Scientometrics (2020) 122:215–247
https://doi.org/10.1007/s11192-019-03275-w

Evaluating technological emergence using text analytics: two case technologies and three approaches

Samira Ranaei[2] · Arho Suominen[1] · Alan Porter[3,4] · Stephen Carley[3]

Received: 13 February 2019 / Published online: 2 November 2019

Abstract

Scientometric methods have long been used to identify technological trajectories, but we have seldom seen reproducible methods that allow for the identification of a technological emergence in a set of documents. This study evaluates the use of three different reproducible approaches for identifying the emergence of technological novelties in scientific publications. The selected approaches are term counting technique, the emergence score (EScore) and Latent Dirichlet Allocation (LDA). We found that the methods provide somewhat distinct perspectives on technological. The term count based method identifies detailed emergence patterns. EScore is a complex bibliometric indicator that provides a holistic view of emergence by considering several parameters, namely term frequency, size, and origin of the research community. LDA traces emergence at the thematic level and provides insights on the linkages between emerging research topics. The results suggest that term counting produces results practical for operational purposes, while LDA offers insight at a strategic level.

Keywords Technological emergence · Topic modeling · Emergence score (EScore) · Text analytics

Introduction

Technological emergence refers to the manifestation of a drastic change to the socio-technological status quo. There has been a long-standing fascination with technological emergence processes. In 1902, Wells (1999) already argued that approaching the implications of new technologies in a systematic manner would enable a better society. Later, Schumpeter

✉ Arho Suominen
arho.suominen@vtt.fi

[1] Innovations, Economy, and Policy, VTT Technical Research Centre of Finland, Vuorimiehentie 3, P.O. Box 1000, 02044 Espoo, Finland

[2] School of Engineering Science, Lappeenranta University of Technology, Lappeenranta, Finland

[3] Search Technology, Inc., 6025 The Corners Pkwy., Peachtree Corners, GA 30092, USA

[4] Science Technology and Innovation Policy (STIP) Program, School of Public Policy, Georgia Tech, Atlanta, GA 30332-0345, USA

explained the phenomenon by defining “creative destruction, which is a cyclical process of new innovations that will be displaced by next generation of improved services or products. Recent literature has highlighted the importance of emerging technologies (ETs) and explored their characteristics. Various authors have argued that ETs offer a wide range of benefits to the economic sectors (Martin 1995), create new or transform existing industries (Day and Schoemaker 2000), have high disruptive potential (Hung and Chu 2006), or can exert economic influence in the future (Porter et al. 2002).

Regardless of the recent efforts to define ETs, analyzing technological emergence can be seen as a pool of methodological approaches rather than a rigorous theory and set of reproducible methods (Suominen and Newman 2017). Theoretically, discussion of technological emergence is linked to the literature on technological change, for example, the evolutionary theory of technological change (ETTC) or technological innovation systems. ETTC offers a biological evolution analogy to understand the development of complex technological systems; other schools of thought apply different lines of reasoning. Focusing on the methodological elements, multiple methods have been used to model the path of technological development. Probably the most well known is the “workhorse” of technological forecasting—trend extrapolation (Lecz and Lanford 1973).

Although we can easily understand how the selection process explains the survival of the fittest, it cannot explain the arrival of the fittest. Arrival, or emergence, as we would rather call it, is a term used to define “the arising of novel and coherent structures, patterns, and properties during the process of self-organization in complex systems (Goldstein 1999). Working toward a robust operationalization of emergence, we look toward a proxy that would explain the arrival of the fittest. Therefore, we are relatively locked-in with the ex-post evaluation of emergence, but we want to be able to identify emergence as quickly as possible.

In the literature (Suominen and Newman 2017), it has been argued that emergence has five distinct characteristics, as follows: ostensivity, global presence, coherence, dynamism, and novelty. These elements of ET require us to identify examples (ostensivity), the large-scale adoption (globality and coherence), newness, and growth of an ET using a proxy measure. Large national and governmental programs, such as the European PromTech project (Roche et al. 2010) or Foresight and Understanding from Scientific Exposition (FUSE) research program, of the US Intelligence Advanced Research Projects Activity (IARPA) in 2011 that targets mining big data related to science and technology (Suominen and Newman 2017), have attempted to do just that. However, robust methodological approaches are seldom shown and tested in this arena.

Several approaches have, however, been proposed in the literature. An emerging cluster model focused on near immediate identification or predictive analysis of emerging clusters based on patent citation data (Breitzman and Thomas 2015). This approach is with limitations due to the different approaches to patent citations in different regions (Criscuolo and Verspagen 2008), which limit the methods applicability. Text mining has been seen as a particularly “effective means” (Kim and Lee 2017) for detecting novelty. A particular advantage is the reliance on the text of the author, which can be taken to often provide the voice of the person carrying out the novelty creation. As seen in Gerken and Moehrle (2012) recent literature in novelty detection has focused on textual data, but limited to a significant extent to keywords or the subject action object (SAO) structures.

This study adds to the existing literature on technological emergence by testing three different approaches to identify elements of emergence in a technological domain. Adopting the framework by (Suominen 2013), we elaborate on how emergence can be measured by simple count-based measures, when additional information is integrated into the simple counts and when a machine learning algorithm is used. This approach enables evaluating at which level of analytical complexity emergence can be detected. Using case studies on two technologies, light-emitting diodes (LEDs) and flash memory, the results highlight how different approaches toward analyzing emergence can yield different outcomes. The objective is not to rank methods based on their superiority, but rather to show how these three approaches to measure ETs could differentially inform science policy, R&D management, and competitive technical intelligence.

The results suggest that methods should be selected based on their intended use, and that even relatively simple approaches can yield practical results. Calculating TF-IDF weighted term delta values by year resulted in highly detailed emergent pattern identification. The emergence score (EScore) builds on the expectation of coherence and stability in picking up novelty, creating needed stability in the measures. Finally, Latent Dirichlet Allocation (LDA) creates a baseline for strategic visual mapping that allows changes in the R&D landscape to be tracked, but has difficulty highlighting rapid small shifts.

