Using Machine Learning Approaches to Identify Emergence: Case of Vehicle Related Patent Data

Samira Ranaei[1], Arho Suominen[2]
[1] School of Business and Management, Lappeenranta University of Technology(LUT), Lappeenranta, Finland
[2] VTT Technical Research Centre of Finland, Espoo, Finland

Abstract—**Bibliometric studies have long used simple search strings, publications count, and word counts to track the emergence of technologies. Novel machine learning methods open new possibilities to study bibliometric data and use algorithmic approaches to uncover emergence of a technology. This study looks at the large and complex dataset of vehicle related patents to uncover emergence indicators. By using machine learning methods this study focuses on if, and to what extent different methods can produce patterns of emergence from the data directly. The data extracted from PATSTAT contains *711296* granted US patent abstracts between the years 1980 and 2014 resulting from a search for "vehicle" creating a complex dataset of technologies from automotive to medical applications. Using Latent Dirichlet Allocation and Dynamic Topic Modeling we show different emergence patterns. Finally, we discuss in detail the possibilities of using machine learning approaches to draw emergence dynamics of technologies.**

I. Introduction

Technological emergence refers to the manifestation of a drastic change to the socio-technological status-quo. The concept of technological emergence dates back to Schumpeterian definition of "creative destruction" which is a cyclical process of new innovations and being displaced by next generation of improved service or products. Modern definitions have characterized emerging technologies with; their offering of a wide range of benefits to economic sectors [1], creating new or transforming existing industries[2], being the core technology with disruptive potentials[3] that may pose economic influence in future [4] or being referred as technologies at very early stage of development[5]. A comprehensive definition of emerging technologies (ET) suggested by recent studies highlight its key features as novelty, fast growth, coherence, prominent impact and uncertainty[6], [7]. In addition to scholar's efforts toward the conceptualization of ETs, national and governmental programs have been designed to predict the emerging science and technological development. Notable among these are; European PromTech project[8] which is designed to define the direction of technological advancement or Foresight and Understanding from Scientific Exposition (FUSE) [1] research program that has been initiated by US Intelligence Advanced Research Projects Activity (IARPA) in 2011 which targets mining big data related to science and technology data sources to identify a pattern of technical emergence. Notable here is the difference of predominantly European tradition of foresight that is seen as an active agent in creating a desired future rather than the passive effort of forecasting the future. There is no surprise that so many technological forecasting methods have been developed due to the inherited uncertainty that surrounded emerging technology's future development and potential applications. From theoretical side, much research efforts have been devoted in monitoring and forecasting future technologies [1], [9], [10]. On practical side, there is a wide spectrum of methods from qualitative to quantitative techniques (take advantage of science and technological databases) have been devised for identifying the technological and scientific advances. Qualitative approaches such as scenario analysis[11] or expert panels are criticized based on their inherited nature of being subjective, expensive or not accessible. Moreover, one could no longer rely on expert opinion since the sheer amount of data stored on data sources are not possible to be analyzed without computer aided tools.

Bibliometric analysis as a quantitative approach is originated from the pioneer works of measuring scientific activity [12], [13]. The invention of bibliographic product – Science citation Index (SCI)- by Eugen Garfield that launched in year 1964 (now owned by Clarivate Analytics), further gave a boost to the bibliometric research field. Furthermore, the quantitative analysis of science and technology advancement gained ground as an independent research field known as Scientometrics where scientific papers and patent literature were used heavily as data sources [14]. The underlying assumption of mining these data origins is that innovation processes may leave a discernable pattern that can be traced and defined as technical emergence. Ever since, various number of bibliometric measures have been introduced; such as basic trend analysis [15,16] by which proliferation of published materials as unit of analysis signals the emergence, co-word analysis [17] that detects the co-occurrence of specific phrase associated with documents, citation analysis [18], co-citation [19] or citation network analysis [20] that identify the emerging structure based on citation patterns, tech-mining technique [9] that applies text analytics on natural language to extract emerging technological trend.

Despite the overall interest toward development of quantitative methods for detection of ET, there are still some obstacles in the operationalization of methods. Extraction of patent or publication data from related database is the first step in the process of detecting ETs in Scientometrics. However,

Supported by the Academy of Finland research grant (288609) "Modeling Science and Technology Systems Through Massive Data Collections"

978-1-890843-36-6 ©2017 PICMET

the scheme of these databases are not designed for Scientometrics research objectives. For example, International Patent classification (IPC) scheme has been designed to ease the process of storing the documents by examiners and not directly facilitates the retrieval process for social scientists. Therefore patent search is challenging when it comes to connecting IPC classes to industry [21] or for product level [22] analysis. Both patent classification schemes and journal article categorizations are subjective due to examiners or publishers' judgments, and might be noisy and inaccurate [23, 24]. Also, utilization of keyword based queries for document retrieval process may not completely improve the precision or recall in information retrieval from databases either. Because the authors, inventors and researchers are not consistent in using scientific terminologies. Moreover, phrases and terms will be outdated at some point as the new concepts, innovative material and process emerge. The established schemes of scientific publication or patent databases may not thoroughly corresponding to the requirements needed for uncovering ETs pattern from those big datasets. The co-word analysis method designed for detection of ET, is based on co-occurrence of "Indexed Term" available on publication database (e.g. Web of Science). These indexed labels were made by human experts, which are biased toward subjective judgment. Co-word analysis method is also criticized due to the impact of "Indexer effect" [25] which refers to the constant change of keywords and contexts while the index labels are not updated on regular basis. Citation based methods for operationalization of ET detection has some disadvantages as well. ETs are defined to be at their early stage and not have yet demonstrated their potentials. Therefore, it may relatively take longer time for a newly published scientific discovery or a disclosed invention to attract attention in form of receiving citations. Another major concern in citation network approach is the inevitable elimination of smaller citation network components for noise reduction, by which any early innovation activities will be ignored. Citation based approaches are also criticized over their applicability for macro level analysis (e.g. country level analysis), as the citation patterns are different among firms, industries and countries[26].

