping it has more adverse effects than dropping both NumCitations and NumClaims combined. However, this is easy to explain; as noted earlier, all the other features differ significantly across technological area thus Patent Class performs a significant role in normalizing values for other features across different technological areas. As would be expected, dropping the Patent Class severely affects the effectiveness of all other features in discerning between ET and NET clusters thus accuracy suffers significantly. This explanation is corroborated by results in Table 4 which indicate that Patent Class is the only single feature that causes abnormally sharp deterioration in accuracy when dropped. For example, when using SVM, dropping the Patent Class caused a decrease of 47.2% in accuracy compared to 16% when CASI, its closest challenger, was dropped.

In fact, among all the 127 possible subsets for each of the six algorithms we tested in this study, it is only the patent class that if singly discarded from the feature set will make accuracy of the remaining subset drop to the worst six performing subsets. On the other hand, NumCitations and NumClaims are the two features that will have the least effect on accuracy if dropped together.

5. Conclusion and recommendations

Whereas most researchers have tackled the problem of predicting emerging technologies using unsupervised learning, this paper has shown it is possible to use supervised learning methods to predict emerging technologies. We overcome the problem of unavailable labeled emerging technologies' data by labeling the data using the patents published in the USPTO patent database. Patent citations are highly regarded by the research community as a rich source of data for studies seeking to trace evolution of technology. However, most studies have been limited by the time lag it takes a patent to attract citations from others and we overcome this limitation by relying on backward citations that are available as soon as a patent is published. Although in this study we chose to use the USPTO, which has been found to have a lot of potential to studies of technological evolution (Valverde, 2014), other patents databases or a combination of them can also be used without significantly affecting the outcome. Our study is part of the growing interest to use patent databases to predict the direction technology is likely to take. The ultimate validation for our model was to test how well it can predict emergence of technologies which is a very difficult task.

Although our method offers a promise of someday being able to make forecasts several years before they happen, this dream is far from being reached as we appreciate the method has still a number of limitations. One major limitation is that as of 2015, the USPC was replaced by a more widely accepted Cooperative Patent Classification (CPC) as the official patent classification scheme of both the USPTO and EPO. It remains to be seen whether the CPC will receive regular creation of new classes as new technologies emerge so as to provide a basis for our method to label clusters of patents. However, it should be noted that new established classes in the USPC only played a role of labeling clusters which can be played by other sources. Intuitively, our method suggests that once we have data in form of labeled clusters of patents, we can use that data to forecast new technologies; several sources, which include new classes established by the USPTO, are available and the source used does not affect our method. Other potential sources of data to link patent clusters to technologies that emerged at a given time include mining major news websites, social networks such as Twitter and Facebook, and tech forums to identify topical technologies discussed at that time, conferences, television and radio networks, and blogs among others. These sources have a great potential (Moed, 2005) since they normally discuss trending technologies and will overcome the problem of technologies that take a long time to find their way into patent databases. There is also a limitation caused by some patent databases that do not require applicants to include citations. Moreover, studies (Alcácer and Gittelman, 2006) have found a qualitative difference between applicant provided citations and examiner provided citations. We believe that separating these two kinds of citations may have a significant improvement on accuracy of our model. Furthermore, we believe that using newly established subclasses as opposed to classes to label patent clusters will make the method more sensitive to detect specific new technologies rather than detecting large clusters that include a broad range of technologies. Addition of more useful features may also improve the accuracy of our model since the seven features used in this study are somewhat very few to discriminate among the large datasets and range of technologies. Studies have suggested several of these features which include source of funding, amount of funding, growth in funding (Moed, 2005).

References

Albert, M.B., Avery, D., Narin, F., McAllister, P., 1991. Direct validation of citation counts as indicators of industrially important patents. Res. Policy 20, 251–259. http://dx.doi.org/10.1016/0048-7333(91)90055-U.

Alcácer, J., Gittelman, M., 2006. Patent citations as a measure of knowledge flows: the influence of examiner citations. Rev. Econ. Stat. 88, 774–779. http://dx.doi.org/10.1162/rest.88.4.774.

Blondel, V.D., Guillaume, J.-L., Lambiotte, R., Lefebvre, E., 2008. Fast unfolding of communities in large networks. J. Stat. Mech: Theory Exp. 10008, 6. http://dx.doi.org/10.1088/1742-5468/2008/10/P10008.

Boyack, K., Klavans, R., 2010. Co-citation analysis, bibliographic coupling, and direct citation: which citation approach represents the research front most accurately? J. Am. Soc. Inf. Sci. Technol. 61, 2389–2404. http://dx.doi.org/10.1002/asi.

Breitzman, A., Thomas, P., 2015. The emerging clusters model: a tool for identifying emerging technologies across multiple patent systems. Res. Policy 44, 195–205. http://dx.doi.org/10.1016/j.respol.2014.06.006.