Background

Operationalizing technological emergence

There is a strong body of literature that considers how we can measure and forecast technological progress. Such research has developed from the work of Ayres (1969) in 1969 to the published book in 2011 (Porter et al. 2011) on managing and forecasting technologies and more recently on tech mining (Porter and Cunningham 2005). To track and forecast technological change a number of methods and data sources have been used (Suominen 2013). Methods have considered technological options (Zheng et al. 2012; Ranaei et al. 2014), technological systems (Guo et al. 2012), and most notably, S-shaped growth curves (Suominen and Seppänen 2014). However, in this body of literature, the specific issue of emergence and how it can be operationalized has received relatively little attention.[1]

Scholars have focused on emergence as a construct, as in Goldstein (1999) and Holland (2000), and as a case study operationalization, as in Small et al. (2014) and Zhang et al. (2016). The managerial utility of emergence comes from a better understanding of the radical and disruptive shifts occurring in industry and society as a whole (for a discussion on radical innovation and emergence, refer to Li et al. (2017)). The textbook approach to analyzing emergence has been using one variable, such as publication count, and extrapolating it to the future (Lecz and Lanford 1973; Porter et al. 2011); although more complex multi-variable and machine learning—based approaches have been proposed, such as in Lee et al. (2018). Developing practical emergence detection methods is very much a work in progress. This prompts us to look at the elements of emergence to better understand how it can be operationalized.

As mentioned, we operationalize emergence in terms of five characteristics, namely ostensivity, global presence, coherence, dynamism, and novelty. It is a measuring problem, as we have uncertainties about what we are measuring. ETs are a complex whole bound in both time and place (for a discussion, refer to Feenberg (2010)); thus, we must rely on the

[1] This was the subject of discussion during the 2017 PICMET conference that hosted a track and round table on technological emergence.

emergent elements revealing themselves. This behavior, referred to as ostensivity, is one of the key aspects of emergence. Ostensivity is based on the emergent being both novel and resulting from a dynamic process in a complex system. Even if we were able to identify parts of the whole and model parts of the system, it is ultimately the whole that creates the emergent. Templeton and Fleischmann (2013) operationalized the ostensive properties of an ET via a map of science and technology.

The characteristics of an ET also require that there is a global presence, although this notion has been critiqued (Small et al. 2014). A global presence does not mean that a technology should be adopted equally everywhere, which may even be impossible (Feenberg 2010). Rather, this characteristic expects an emergent to be known broadly, rather than just within a micro-level social structure. This point should hold even if the ET's applicability is only in niches. The expectation of macro-level knowledge of an emergent also links to the expectation of coherence. That is to say, post-emergence, the emergent science or technology should be somewhat stable and reflect a shared view on how people understand it. The five characteristics described above lay the foundation for a framework for analyzing technological emergence.

Text mining and technological emergence

Text analytics based methods have opened up new opportunities in the process of detecting ETs. Text mining refers to the process of extracting the knowledge or nontrivial patterns from text documents (Tan 1999) and converting high-dimensional text to representable units with fewer dimensions, while keeping the important information. The concept of "tech-mining" (Porter et al. 2011) has been defined as "the application of text mining tools to science and technology information, informed by understanding of technological innovation processes. The core functionality of a text-mining system lies in the identification of concept occurrence patterns across a document collection (Feldman and Sanger 2006). In practice, text mining utilizes term count or algorithmic approaches to identify distributions, frequent sets, and various associations of concepts at an inter-document level, thereby illustrating the structure and relationships of concepts as reflected in the corpus (Feldman and Sanger 2006). The major challenge in text mining arises from the high dimensionality associated with natural language, where each word from the text is considered a variable representing one dimension. To model technological emergence, a text mining process is required to reduce the dimensionality of the data to proxy measures that can highlight the characteristics of an ET.

The most elementary approach of using text mining to track technological emergence is the identification of novel terms that become prevalent as time goes on Suominen (2013). This type of elementary approach focuses simply on the emergence of novel concepts. The assumption is that these terms can be used as proxies that are able to describe the emergent aspects within a technological field. More complex approaches to understand technological emergence either add additional data or use more advanced methods to analyze the count of terms in a given set of documents. For instance, text-based similarity measures combined with citation information used to identify the degree of knowledge flow between documents (Joung et al. 2015). Patent text analysis using patent co-classification was used to expose the uncertainty of ETs in the field of cellulosic bioethanol (Gustafsson et al. 2015). Methodological complexity has been added by, for example using "k nearest neighbor" to conduct technology opportunity analysis using patent text (Lee et al. 2015). Text summarization and angle similarity measures have been applied to map patents and technological pathways (Tseng et al. 2007; Lee et al. 2015). Latent semantic indexing (LSI) was used to grasp the patent and paper concept similarity, which may provide insights on the detection of technological opportunities (Magerman et al. 2010).

As computational capabilities have increased more complex analysis processes that depart from simple term count based measures have become widely used. The most common approach might be LSI (Deerwester et al. 1990). This is based on singular value decomposition (SVD), and it represents an extension of vector space model (VSM). Another classical approach is PCA (Wold et al. 1987). However, these methods suffer from excessive information loss while pruning the data dimensions; moreover, they cannot account for the correlated words in the corpus's given lexicon. In other words, the methods are less accurate because they cannot address polysemy (words with multiple meanings) and synonymy (multiple words with similar meaning).

The focus of more recent dimensionality-reduction algorithms has shifted from traditional models to probabilistic methods. Probabilistic LSI (PLSI), a method proposed by Hofmann (1999), was a significant step forward in text analytics. It provides a probabilistic structure at the word level as an alternative to LSI. Its modeling draws each word of a document from a mixture model specified via a multidimensional random variable. The mixture model represents the "topics". Therefore, each word originates from a single topic, and different words of one document can be drawn from various topics. However, PLSI lacks a probabilistic model at the document level. Documents in PLSI are represented as a list of numbers, with no generative probabilistic model for these numbers. This causes problems like over-fitting, as the number of parameters will grow linearly with the corpus size. Another problem is PLSI's inability to model documents outside the training set.