Text-analytics based methods have opened up new opportunities in the process of detection of ETs. Patent and publication documents comprise of textual data which are the ample source of technical information. As textual data are high dimensional and noisy, the first target of using text analytics is reducing the dimension and extracting the important information. In text analytics literature each word from the text stands for a variable (dimension). Many scholars have applied dimension reduction methods for ETs detection in Scientometrics research. Porter [9] as a pioneer scholar in technology mining has introduced "Tech-mining" concept which is defined as" the application of text mining tools to science and technology information, informed by understanding of technological innovation processes". Several dimension reduction methods have been applied for various purposes in Scientometrics research; for instance citation analysis has been combined with text-based similarity measures to identify degree of knowledge flow between the documents[27], text summarization and angle similarity measures have been applied to map and classify patents[28], the distance based clustering approach called "k nearest neighbor" was used to conduct technology opportunity analysis using patent text (description and claims) [29], and the applicability of Latent semantic Indexing (LSI) algorithm was examined to grasp the patent and paper concept similarity[30].

Most of the dimension reduction methods in previous text analytics research are built upon the co-occurring keywords (that their probabilities in a vector space represents the meaning of the corpus) and then geometric methods were applied to show the similarity of meanings. In computational linguistic and machine learning literature, these methods are called as count-based models [31] which are based on the frequency of given keywords appear in a particular context as vector of keywords. For example, LSI, vector space model (VSM) or principle component analysis (PCA) can be categorized as count-based techniques that assume words are independent. But words are sometimes dependent via synonymy or polysemy. Synonymy refers to several words with similar meaning, and polysemy can be defined as one words with various meaning. However, recent studies in machine learning [31]shows that predictive models can outperform count-based models in natural language processing context, as they directly predict a word from its neighbor terms and detect the semantic structure. One of the advantage of predictive approaches over count-based models is they consider the synonymous, polysemous words and the semantic relations between words. This paper examines the application of two predictive models in detecting the emergence of new concept from patent abstracts. Latent Dirichlet Allocation (LDA) algorithm [32], and its further extension Dynamic Topic Modeling (DTM)[33] by their definition are expected to uncover the latent pattern (topic) among huge number of documents during the time.

In this paper we are investigating whether the predictive models are able to detect emergence patterns from patent dataset. The assumption is that content of documents related ETs are deviating from the rest of dataset in terms of content and can be considered as outliers. The paper focuses on how existing topic model algorithm perform in capturing emerging technologies. For the purpose of experiment we collected patents related to vehicles, focusing on automotive ´technologies. The resulted collection is, however, a complex dataset of technologies from automotive to medical applications.

II. Background

Patents are a unique source of technical information, describing the state-of-the-art [34] and not published anywhere else. Numerous scholars in Scientometrics have utilized patent databases as one of the major data source for detecting emerging technologies; Reference [35] applied bibliometrics coupling on patent clusters to identify the vital technological paradigms in nanotechnology, Yoon[36] used patent data to detect trend in technological innovation by applying co-word analysis method, Kay and his colleagues [37] used patent citation network to identify the technological distance globally, Erdi [38] provided prediction on emerging technologies in agriculture and food industry by performing co-citation analysis on US patents.

A. Text mining and technological emergence

Due to limitations of current established methods; such as partial loss of citation network components contain less or zero citation and adverse effect of outdated words in implementation of "co-words analysis [25], text-mining based methods have been utilized as a complementary approach. Text-mining refers to the process of extracting the knowledge or non-trivial patterns from text documents [39] and convert high dimension text to representable units with less dimensions while keeping the important information. Text mining techniques has been successfully implemented by many scholars to address technology management and Scientometrics research questions. A recent study in 2015, has combined patent text analysis with patent co-classification to expose the uncertainty of emerging technologies in field of cellulosic bioethanol [40]. Reference [27] also presents a hybrid approach by integration of citation analysis with text-based similarity measures to identify degree of knowledge flow between the documents. Text summarization and angle similarity measures have been applied to map and classify patents[28]. Reference [29] used “k nearest neighbor” algorithm which is a distance based approach to conduct technology opportunity analysis on patent description and claims, [30] has examined the applicability of Latent semantic Indexing (LSI) to grasp the patent and paper similarity. Latent Dirichlet Allocation has been used by [41] to structure firms’ knowledge profiles with a case study in the telecommunication industry.

Studies have utilized text-mining methods on patent to detect the technological emergence. The majority of text analytics approaches are based on count-based models. Count-based methods [31], are dependent on the “distributional semantic models (DSMs)” [42], where similar words are placed near each other since they tend to have similar contextual distribution. Count-based methods compute the statistics of how often some word co-occurs with its neighbor words in a large text corpus, and then map these count-statistics down to a small, dense vector for each word. The issue with count-based methods in dimension reductions is “data sparseness” that causes incorrect classification of documents. Recent studies [31,43] show that predictive models outperform count-based models.