Chen, C., Ibekwe-SanJuan, F., Hou, J., 2010. The structure and dynamics of cocitation clusters: a multiple-perspective cocitation analysis. J. Am. Soc. Inf. Sci. Technol. 61, 1386–1409. http://dx.doi.org/10.1002/asi.21309.

De Solla Price, D.J., 1965. Networks of scientific papers. Science (80-.) 149, 510. http://dx.doi.org/10.1126/science.149.3683.510.

Érdi, P., Makovi, K., Somogyvári, Z., Strandburg, K., Tobochnik, J., Volf, P., Zalányi, L., 2013. Prediction of emerging technologies based on analysis of the US patent citation

network. Scientometrics 95, 225–242. http://dx.doi.org/10.1007/s11192-012-0796-4.
Ester, M., Kriegel, H.P., Sander, J., Xu, X., 1996. A density-based algorithm for discovering clusters in large spatial databases with noise. In: Proceedings of the 2nd International Conference on Knowledge Discovery and Data Mining, pp. 226–231 (doi:10.1.1.71.1980).
Fleming, L., 2001. Recombinant uncertainty in technological search. Manag. Sci. 47, 117–132. http://dx.doi.org/10.1287/mnsc.47.1.117.10671.
Fleming, L., Sorenson, O., 2004. Science as a map in technological search. Strateg. Manag. J. 25, 909–928. http://dx.doi.org/10.1002/smj.384.
Fleming, L., King, C., Juda, A.I., 2006. Small worlds and regional innovation. SSRN Electron. J. http://dx.doi.org/10.2139/ssrn.892871.
Furukawa, T., Mori, K., Arino, K., Hayashi, K., Shirakawa, N., 2014. Identifying the evolutionary process of emerging technologies: a chronological network analysis of World Wide Web conference sessions. Technol. Forecast. Soc. Chang. 91, 280–294. http://dx.doi.org/10.1016/j.techfore.2014.03.013.
Garfield, E., 1979. Citation indexing: its theory and application in science, technology, and humanities. Humanities 1983. http://dx.doi.org/10.1086/601003.
Ittipanuvat, V., Fujita, K., Sakata, I., Kajikawa, Y., 2014. Finding linkage between technology and social issue: a literature based discovery approach. J. Eng. Technol. Manag. 32, 160–184. http://dx.doi.org/10.1016/j.jengtecman.2013.05.006.
Karvonen, M., Kässi, T., 2013. Technological Forecasting & Social Change Patent citations as a tool for analysing the early stages of convergence. Technol. Forecast. Soc. Chang. 80, 1094–1107. http://dx.doi.org/10.1016/j.techfore.2012.05.006.
Kayal, A.A., Waters, R.C., 1999. An empirical evaluation of the technology cycle time indicator as a measure of the pace of technological progress in superconductor technology. IEEE Trans. Eng. Manag. 46, 127–131. http://dx.doi.org/10.1109/17.759138.
Kessler, M.M., 1963. Bibliographic coupling between scientific papers. Am. Doc. 14, 10. http://dx.doi.org/10.1002/asi.5090140103.
Klavans, R., Boyack, K.W., 2015. Which type of citation analysis generates the most accurate taxonomy of scientific and technical knowledge? J. Assoc. Inf. Sci. Technol. 68, 984–998.
Lai, K.K., Wu, S.J., 2005. Using the patent co-citation approach to establish a new patent classification system. Inf. Process. Manag. 41, 313–330. http://dx.doi.org/10.1016/j.ipm.2003.11.004.
Lampe, R., 2012. Strategic citation. Rev. Econ. Stat. 94, 320–333. http://dx.doi.org/10.1162/REST_a_00159.
Michel, J., Bettels, B., 2001. Patent citation analysis: a closer look at the basic input data from patent search reports. Scientometrics 51, 185–201. http://dx.doi.org/10.1023/A:1010577030871.
Moed, H.F., 2005. Citation analysis in research evaluation. Cit. Anal. Res. Eval. 323–346. http://dx.doi.org/10.1007/1-4020-3714-7.
Moore, K.A., 2004. Worthless patents. http://dx.doi.org/10.2139/ssrn.566941.
Newman, M., 2010. Networks: An Introduction. Oxford University Press.
Park, G., Shin, J., Park, Y., 2006. Measurement of depreciation rate of technological knowledge: technology cycle time approach. J. Sci. Ind. Res. (India) 65, 121–127.
Park, H., Ree, J.J., Kim, K., 2013. Identification of promising patents for technology transfers using {TRIZ} evolution trends. Expert Syst. Appl. 40, 736–743. http://dx.doi.org/10.1016/j.eswa.2012.08.008.
Rotolo, D., Hicks, D., Martin, B.R., 2015. What is an emerging technology? In: SPRU Work. Pap. Ser. 2015–6, pp. 1–40. http://dx.doi.org/10.1016/j.respol.2015.06.006.
Seung-wook, C., Yen-yoo, Y.O.U., Kwan-sik, N.A., 2014. Forecasting promising technology using analysis of patent information: focused on next generation mobile communications. doi:http://dx.doi.org/10.1007/s11771-014-2429-y.
Shibata, N., Kajikawa, Y., Takeda, Y., Matsushima, K., 2008. Detecting emerging research fronts based on topological measures in citation networks of scientific publications. Technovation 28, 758–775. http://dx.doi.org/10.1016/j.technovation.2008.03.009.
Shibata, N., Kajikawa, Y., Takeda, Y., Matsushima, K., 2009. Comparative study on methods of detecting research fronts using different types of citation. J. Am. Soc. Inf. Sci. Technol. 60, 571–580. http://dx.doi.org/10.1002/asi.20994.
Shibata, N., Kajikawa, Y., Sakata, I., 2010. Extracting the commercialization gap between science and technology — case study of a solar cell. Technol. Forecast. Soc. Chang. 77, 1147–1155. http://dx.doi.org/10.1016/j.techfore.2010.03.008.
Smalheiser, N., 2001. Predicting emerging technologies with the aid of text-based data mining: the micro approach. Technovation 21, 689–693. http://dx.doi.org/10.1016/S0166-4972(01)00048-7.
Small, H., 1973. Co-citation in scientific literature: a new measure of the relationship between two documents. J. Am. Soc. Inf. Sci. http://dx.doi.org/10.1002/asi.4630240406.
Sorenson, O., Rivkin, J.W., Fleming, L., 2006. Complexity, networks and knowledge flow. Res. Policy 35, 994–1017. http://dx.doi.org/10.1016/j.respol.2006.05.002.
Swanson, D.R., 1987. Two medical literatures that are logically but not bibliographically connected. J. Am. Soc. Inf. Sci. 38, 228–233.
Tang, J., Wang, B., Yang, Y., Hu, P., Zhao, Y., Yan, X., Gao, B., Huang, M., Xu, P., Li, W., Usadi, A.K., 2012. PatentMiner: topic-driven patent analysis and mining. In: Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '12. ACM, New York, NY, USA, pp. 1366–1374. http://dx.doi.org/10.1145/2339530.2339741.
Tseng, Y.-H., Lin, C., Lin, Y., 2007. Text mining techniques for patent analysis. Inf. Process. Manag. 43, 1216–1247. http://dx.doi.org/10.1016/j.ipm.2006.11.011.
Valverde, S., 2014. Evolution of patent citation networks. In: Complexity in Engineering (COMPENG). 2014. pp. 1–5.
Wong, C.-Y., Wang, L., 2015. Trajectories of science and technology and their co-evolution in BRICS: insights from publication and patent analysis. J. Inf. Secur. 9, 90–101. http://dx.doi.org/10.1016/j.joi.2014.11.006.
Yoon, J., Kim, K., 2011. Identifying rapidly evolving technological trends for R & D planning using SAO-based semantic patent networks. Scientometrics 88, 213–228. http://dx.doi.org/10.1007/s11192-011-0383-0.