The LDA method, proposed by Blei et al. (2003), can overcome limitations of PLSI, providing a probabilistic model for document- and word-level analysis. LDA is a generative probabilistic model that draws latent topics from discrete data, like textual data. LDA relates documents, which are represented as random mixtures of latent topics, to each topic, and the topics are based on distribution of the words. The LDA probabilistic model and its extensions have been applied by several scholars to address scientometric research questions. In Rosen-Zvi et al. (2004), the authors extend LDA by adding author information to create author-topic models. The primary benefit of the model is predicting the future research theme of specific scientists. Lu and Wolfram (2012) showed that topic modeling outperforms the co-citation approach in producing distinctive maps of author-research relatedness. The classification of large text corpora is another stream of scientometric research that has been applied via LDA for mapping scientific publicationsåö (Yau et al. 2014; Suominen and Toivanen 2015), topic based classification of patents (Suominen et al. 2016; Venugopalan and Rai 2015), and clustering biomedical publications (Boyack et al. 2011).

This paper examines to what extent emergent pattern are visible through the use of term proxies derived from scientific literature. The analysis looks at ostensivity, the appearance of an emergent, and growth of any emergent pattern and how these are visible with different methods. We analyze, as a baseline, (1) an elementary term count based approach, (2) a more complex approach incorporating control parameters for term emergence and finally (3) a probabilistic based method. The capabilities of the methods are evaluated via two well-documented case studies, looking to qualitatively identify expected patterns of emergence.

Case technologies

Light-emitting diodes

Light-emitting diodes (LEDs) represent an application of semiconductor technology that emits light when activated. The technology has had several advantageous applications over several decades. As a component, LEDs have been available since the 1960s, but due to the limitation of the technology, LED applications have been restricted to small indicator lights. The first LED, presented in 1962 by Holonyak and Bevacqua (1962), had a luminous efficiency of 0.1 1 m/W. More recent developments have led to white LEDs, which have a greater luminous efficiency, enabling LEDs to be used for lighting.

The technological pathway of LED technology was founded on advancements in semiconductor technology. While the capacity to emit visible light was well known, the ability to create LEDs that could be utilized in practical applications required stable processes for manufacturing semiconductors. Although LEDs were used as indicators during the late 1960s, it was the rapid development in semiconductor technology that resulted in a near order of magnitude development in the lm/W efficiency of LEDs (Craford 1997).

Even the order of magnitude development was not enough to create the white light that could serve as a replacement for the dominant lighting technologies. White light was produced by either combining red, green, and blue LEDs or using phosphorous material to convert a blue or ultraviolet (UV) LED to a white light-emitting one (Yam and Hassan 2005). The required technological breakthrough came with the work of Nakamura and colleagues (1991, 1993); this advancement enabled a gallium nitride-based blue and green LED. This invention enabled the development of white LEDs with the technological capability of replacing the existing dominant technology. Haitz's law was used to model the exponential rate of lumen/watt efficiency development of LEDs; this law expects the efficiency of LEDs to double every 36 months, but about 2020, the development of LEDs is expected to reach a phase where they "approach the end of the efficacy ladder and meet or exceed the market's needs with respect to cost and quality" (Haitz and Tsao 2011).

LEDs offer an interesting scientometric case technology, as we can easily identify interesting points in the technological trajectory. We can pinpoint the beginning of the technology to Russia in 1927 (Zheludev 2007), a major inflection point that started with the Nakamura group's invention about 1991 (Nakamura et al. 1991) and technological advancement that is now nearing its end (Haitz and Tsao 2011).

At the current status of the LED, or solid state lighting (SSL) technology, much seems to have been accomplished, but the literature still expects new patterns to emerge. Haitz and Tsao (2011) suggested a major current advancement in the domain relating to improving the lifetime of LEDs, for example, by reducing efficiency losses due to semiconductor heating. Another avenue of LED development consists of broad complementarities where LEDs are either applied or enable other technologies to evolve. These include a "distributed last-meter for sensors, actuators, communications, and intelligence for the Internet of Things," water-purification, fiber-optic communication, or power electronics (Tsao et al. 2015).

Flash memory

Flash memory is one technology in the continuum of development among memory cell technologies. Semiconductor memory cells can be divided into two main categories, namely volatile and nonvolatile memory. Volatile memory, including SRAM and DRAM, enables fast reading and writing, but it loses its data when the power supply is turned off. Nonvolatile memory (NVM), such as flash memory, can sustain data even without a power supply. Mainly due to this characteristic, NVM has several applications. Flash memory has enabled the growth of portable electronic adoption, as it offers a practical compromise between size and flexibility. Flash memory is used for two major applications, namely code storage and data storage. As the need for these applications increases, the demand for flash memory also increases (Bez et al. 2003).

Its origin can be tracked to Masuoka and IIzuka's 1985 patent (Masuoka and Iizuka 1985). The technology was seen to overcome significant challenges of EPROM, the dominant technology at the time, as it is much more reliable. Soon after its invention, several publications expressed the expectation that flash memories would be rapidly adopted (Lineback 1988; Cole 1988). The technology faced reliability issues that kept market penetration relatively low (Pavan et al. 1997). Early market predictions were modest, keeping expectations regarding flash memory relatively limited. In the mid-1990s, flash memory was expected to occupy a market share of 6% by 2000. Since then, two dominant flash architectures have emerged, as follows: NOR flash, designed for code and data storage, and NAND flash for data storage (Bez et al. 2003). These developments changed the the game for flash memory.

The abilities of NOR and NAND flash memory increased the market size of the technology. Flash memory is currently used, for example, in solid-state disks, which take advantage of its small dimensions, low power consumption, and lack of mobile parts. With an increase in the number of applications exploiting the benefits of the technology, flash memory's market share has increased. Since 2000, flash memory technology has been seen as a mature technology, which has increased rapidly in market size (Bez et al. 2003).