B. Dimension reduction and unsupervised predictive models

The core functionality of text mining system lies on the identification of concept co-occurrence patterns across documents collection [44]. In practice, text mining utilizes algorithmic approaches to identify distributions, frequent sets, and various associations of concepts at an inter-document level to illustrate the structure and relationships of concepts as reflected in the corpus [44]. The major challenge in text mining rises from the high dimensionality associated with natural language, where each word from the text considered as a variable and represents one dimension.

Several dimensionality reduction methods have been introduced for text clustering, the most common being LSI [45] which is based on singular value decomposition (SVD) and an extension of vector space model (VSM). Another classical approach is principle component analysis (PCA) [46]. These methods suffer from excessive information loss while pruning the data dimensions, and moreover, they cannot account for the correlated words within the given lexicon of the corpus. In other word, the methods are not very accurate due to not being able to address the polysemy (words with multiple meaning) and synonymy (multiple words with similar meaning).

The latest study trend in dimensionality reduction algorithms has been shifted from traditional count-based models to predictive probabilistic methods. Probabilistic Latent Semantic Indexing (PLSI) method proposed by [47] was a significant step forward in text analytics. It provided a probabilistic structure at word level as an alternative to LSI. PLSI model draws each word of a document from a mixture model specified via multidimensional random variable. The mixture model represents the “topics”. Therefore, each word originated from a single topic, and different words of one document can be drawn from various topics. However, PLSI lacks the probabilistic model at the document level. Since documents in PLSI are represented as the list of numbers without any generative probabilistic model for these numbers. This causes problems such as overfitting, as the number of parameters would grow linearly with the size of corpus. Another problem is PLSI inability to model documents outside of the training set.

LDA method, proposed by [32], is able to overcome the limitations of PLSI providing a probabilistic model for document level and word level analysis. LDA is a generative probabilistic model that draws latent topics from discrete data, like textual data. In LDA all documents are represented as random mixture of latent topics, and each topic is based on distribution of the words. LDA probabilistic model and its extensions have been applied by several scholars to tackle the research problems in Scientometric. Reference [48] has extended LDA by adding author information to create author-topic model. The primary benefit of the model is predicting the future research theme of specific scientists. Reference [49] showed that the topic modeling outperforms co-citation approach in producing distinctive map of author-research relatedness. Classification of large text corpora is another stream of Scientometric research that have been applied LDA for mapping the scientific publications [50, 51], topic based classification of patents [41,52] and clustering biomedical publications[53].

Dynamic Topic Model (DTM) [33] is an extension to LDA, fitting the topic model to a subset of documents with reference to time slices. DTM enables a temporal analysis, showing the rise and death of specific topics characterized by appearance or disappearance of keywords during the specified time period. The underlying assumption for DTM is that the order of documents illustrates the evolution of topics. Unlike LDA, documents analyzed by dynamic topic models are not exchangeable between similar topic. The dataset will be sliced by time and the topics of documents from each time slices will be modeled. Mathematically each topic in DTM is represented as a distribution of words, and in parallel to change of time slices, the word distribution will be changed. In Scientometrics DTM has been used as a new proxy for measuring the impact of scholars work [43].

The general hypothesis in this paper is that the topic of documents related to the emerging technologies might be slightly different from the rest of the dataset. Therefore, the main goal is detecting the topics (feature) from textual data and cluster documents according to extracted topics. For the purpose of experiment, the two algorithm of LDA and DTM will be applied on a set of patent data abstracts to detect the emerging concepts.

III. DATA COLLECTION AND METHODOLOGY

This study used the vehicle industry as a case study, and the target is examining the performance of two predictive algorithms in detection of relevant topics or possible emerging technologies. The granted patent documents have been collected from PATSTAT database requiring on the the search terms “car”, “vehicle” or “automobile” appearing in the abstract field. The data retrieval was also restricted to a period from 1980 to 2014. This resulted in 711 296 patent documents being extracted for analysis. The dataset selected for this study is relatively large, and also complex in the sense that the search retrieved documents that contain the term “vehicle”, which is not specific to automobiles but is widely used in several contexts. This creates heterogeneous dataset not solely focusing on the automobile use of the term vehicle. This adds additional complexity as our effort is to identify automobile related emergent topics.

Prior to analysis, the documents abstracts were preprocessed using natural language processing library (NLTK) Python library. The textual data was manipulated by removing all punctuations, stop-words, tokens containing numbers or consisting solely of numbers. Then the sentences were transformed to tokens (uni-grams) and save as a dictionary of words. In the further analysis, the dictionary (bag of words-BOW) has been used as an input for the two predictive unsupervised models.

A. LDA procedure

The tokenized data was analysed with an LDA algorithm implemented in Python, using an online variational Bayes algorithm for LDA [54]. The algorithm goes through the input data in chunks, updating the model as new data is analysed, allowing for a relatively large corpus being run with a relatively small computation effort. LDA relies on its formal framework to model the input data, but requires the user to set the number of topics produced as an output. Although statistical methods are available to evaluate the number of topics [32] the validity of these measures can be questioned [55]. In this study, the number of topics was evaluated through a trial and error, at each stage evaluating the word to topic probabilities as described in the following. The number of topics used for the analysis is 25.