Mr **Moses Ntanda Kyebambe** received his B.S. (Education) and M.S. (Computer Science) degrees in March 2003 and January 2009 respectively from Makerere University, Kampala. He is currently pursuing his Ph.D. degree at Xiangtan University, China. He has a teaching experience of 10 years. His research interests include Machine Learning and Big data Analytics. Orcid ID: http://orcid.org/0000-0001-7671-1982.

Assoc. Prof. **Ge Cheng** received his B. S. and M. S. degrees (both in Mathematics and Computational Science) from Xiangtan University, China in 2000 and 2005 respectively. He received his Ph.D. in Computer System Architecture from Huazhong University of Science and Technology in 2011. He is currently an Associate Professor at the College of Information Engineering, Xiangtan University, China. His research interests include Statistical learning method and Computer Security. Orcid ID: http://orcid.org/0000-0002-4342-8029.

Prof. **Yunqing Huang** received his B. S. and M. S. degrees (both in Computational Mathematics) from Xiangtan University, China in 1982 and 1984 respectively. He received his Ph.D. in Applied Mathematics from the Chinese Academy of Sciences in1987. He is a Professor at the School of Mathematics and Computational Science Xiangtan University and currently the President of the University. His research interests include Numerical analysis and Scientific Computing.

Mr. **Chunhui He** received his B.S. degree from Hunan City University, China in 2014. He is currently pursuing a Master's of Science degree in Xiangtan University, China. His research interests include Data mining, Natural Language Processing and Deep Learning. Orcid ID: http://orcid.org/0000-0003-1505-1620.

Mr. **Zhenyu Zhang** received his B.S. (Mathematics) degree from Lanzhou University, China in 2014. He is currently pursuing a Master's of Science degree in Xiangtan University, China. His research interests include Machine Learning and Deep Learning.