Whereas SSL technology seems to be the end of the line for lighting technology (Haitz and Tsao 2011), this is not the case for flash memory. With the ever-increasing demand for data storage, the sizes of flash memories have increased. Unfortunately, the size of a flash memory has been seen to correlate with an increase in latencies and data errors, creating a nearly unbreakable roadblock for the technology. Some suggest that the technology will be unable keep up with users' needs by 2024.

Several options have arisen to take the place of flash memory, within the family of Non-volatile memories. Among these are magnetoresistive random-access memory (MRAM), phase-change random-access memory (PRAM), silicon–oxide–nitride–oxide–silicon (SONOS) and resistive random-access memory (RRAM). Of these technologies, MRAMs represent a near-to-market technology that is being manufactured and applied in niche applications (Bhatti et al. 2017). PRAMs are also near to market, but they are also undergoing heavy research (Simpson et al. 2011), and they are expected to solve the key issues currently restricting flash memory technology. Finally, RRAM is at the demonstration phase. Research suggests that it will have a simpler structure than MRAMs, while it will be faster than PRAMs and offer smaller latencies and lower power consumption compared with flash memory (Zhang et al. 2013; Mittal et al. 2015; Lu et al. 2009).

Data collection and methodology

This study uses LED and flash memory technologies as two case studies. The aim is to examine the performance of three approaches for the detection of relevant topics or possible ETs. The scientific publication data related to LED and flash memory were gathered via a Boolean search algorithm from the Web of Science (WOS) core collection database in December 2017. The search query for LED technology was formed on variations of LED,[2] while that for flash memory[3] was based on keywords. The search queries were applied to titles, keywords, and abstracts of publications in the English language. The timespan of the search query was through 2016, without any restriction on the starting publication year. The searches returned 65,222 and 13,507 abstract records related to LED and flash technology, respectively. The retrieved data were imported into VantagePoint software (www.theVantagePoint.com) to consolidate duplications and author name variations, as well as to prepare the data for running the EScore algorithm. After cleaning, especially in terms of eliminating non-flash-related biomedical papers, the flash dataset was reduced from 13,507 to 10,968 records.

Central to the evaluation of the selected methods is to understand the cases at hand. The previous sections gave a brief introduction to the technologies and the emergent factors expected to occur in the data. Evaluation of the different methods is done using a qualitative approach relying on a framework of expected emergent behavior in the data. Each method is evaluated separately and compared on the basis of theoretical and managerial implications.

The evaluation framework is arrayed in Table 1. This table is used to identify, if an emergent topic is identified by the three utilized methods or not. For both LED and flash memory technologies, the Table 1 categories concepts and probable terms associated with these technologies. These concepts have been created based on the authors careful reading of recent publications in the case study domains and knowledge of the technologies. The probable terms have not been used as an exhaustive list, but terms identified to be emergent by different methods were evaluated individually by relying on the data (articles abstracts) downloaded and online searches, if needed. For LDA this meant looking a high probability terms in topics as a whole. Overall, the term mentioned guide the evaluation process to look for particular emergent patterns in the case technologies within the last ten years.

Emergence score algorithm

Previous text mining-based indicators designed for tracing ETs were formulated using citations or textual parts of documents. Indeed, these methods may have overlooked the role of other bibliometric variables, such as author-ship information. The emergence score (EScore) (Carley et al. 2018) used in this study traces ETs at a micro-level (technology level), incorporating authorship information and thresholds for time of publication. The EScore's functionality corresponds to the major attributes of ETs defined in the literature (Rotolo et al. 2015; Small et al. 2014; Day and Schoemaker 2000; Porter et al. 2002).

The EScore is a built-in module in the VantagePoint software program (Garner et al. 2017). The EScore algorithm is developed based on four major attributes of an ET, which

[2] TS = ((“light-emitting diode*” OR “light emitting diode*”) OR (“diode*” AND “LED*”).

[3] TS = (“flash memor*” or “memory cell*”).

Table 1 LED and FLASH technology emergent categories and a non-exhaustive list of concepts

Technology	Category	Concepts
LED	Efficiency	Thermal, heat, management, dissipation, efficiency loss and different chemical compounds, spectrum
	Sensors and actuators	Sensors, internet of things, actuators
	Communication	Optic, sensor, transimission, communication, transmission speed, transmission efficiency
	Purification	UV range leds, water
FLASH	MRAM	Magnetic, ferromagnetic, magnetoresistive effect, spin-transfer torque (STT), thermal assisted switching
	PRAM	PCM, PCME, PRAM, PCRAM, OUM, C-RAM, CRAM, chalcogenide, phase-change, memristor
	RRAM	ReRAM, memristor, dielectric
	SONOS	Silicon nitride, SONOS, MONOS, charge trap flash

are derived from the FUSE project. The first three are: novelty, persistence, and growth. The concept of novelty pertains to originality, discontinuous innovations (Day and Schoemaker 2000) or putting an existing technology to a new use (Adner and Levinthal 2002) or putting an existing technology to a new use (Rehurek and Sojka 2010). We acknowledge the difficulty inherent to predicting a concept which does not yet fully exist. Measuring the degree to which novel concepts have emerged within a given domain from a previous time period, however, is feasible and useful for purposes of forecasting. The concept of persistence considers a concept's ability to endure over time. To operationalize the persistency attribute, the EScore algorithm considers terms that occur in a minimum of seven records published over at least 3 years; thus, the persistency measure ensures that the term is not a one time hit. The growth concept is concerned with increase over time. This measure focuses on the change rates of terminology with the mentioned controls. To assure the novelty and growth characteristics, the ratio of records containing the term in the active period to those in the base period must be at least 2:1.