LDA creates two matrices, document probabilities and word probabilities. These probabilities are probability distribution for each patent to belong to one of the topics (a document probability matrix and for each word in the corpus to belong to a certain topic (a word probability matrix). The topic probability distribution of each document, omitting small probabilities, was used to create a directed network. In the network, nodes are latent topics created by the algorithm and individual documents in the dataset. The edges between the nodes are directed from document to topic and the weight of an edge defined by the probability of the document belonging to a certain topic. The word probability distributions were used to create word clouds used to evaluate the content of each topic. The 50 top words are used to create the word clouds. The content and quality of the topics are evaluated using the word clouds. This is done by evaluating how concentrated the topics are to having high probability in only one or a few words, how semantically “cohesive” the topics are, and how well a “topic model’s decomposition of a document as a mixture of topics agrees with human associations of topics with a document” [55].

Compared to dynamic topic modelling, the formal framework of LDA does not include a temporal constraint. The algorithm is unaware if the documents in the corpus are spaced differently in time. Thus information on the year of publication of each document is included to the results ex post. Using the publication year of each patent document and the document topic probability matrix, the results are aggregated to a year to topic matrix A, where A_{ij} is the sum of probabilities of year i patents over latent topic j. Finally, the year to topic matrix is aggregated using the hard clustering of topics creating a year to cluster matrix. The year to topic and year to cluster matrices are used to uncover topics that grow over time in order to find potentially emerging topics.

B. DTM procedure

The key different in implementation of DTM in comparison to LDA is that the order of documents is considered per time slices, which means that in contract to LDA, documents are not exchangeable in the corpus. The patents filed in 1980 related to vehicle industry are of course in the same context of patents filed in year 2000, but the manufacturing and production process of vehicles are much different than the past. The content and topics of patent documents are evolving over time. Therefore, it is interesting to model the dynamic of contents using the sequentially organized corpus of patent documents. The same patent data set has been used to perform the DTM. The data set is divided to 35 time slices (each time slice represent a year). Then the document of each time slice will be modeled with three number of topics using DTM algorithm from Gensim library in Python. The topics affiliated with time slice t evolve from topics related to time slice t-1.

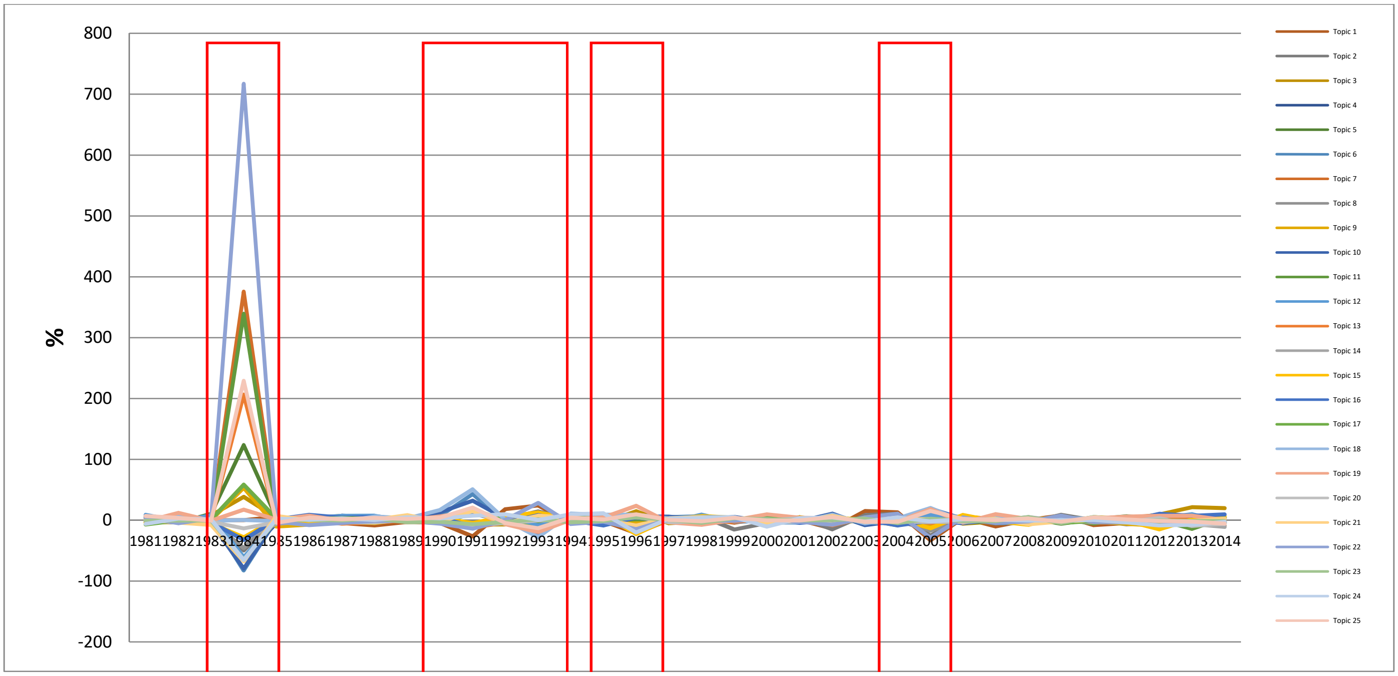

Fig. 1. The sum of probability percentage change between years. The figure is show years in the x-axis and change on y-axis. Lines are topics. The topic is only used to highlight years that have signficant changes in important topics, highlighted in the figure.

IV. Results

A. LDA results

The output document to topic matrix creates a network with 708 003 nodes and 5 307 977 edges. The difference between patent retrieved and nodes in the result are a result of filtering patent with a shorter than 100 character abstract. Aggregated to year to topic matrix the result creates a sum of probability matrix showing the growth patterns of each topic through 1981 to 2014. Normalized to the overall growth of patents yearly, the topics grow on average on 17 years (S=4, N=25). The average of change per topic is 1.96 percent (S=5.24, N=25). The highest number of growth years is 22, out of the 34 years in the time series, the lowest number of growth years is nine. The dynamics of topics produced by the LDA algorithm shows an interesting variation from near plateau changes to years of high relative change in the time series. Seen in Figure. 1, at five specific times, the sum of probability distribution across topics have a significantly high standard deviation, namely 1984, 1991, 1993, 1996 and 2005. These changes, at each time, results in a number of topics growing in their relative importance of the whole data set, while others lose their importance.

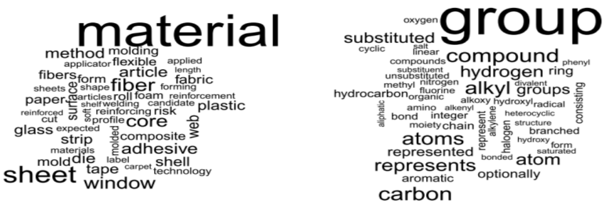

Fig. 2. The 1984 was the year with most dynamic changes in the dataset. Word cloud of topics 18 (left) 25 (right).

Looking at the most relative sum of probabilities, all of the five largest topics changed between 1983 and 1984. Later, in 1991, the change was less dramatic with two new topics emerging to the group of five largest topics, namely Topic 18 and Topic 25. The discontinuity in 1993 only meant that there were smaller changes in the largest topics, only in 1996 introducing a topic that had never before been included to the five largest topics (Topic 19). Finally, in 2005 the changes are again only among the group of topics that had previously been in the group of five largest topics. Qualitatively there has been several major transition points in the time series with new, emergent, topics arising. These are the emergence of Topics 22, 1, 5, 3 and 7 in 1984, emergence of Topics 18 and 25 in 1991, and the emergence of Topic 19 in 1996. Appendix A list the five largest topics at each year, illustrating changes between topic significance.

We use word clouds to describe the content of the topics. Seen in Figure 2. is the word cloud of Topic 18 and 25 emerging in 1991. These topics highlight the emergence of two materials related themes in the dataset; Topic 18 on sheet materials and Topic 25 on chemical compounds related to hydrogen and hydrocarbon. The emergence of hydrogen topics 1991 can be argued to be timely, as for example the US Energy Policy Act of 1992 encouraged the development of alternative fuel vehicles. Topic 18 is more challenging to qualitatively analyse, but through random sample of patents analyse qualitatively relates to the used of novel material or existing material in novel ways in vehicles.

B. *DTM results*

The output of DTM analysis is a three-dimension matrix that represent probability of words per topic for each time slices. To run the further analysis, the matrix was converted to a two-dimension matrix where words related to each topic represent the rows (53317 words) and times slices are columns (35 years). To facilitate interpretations, another matrix was created with words probabilities transformed to rankings for each time slices. The number of words for each three generated topics are: 17520, 18132 and 17665. For instance, words for topic number one were ranked from 1 to 17520 per time slices.

In order to show emerging words, the matrix filtered by showing only the top ranked 10000 of words per topic. The percentage of growth rates were calculated for each word per time slices, to compare the word behaviors during the years. A conditional heat map was applied to the growth rate matrix to detect the changes for each word per time slices. The matrix shows the occurrence and disappearance of each word used in documents for all three topics per year. Figure 3 represents an example in topic number two the word "coal" appeared since the beginning of the time series, kept a steady growth rate till year 2000 when its growth rate started to decrease. So it means the word has been less used in the documents since 2000. Moreover, the word "hydrogen" from topic two experiences its ups and downs on several subsequent years but it's growth rate is on rise in the last four years. There are also many words with a steady growth rate during all the years. The word "wheel" in topic number three has appeared in all 35 years with more or less similar rate of growth.

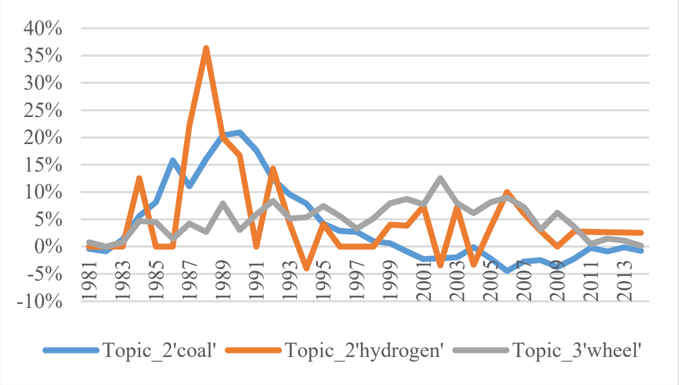

Fig.3. Growth percentages of the three randomly chosen term from DTM result table