The last defining attribute of emergence is community, which is explained as coherence by Rotolo and colleagues (2015). The concept of community suggests that a number of professional researchers coalescing on a common topic and connections among them are necessary while an emergent technology is evolving. The EScore algorithm selects those terms associated with more than one author who does not share the same record set. In this study, the EScore serves as a metadata base emergence indicator. Why are the preceding four criteria emphasized? We reason that, to be truly emergent, a concept must embody all of these attributes (Rotolo et al. 2015). Making use of the preceding criteria not only ensures strong growth, but favors such growth continuing into subsequent time periods accompanied by a network of scholars cognizant of one another.

Text preprocessing for count based analysis and LDA

The preprocessing for the LDA and term count based analysis was implemented in the Python programming language using existing software packages to clean the textual data. First, the datasets were processed using the Part-of-Speech tagging implemented in the NLTK package. The abstracts of the documents were analyzed and only the tags NN (noun, common, singular or mass), NNP (noun, proper, singular), and JJ (adjective or numeral, ordinal) were kept in the analysis. Thereafter, the abstracts were analyzed to remove terms containing numbers, stopwords, and punctuation. The size of the token bag for LED is 94,885 unique tokens and for flash 21,320 unique tokens.

Count and TF-IDF weight based analysis

Term count based analysis is used as the most elementary approach to analyzing emergence using terms as the unit of analysis. The analysis was implemented using the preprocessed abstract and calculating Δ for all word tokens w appearing in year t and $t-1$. For each w Δ is word frequency $f_t - f_{t-1}$. Words that do not appear at two consecutive years are excluded.

Similar calculation is done using term frequency and inverse document frequency (TF-IDF) weighting scheme. Prior to calculating Δ for each w the frequency calculation were transformed to TF-IDF score weights using the Gensim package (Rehurek and Sojka 2010) in Python. After calculating the Δ values for LED and flash token bags yearly, the tokens were sorted by to highlight the highest Δ values.

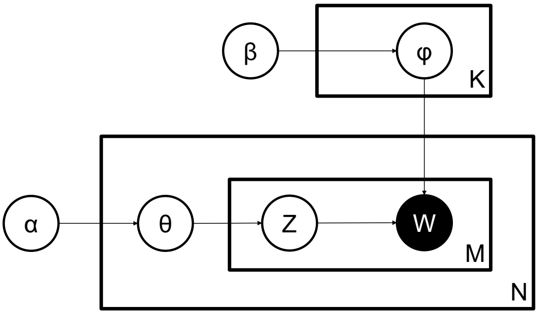


Fig. 1 Graphical representation of the Latent Dirichelt Algorithm (LDA). Square boxes represent sampling done as many times as given by the value of *K*, *M* and *N* Adopted from Blei et al. (2003)

LDA procedure

LDA is a probabilistic model where each document in a corpus is decribed by a random mixture over latent topics. Each of the latent topics is characterized by a distribution over words. Figure 1 shows the plate diagram of LDA, where

- K is the number of topics
- M is the number of documents
- N is the number of words in the document
- α is the parameter of the Dirichlet prior on the document topic distributions,
- β is the parameter of the Dirichlet prior on the topic word distribution
- θ_i the topic distribution for document i
- φ_k is the word distribution for topic k
- z_{ij} is the topic assignment for w_{ij}
- w_{ij} is the word

Each topic is a multinomial distribution over the vocabulary, thus the topics are described through high probability words in each topic. Document also have probability to be associated with topics. LDA is decomposed to two parts, the distributions over words and the distributions over topics. The unsupervised process balances two goals, allocating documents words to as few topics as possible and for topics assign high probability to as few terms as possible.

The LDA algorithm was implemented in Python using an online variational Bayes algorithm (Rehurek and Sojka 2010). The algorithm goes through the input tokenized data in chunks, updating the model as new data are analyzed, 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 it requires the user to set the number of topics produced as an output. Selecting a practical number of topics has been discussed in the literature. Chang et al. (2009) looked for a trial-and-error method, testing different numbers of topics with given input data to produce the results that would be most convenient for human interpretation. Furthermore, research has shown a number of other mathematical approaches, such as using Kullback–Leibler (KL) divergence (Arun et al. 2010) to estimate the input. In this study, we implemented KL divergence to estimate the number of topics in the corpus.

For the datasets, the KL divergence was calculated for topic values ranging from 1 to 100. The upper bound was set somewhat arbitrarily based on the experience of running an analysis with different corpus sizes. Figures 2 and 3 show the plot of values returned by the KL divergence function . Estimating the number of topics requires human intervention, as