V. Discussion

The detection of emerging scientific or technological advances to make prediction about possible future direction have been always the concern of managers at national, industry or firm level. The purpose of this study was examining the applicability of the two advance machine learning algorithms in detection of emerging topics over the years. For the purpose of demonstration both methods were applied on a patent data set related to vehicle. It is heterogeneous dataset not solely focusing on the automobile use of the term vehicle. This adds additional complexity as our effort is to identify automobile related emergent topics. The previous established methods for detection of emerging patterns in datasets are bound to limitations. First, qualitative methods (e.g. Delphi) are heavily relying on expert biased toward human judgments. In general, obtaining and analysis are expensive and time consuming. Second, the classic quantitative method is patent citation based approaches which are not applicable at macro level research. Third, co-word analysis methods are may not be accurate as the method is based on the indexed terms that are rarely updated on databases, while the content documented on patents are on constant change. Forth, most of the previous text mining approaches used for detection of emerging topics were based on count-based models that are not accurately overcome the ambiguity in natural language processing (e.g. polysemy, synonymy). The novel predictive models [31,32, 56] have opened up new opportunities in computational linguistics and machine learning areas. This study uses LDA and DTM to examine their applicability in detection of emerging patterns. The results of the study show, that the algorithms used are applicable but yield results in a different manner. LDA is unaware of the temporal dynamics of the data and through the ex-post introduction of the time series shows interesting behavior, rise and decline, of topics. The results were able to show several years with major dynamism in the probability distribution of documents. This means, that at a given time of change a new set of topics become more important, or less important, to the whole set of document. We can argue that the rising topics are interesting, but further analysis should be done to conclude if the dynamic behavior actually described novel, persistent and growing topics with a significant community that develops the technologies behind the topics.

For DTM, the results of the analysis are provided in a totally different level of aggregation. More consistent with traditional approached of looking at emerging terms, DTM focuses our attention to the word probabilities within topics and changes between the time slices. In practice, the focus of the qualitative interpretation of the result is on understanding why some words have suddenly become more, or less, important to a given topic and if this change describes some emergent behavior. This study has applied a ranking system from Pandas python library to rate words based on their probability of occurrences in each topic per year. From the resulted matrix we would able to characterize and categorized words based on their growth rate during the 35 years. For instance, the decreasing growth rate of word "coal" since 2000 shows it has been less used as a material in recent patent documents. Or the word "wheel" with a steady growth during the time slices can be classified as general word in the dataset. However, the rising growth rate of "hydrogen" in the last few years makes it a very interesting in a vehicle related topic.

This study has limitations. The quality and accuracy of the results are highly associated with cleaning and text pre-processing procedure that has been implemented prior to application of the algorithms. Specifically, with patent as legal document that contain a significant portion of text that does not have a semantic meaning relating to the technological context at hand. Another limitation would be including only the granted patents, because it may take many years for a patent to be granted. For detection of emerging topics, it might be more productive to conduct an analysis on all the filed patents regardless of their legal status in patent database. But this bring an issue of data quality that should be managed. Also, due to the limited processing power of utilized computer we could not generate and analyze more than three topics for DTM. Last but not least, the overall quality of patent abstract as a source of data can be questioned. Reference [41]discusses the problem in detail and uses full-text to run the patent analysis, arguing that the poor information value in patent abstract makes for a poor source of data for machine-learning algorithms. For future research, it would worth to examine the sub-topics and uncover smaller emerging patterns. The paper format restricts the capabilities of LDA or DTM, as it is clear that for the analysis to be practical more topics, and shorter time slices should have been produced. This would however resulted in such extensive results that are hard to communicate through print. This point highlight the need of a visualization platform or calculation framework designed for better understanding of the results. Also, it would be interesting to assess the performance of these predictive models on the datasets with smaller size to test their sensitivity toward the size of datasets. Further work should also focus on producing a predictive component to the results. Already tested by Suominen [41], the results matrices could with relatively ease be extend to the future to truly show future emergent behavior.

References

[1] B. R. Martin, “Foresight in science and technology,” *Technology Analysis & Strategic Management*, vol. 7, no. 2, pp. 139–168, Jan. 1995.

[2] G. S. Day and P. J. H. Schoemaker, “Avoiding the Pitfalls of Emerging Technologies,” *California Management Review*, vol. 42, no. 2, pp. 8–33, Jan. 2000.

[3] S.-C. Hung and Y.-Y. Chu, “Stimulating new industries from emerging technologies: challenges for the public sector,” *Technovation*, vol. 26, no. 1, pp. 104–110, 2006.

[4] A. L. Porter, J. D. Roessner, X.-Y. Jin, and N. C. Newman, “Measuring national ‘emerging technology’ capabilities,” *Science and Public Policy*, vol. 29, no. 3, pp. 189–200, Jun. 2002.

[5] W. Boon and E. Moors, “Exploring emerging technologies using metaphors–a study of orphan drugs and pharmacogenomics,” *Social science & medicine*, 2008.

[6] H. Small, K. W. Boyack, and R. Klavans, “Identifying emerging topics in science and technology,” *Research Policy*, vol. 43, no. 8, pp. 1450–1467, Oct. 2014.

[7] D. Rotolo, D. Hicks, and B. R. Martin, “What is an emerging technology?,” *Research Policy*, vol. 44, no. 10, pp. 1827–1843, Dec. 2015.

[8] I. Roche, D. Besagni, C. François, M. Hörlesberger, and E. Schiebel, “Identification and characterisation of technological topics in the field of Molecular Biology,” *Scientometrics*, vol. 82, no. 3, pp. 663–676, Mar. 2010.

[9] A. Porter and S. Cunningham, *Tech mining: exploiting new technologies for competitive advantage*. New Jersey: Wiley, 2005.

[10] A. L. Porter, S. W. Cunningham, J. Banks, Roper Thomas, Mason Thomas, and R. Frederick, *Forecasting and Management of Technology*. Wiley; 2 edition, 2011.

[11] K. Van der Heijden, *Scenarios : the art of strategic conversation*. John Wiley & Sons, 1996.

[12] A. Pritchard, “Statistical bibliography or bibliometrics,” *Journal of documentation*, 1969.

[13] D. J. D. S. Price, *Little Science, Big Science...and Beyond*. Columbia University Press, 1965.

[14] W. Glänzel and U. Schoepflin, “Little scientometrics, big scientometrics ... and beyond?,” *Scientometrics*, vol. 30, no. 2–3, pp. 375–384, Jun. 1994.

[15] A. L. Porter and M. J. Detampel, “Technology opportunities analysis,” *Technological Forecasting and Social Change*, vol. 49, no. 3, pp. 237–255, Jul. 1995.

[16] M. Bengisu, “Critical and emerging technologies in Materials, Manufacturing, and Industrial Engineering: A study for priority setting,” *Scientometrics*, vol. 58, no. 3, pp. 473–487, 2003.

[17] M. Callon, J.-P. Courtial, W. A. Turner, and S. Bauin, “From translations to problematic networks: An introduction to co-word analysis,” *Social Science Information*, vol. 22, no. 2, pp. 191–235, Mar. 1983.

[18] E. Garfield, I. H. Sher, and R. J. Torpie, “THE USE OF CITATION DATA IN WRITING THE HISTORY OF SCIENCE,” no. 64, 1964.

[19] H. Small, “Co-Citation in Scientific Literature: A new measure of the relationship between two documents,” *Journal of the American Society for Information Science*, vol. 24, no. 4. pp. 265–269, 1973.

[20] N. Shibata, Y. Kajikawa, Y. Takeda, I. Sakata, and K. Matsushima, “Detecting emerging research fronts in regenerative medicine by the citation network analysis of scientific publications,” *Technological Forecasting and Social Change*, vol. 78, no. 2, pp. 274–282, Feb. 2011.

[21] U. Schmoch, “Concept of a technology classification for country comparisons,” *Final report to the world intellectual property*, 2008.

[22] A. Pilkington, R. Dyerson, and O. Tissier, “The electric vehicle:Patent data as indicators of technological development,” *World Patent Information*, vol. 24, no. 1, pp. 5–12, Mar. 2002.

[23] G. F. Nemet, “Inter-technology knowledge spillovers for energy technologies,” *Energy Economics*, vol. 34, no. 5, pp. 1259–1270, 2012.

[24] K. B. Dahlin and D. M. Behrens, “When is an invention really radical?: Defining and measuring technological radicalness,” *Research Policy*, vol. 34, no. 5, pp. 717–737, 2005.

[25] L. Leydesdorff, “Why words and co-words cannot map the development of the sciences,” *... of the American society for information science*, 1997.

[26] J. Alcacer and M. Gittelman, “Patent citations as a measure of knowledge flows: The influence of examiner citations,” *The Review of Economics and Statistics*, 2006.

[27] H. Joung, Y. An, and Y. Park, “A structured approach to explore knowledge fl ows through technology-based business methods by integrating patent citation analysis and text mining,” *Technological Forecasting & Social Change*, vol. 97, pp. 181–192, 2015.

[28] Y.-H. Tseng, C.-J. Lin, and Y.-I. Lin, “Text mining techniques for patent analysis,” *Information Processing & Management*, vol. 43, no. 5, pp. 1216–1247, Sep. 2007.

[29] C. Lee, B. Kang, and J. Shin, “Novelty-focused patent mapping for technology opportunity analysis,” *Technological Forecasting and Social Change*, vol. 90, pp. 355–365, Jan. 2015.

[30] T. Magerman, B. Van Looy, and X. Song, “Exploring the feasibility and accuracy of Latent Semantic Analysis based text mining techniques to detect similarity between patent documents and scientific publications,” *Scientometrics*, vol. 82, no. 2, pp. 289–306, 2010.

[31] M. Baroni, G. Dinu, and G. Kruszewski, “Don’t count , predict ! A systematic comparison of context-counting vs . context-predicting semantic vectors,” *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics.*, pp. 238–247, 2014.

[32] D. M. Blei, A. Y. Ng, M. I. Jordan, and J. Lafferty, “Latent Dirichlet Allocation,” *Journal of Machine Learning Research*, vol. 3, p. 2003, 2003.

[33] D. Blei and J. Lafferty, “Dynamic topic models,” *Proceedings of the 23rd international conference*, 2006.

[34] D. Hunt, L. (Long B. . Nguyen, and M. (Matthew E. . Rodgers, *Patent searching : tools &amp; techniques*. Wiley, 2007.

[35] O. Kuusi and M. Meyer, “Anticipating technological breakthroughs: Using bibliographic coupling to explore the nanotubes paradigm,” *Scientometrics*, vol. 70, no. 3, pp. 759–777, Mar. 2007.

[36] J. Yoon, S. Choi, and K. Kim, “Invention property-function network analysis of patents: a case of silicon-based thin film solar cells,” *Scientometrics*, vol. 86, no. 3, pp. 687–703, Mar. 2011.

[37] L. Kay, N. Newman, J. Youtie, A. L. Porter, and I. Rafols, “Patent

Overlay Mapping: Visualizing Technological Distance," Aug. 2012.

[38] P. Érdi, K. Makovi, Z. Somogyvári, K. Strandburg, J. Tobochnik, P. Volf, and L. Zalányi, "Prediction of emerging technologies based on analysis of the US patent citation network," *Scientometrics*, vol. 95, no. 1, pp. 225–242, 2013.

[39] A. Tan, "Text mining: The state of the art and the challenges," *Proceedings of the PAKDD 1999 Workshop on*, 1999.