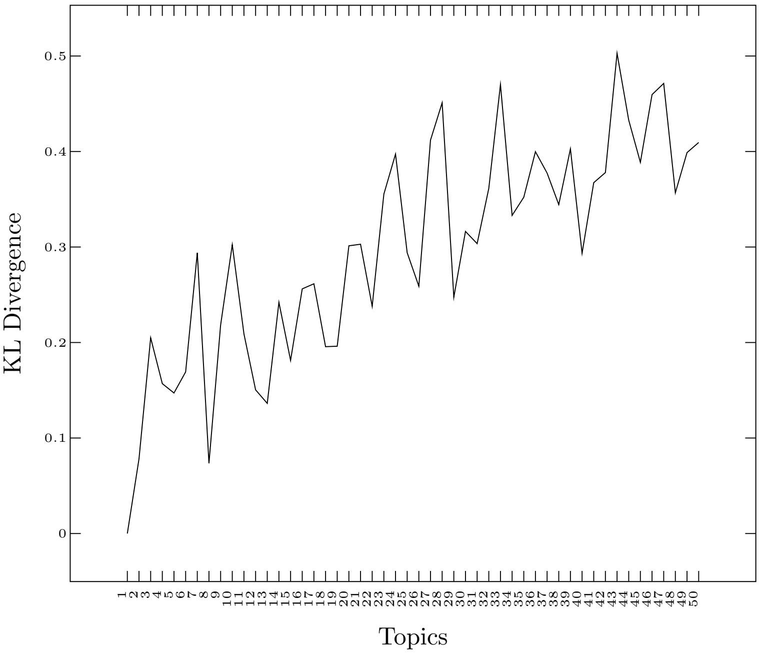


Fig. 2 KL divergence for LED dataset

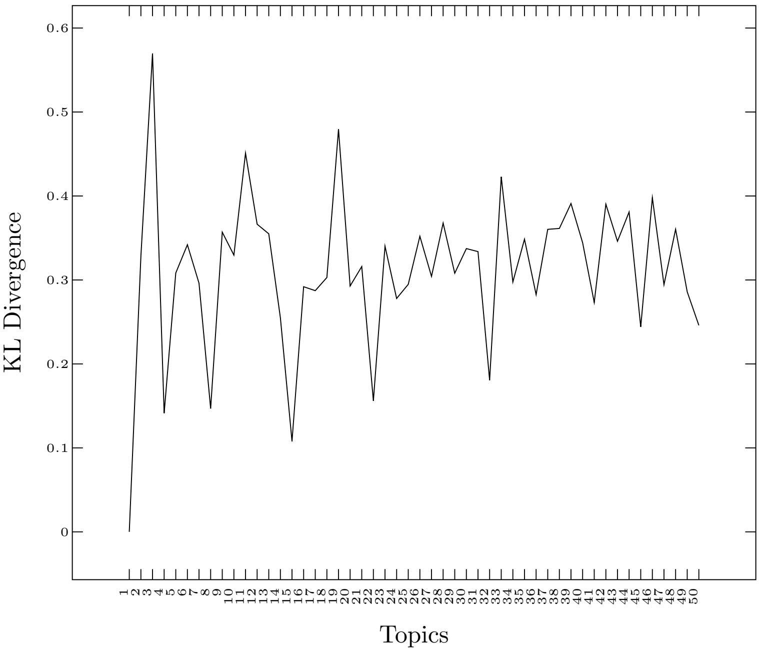


Fig. 3 KL divergence for flash memory dataset

simply taking the smallest value of the series is not sufficient. Even if the researcher has a relatively narrow window of expected topics, automating the evaluation of a KL divergence plot can be challenging. In the case at hand, the number of topics selected for the analysis was 15 for flash memory and 8 topics for LEDs. This selection was based on the sharp value decrease of the KL divergence and subsequent testing of different topic parameters.

LDA creates two matrices, one for document probabilities and one for word probabilities. The former matrix contains the probability distribution of each record that belongs to one of the topics, while the latter contains the probability distribution of words in the corpus and their association with each topic. 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 is defined by the probability of the document belonging to a certain topic. The word probability distributions were used to create word clouds to evaluate the content of the generated topics. The top 50 words were used to create the word cloud. In the word cloud visualization, the size of the words is based on their probability in a topic. The content and coherence of the topics are also evaluated using the word cloud plots. The assessment of the topical coherence is done by, first evaluating how concentrated the topics are to have high probability in only one or a few words. Second the topics are screened against expert human judgment to evaluate how semantically "cohesive" the topics are (Chang et al. 2009).

The LDA framework does not include a temporal constraint. The algorithm is unaware if the documents in the corpus are spaced differently in time. Thus, information on each document's year of publication is included in the results ex post. Using the publication year of each record and the document topic probability matrix, the results are aggregated to a year to topic matrix, A, where Aij is the sum of probabilities of year i records over latent topic j. Finally, the year to topic matrix is aggregated using the soft 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 to find potentially emerging topics.

The LDA soft classification is used to create a network representation. The result can be defined as an undirected bipartite graph. In a bipartite graph, nodes can be divided into two disjoint sets, U and V. Within the disjoint sets, each edge connects a node in U to one or more nodes in V. By definition, a bipartite graph is a graph that does not contain odd-length cycles. Focusing on the documents to topic probabilities produced by LDA, we can define $G = (U, V, E)$, where U represents documents, V topics, and E document topic probabilities. Focusing specifically on communities and the interaction between either set, we can use existing network analysis methods to transform the network to monopartite (one-mode) projection and find communities in either the bi- or monopartite graph.

Results

Emergence score results

The results of the EScore algorithm for flash memory data are presented in Table 2. The top 10 emerging terms are illustrated in the first column. The keyword solid state drive (SSD) is at the top of the list. SSD is an application of the flash memory cell, and it is currently fully commercial. Arguably, the term appears here due to the overall increase and

Table 2 Emergence score results for flash memory

	Emerging technologies	EScores	Top organisations	Top countries	Top authors
1	Solid-state drive	6.043	Chuo Univ	China	Takeuchi, Ken
2	Rank modulation	5.651	Ben Gurion Univ Negev	USA	Schwartz, Moshe
3	Codes	5.01	CALTECH	Israel	Tanakamaru, Shuhei
4	Phase-change memory	5.004	Univ Tokyo	South Korea	Sun, Chao
5	RRAM	4.691	Technion Israel Inst Technol	Japan	Bruck, Jehoshua
6	Memristor	3.624	Chongqing Univ	Taiwan	Yamazaki, Senju
7	SRAM cell	3.467	Univ Sci and Technol China	India	Dolecek, Lara
8	Flash-memory	3.466	Seoul Natl Univ	Singapore	Zuolo, Lorenzo
9	Error-correction	3.214	Texas A&M Univ	Canada	Matsui, Chihiro
10	Parity-check codes	2.96	Univ Calif Los Angeles	Iran	Tokutomi, Tsukasa

Table 3 Emergence score results for LED dataset

	Emerging technologies	EScores	Top organisations	Top countries	Top authors
1	Visible light communication	26.275	Chinese Acad Sci	China	Cao, Renping
2	Organometal halide perovskite	18.002	Pukyong Natl Univ	South Korea	Jeong, Jung Hyun
3	Sensitized solar-cell	17.884	Univ Chinese Acad Sci	India	Rajbhandari, Sujan
4	Delay fluorescence	17.881	Jinggangshan Univ	USA	Ghassemlooy, Zabib
5	Eu3+ ion	12.282	China Univ Geosci	Taiwan	Lin, Jun
6	Graphene	11.992	Northumbria Univ	UK	Shang, Mengmeng
7	Perovskite	10.645	Soochow Univ	Japan	Luo, Zhiyang
8	Phosphorescent OLED	10.552	Jilin Univ	Singapore	Yu, Xiaoguang
9	Perovskite solar cell	10.091	Hebei Univ	Germany	Haas, Harald
10	Eu3+ phosphor	9.91	Tsinghua Univ	France	Haigh, Paul Anthony

terminological cohesion in memory technologies. More interesting are the new technologies, such as the PRAM and RRAM, that are also highlighted in the results. This suggests that EScores can pinpoint emergent patterns. However, as relatively well known and common terms also appear, the analyst is required to handpick novel terms. The top organizations active in the mentioned emerging fields, for instance, are Chuo University, Ben Gurion University, Caltech, and the University of Tokyo. The top countries in order of EScores are China, the United States, Israel, and South Korea. The top three researchers are Ken Takeuchi, Moshe Schwartz, and Shubei Tanakamaru.

The implementation of the EScore algorithm for the LED dataset provided the results shown in Table 3. Several word combinations appears at the top of the list; for instance "visible light communication" (VLC) is a data transmission system that uses LED for communication rather than illumination. The idea behind building VLC systems using LEDs is to transmit data as fast as the speed of light. The second emerging term is "organometal halide perovskite," which was used for the fabrication of bright LEDs in a 2014 report in *Nature* (Tan et al. 2014). Here, organometal halide perovskite was used instead of the old direct-bandgap semiconductors, which are not economical for use in large-area displays. “Graphene,” the sixth emerging term, allows for tuning the color of LED light (Wang et al. 2015). The next emerging term is “phosphorescent organic LED,” which is still a technology under research and development. The top three active participants in ETs in the LED domain are the Chinese Academy of Science, Pukyong National University, and the University of the Chinese Academy of Science. The top countries are China, South Korea, India and the United States. According to the EScores, the top authors are Remping Cao, Jung Hyun Jeong, and Sujan Rajbhandari.

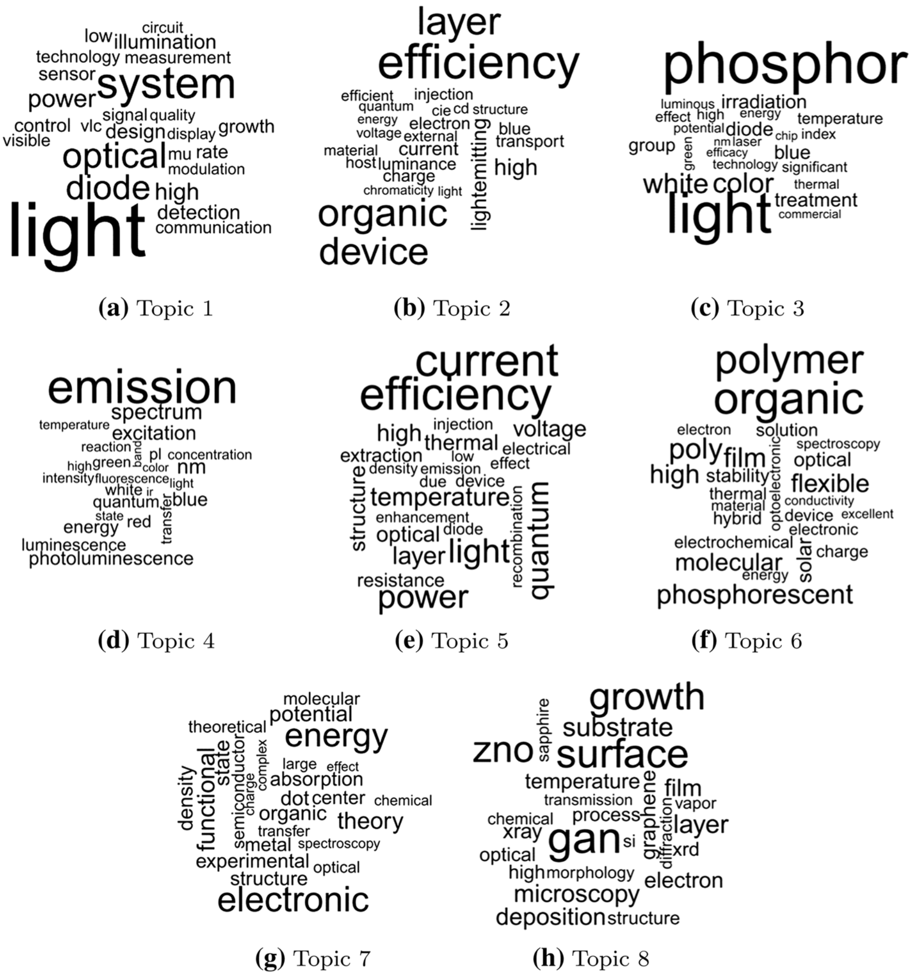
Term count based results

The analysis of LED and flash memory data produced two tables for both technologies (See “Appendix”), one with simple term count values and one with values based on TF-IDF weighting. For flash memory, the results for TF-IDF (Table 5) weighted and count based are very different. For TF-IDF results, the table highlights several technical aspects of flash technology that could be regarded as emerging at their time. For example between 2006 and 2007 “tcnq”, used as silver–tetracyanoquinodimethane, has been noted in the development of novel memory types (Zhang et al. 2014). Similar examples can be highlighted for several years such as “nanocrystal based” (Chang et al. 2011), “SiNx”, “Fullerene” (Ko et al. 2014) among other detailed material options that have been reported to change flash memory technology. The TF-IDF weighted results also have a number of references to new technologies, such as STT RAM or Spintronics referring to spin-transfer-torque memories (Nanotechnology 2015) and silicon–oxide–nitride–oxide–silicon (SONOS) memory cell (Lu et al. 2009), error correction issues such as “RTN” (Random Telegraph Noise) (Fukuda et al. 2007) or “SEUs” (Single-Event Upset) (Schwartz et al. 1997). The results for the TF-IDF weighted Δ values show clear emergent terms with significant detail.