[40] R. Gustafsson, O. Kuusi, and M. Meyer, "Examining open-endedness of expectations in emerging technological fields: The case of cellulosic ethanol," *Technological Forecasting and Social Change*, vol. 91, pp. 179–193, 2015.

[41] A. Suominen, H. Toivanen, and M. Sepp??nen, "Firms' knowledge profiles: Mapping patent data with unsupervised learning," *Technological Forecasting and Social Change*, vol. 115, pp. 131–142, 2016.

[42] G. A. Miller and W. G. Charles, "Contextual correlates of semantic similarity," *Language and Cognitive Processes*, vol. 6, no. 1, pp. 1–28, Jan. 1991.

[43] S. Gerrish and D. Blei, "A language-based approach to measuring scholarly impact," *Proceedings of the 27th …*, 2010.

[44] R. Feldman and J. Sanger, "Text Mining Handbook: Advanced Approaches in Analyzing Unstructured Data," Jun. 2006.

[45] S. Deerwester, S. T. Dumais, and R. Harshman, "Indexing by latent semantic analysis," *Journal of the American society for information science*, vol. 41, no. 6, pp. 391–407, 1990.

[46] S. Wold, K. Esbensen, and P. Geladi, "Principal component analysis," *Chemometrics and intelligent laboratory*, 1987.

[47] T. Hofmann, "Probabilistic latent semantic indexing," in *Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval - SIGIR '99*, 1999, pp. 50–57.

[48] M. Rosen-Zvi, T. Griffiths, M. Steyvers, and P. Smyth, "The author-topic model for authors and documents," *Proceedings of the 20th conference on Uncertainty in artificial intelligence*, pp. 487–494, 2004.

[49] K. Lu and D. Wolfram, "Measuring author research relatedness: A comparison of word-based, topic-based, and author cocitation approaches," *Journal of the American Society for Information Science and Technology*, 2012.

[50] C.-K. Yau, A. Porter, N. Newman, and A. Suominen, "Clustering scientific documents with topic modeling," *Scientometrics*, vol. 100, no. 3, pp. 767–786, 2014.

[51] A. Suominen and H. Toivanen, "Map of science with topic modeling: Comparison of unsupervised learning and human-assigned subject classification," *Journal of the Association for Information Science and Technology*, no. October, p. n/a-n/a, 2015.

[52] S. Venugopalan and V. Rai, "Topic based classification and pattern identification in patents," *Technological Forecasting and Social Change*, vol. 94, pp. 236–250, Nov. 2015.

[53] K. W. Boyack, D. Newman, R. J. Duhon, R. Klavans, M. Patek, J. R. Biberstine, B. Schijvenaars, A. Skupin, N. Ma, and K. Börner, "Clustering more than two million biomedical publications: Comparing the accuracies of nine text-based similarity approaches," *PLoS ONE*, vol. 6, no. 3, 2011.

[54] D. Blei and M. Hoffman, "Online Learning for Latent Dirichlet Allocation," *Neural Information Processing Systems*, 2010.

[55] J. Chang, S. Gerrish, and C. Wang, "Reading tea leaves: How humans interpret topic models," *Advances in*, 2009.

[56] D. M. Blei and J. D. Lafferty, "A correlated topic model of Science," *The Annals of Applied Statistics*, vol. 1, no. 1, pp. 17–35, 2007.

APPENDIX A

TABLE 1 LDA RESULT – WORDS PER TOPIC

#Topic	*Top keywords*
1	control, device, unit, power, system, battery, energy storage, electric, electrical, charging
2	cell, cells, membrane, transistor,
3	light, emitting, lamp, element, beam, lens
4	acid, catalyst, solvent, compound, water, mixture, metal, carbonate, reaction, carboxylic,
5	cardiac, diagnostic, blood, strap, protection, belt
6	pressure, fluid, valve, pump, hydraulic, cylinder, chamber, brake, vehicle intake, compression, passage,
7	vehicle, engine, motor, steering, driving, control, electric, gear, wheel, hybrid
8	data, signal, information, network, communication, transmitting, wireless, antenna, digital
9	image, display, optical sensor, light, camera, radiation
10	panel, container, opening, compartment, segments, route, wall
11	treatment, disease, compound, pharmaceutical, skin, enzyme
12	fuel, chamber, air, water, gas, flow, exhaust, combustion, pipe, temperature, liquid, nozzle, filter, outlet, tank, cooling,
13	shaft, drive, transfer, carriage, mechanism, track, ink, roller, machine,
14	module, circuit, plurality, board, exchanger, chip,
15	layer, substrate, carbon, semiconductor, film, surface, material, metal, electrode,
16	user, services, media, payment, patient, tissue, credit, delivery, gate, key, system
17	carbon, gas, dioxide, hydrogen, stream, catalyst
18	material, sheet, molding, method, flexible, adhesive, plastic, window, glass
19	body, portion, surface, outer, part, lower, main
20	card, computer, memory, account, machine, smart, reader, game, interface
21	member, housing, plate, assembly, cartridge, mechanism, device, connector, configured
22	particles, energy, wafer, absorption, charge, diameter, powder, domain, chemical material, prevention
23	vehicle, seat, upper, door, support, structure, rear, lower, assembly, body, base, mounting, system
24	component, polymer, rubber, copolymer, mass, resin, parts, functional, molecular, polycarbonate, polyester, thermoplastic
25	group, hydrogen, compound, alkyl, consisting, hydrocarbon, alkaline,