This is not the case for the count based values. Seen in Table 6, the count based values only produce general terms like “memory”, “flash” or “voltage”. In addition to the general terms, the count based values yield some terms that could be regarded as stop-words. A notable exception to the general terms are SEUs, also seen in the TF-IDF weighted result.

For LEDs the case is similar to the results of the flash memory analysis. Using TF-IDF weighting the results show the individual materials used for enhancing of LED performance. Examples in Table 7 are ZnTe (Tanaka et al. 2009), quinacridone (Liu et al. 2008), and Organoboron (Nagai and Chujo 2010). In addition to materials that would yield better efficiency, the table contains multiple LED components enabling better performance, such as misorientation in LED manufacturing (Nakamura et al. 2008) or microball lenses (Kim et al. 2013). Table 7 also highlights application areas of LEDs, such as LEDs in microscopy (Bhadade et al. 2015) and communication (Farshchi et al. 2011).

Similar behavior as with the flash memory dataset (see Table 8), the count based analysis of the LED data remains at a high abstraction level . Although issues highlighted in the background as being emergent, such as “temperature” appears in the Table 8. Table 8 yields little value in identifying detailed emergent issues. It seems that a relatively simple rate of change based analysis yields relatively detailed results when using TF-IDF weighting. Obvious from the Table is that the count based and TF-IDF tables for either of the technologies share little terminology. Also, TF-IDF terms change significantly between the evaluated time slices. This raises some questions of the stability of the approach.



(a) Topic 1 **(b)** Topic 2 **(c)** Topic 3

(d) Topic 4 **(e)** Topic 5 **(f)** Topic 6

(g) Topic 7 **(h)** Topic 8

Fig. 4 Word clouds of eight topics created for LED technology using LDA

LDA results

The analysis of the LED dataset produced eight topics and a bipartite graph linking the documents to topics. Word clouds were used to describe the content of each node. The eight topics produced via LDA showed different thematic areas of research, described by the word clouds in Fig. 4. Topic 1 is related to the light-emitting system, design, and different relevant modules. Topic 3 focuses on phosphor, which is a key material related to LED performance. Phosphor's material, chemistry, and composition determine the efficiency, light quality, and stability of the LED light. Topic 4 focuses on the emission spectrum, that is, the spectrum of frequencies emitted by the LED. This is central to understanding of how an LED operates. Topic 5 focuses on the interplay current and efficiency of LEDs. When an LED's current increases, its efficiency drops; this is due to electron leakage, and it represents one of the major obstacles for creating long-lasting, and high–lumen output LEDs. The cause of this inefficiency was identified in the late 1990s and the solution was

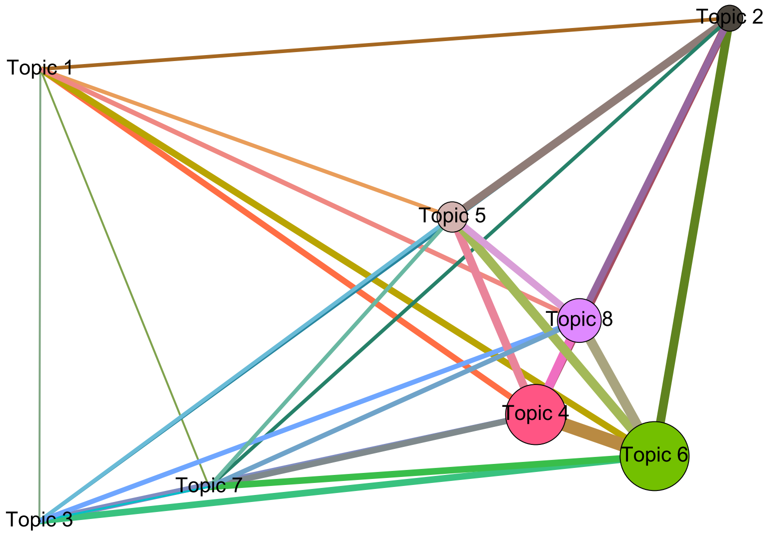


Fig. 5 Bipartite graph showing the relationship between the light-emitting diode (LED) topics. The graph consists of 56,985 nodes and 455,816 edges connecting the document nodes to topic nodes

only discovered in 2010s (Meyaard et al. 2013). Topics 2 and 6 look at a particular stream of LED research that focuses on organic LEDs (OLEDs). These are LEDs in which an organic compound is the source of the emission. In OLEDs, there is a specific polymer-based solution. The term “polymer” can also be seen in Topic 6 (4f). Both Topics 7 and 8 focus especially on the LED substrate. In these topics, we can identify two central materials, zinc oxide (ZnO) and gallium nitride (GaN), which are the core of making better white LEDs. The substrate material used is an ongoing research topic in the LED industry, and the search for a dominant solution continues.[4]

The soft classification of LDA was converted to a bipartite graph, embedding the metadata of publication year into each document node. This enabled visualization of an overlay of what has changed in the graph from overall to 2015 onward. The bipartite graph consists of 56,985 nodes and 455,816 edges connecting document nodes to topic nodes. The bipartite network was converted to the one mode, representing only linkages and sizes of topics. The one-mode transformation showed three central nodes, namely Topics 4, 6, and 8. These are large topics, but they are also heavily linked via shared probabilities among the documents. The network created is shown in Fig. 5. The one-mode network illustrates significant interest in OLED technology (Topic 6), analyzing its capabilities (Topic 4), and material options related to the technology (Topic 8).

Focusing on emergence, for each topic, we calculated its relative growth in importance. Thus, the relative share of growth in the topic probability assignments per year is seen in Fig. 6. Here, we identify the emergent pattern, as only two topics increase in relative importance, namely Topics 6 and 4. From these, the important topic is Topic 6, which focuses on OLED. This is to say, by using LDA, the emergent technological option highlighted is OLED technology.

[4] https://www.ledinside.com/showreport/2015/4/manufacturers_divided_about_gan_led_chip_technology.

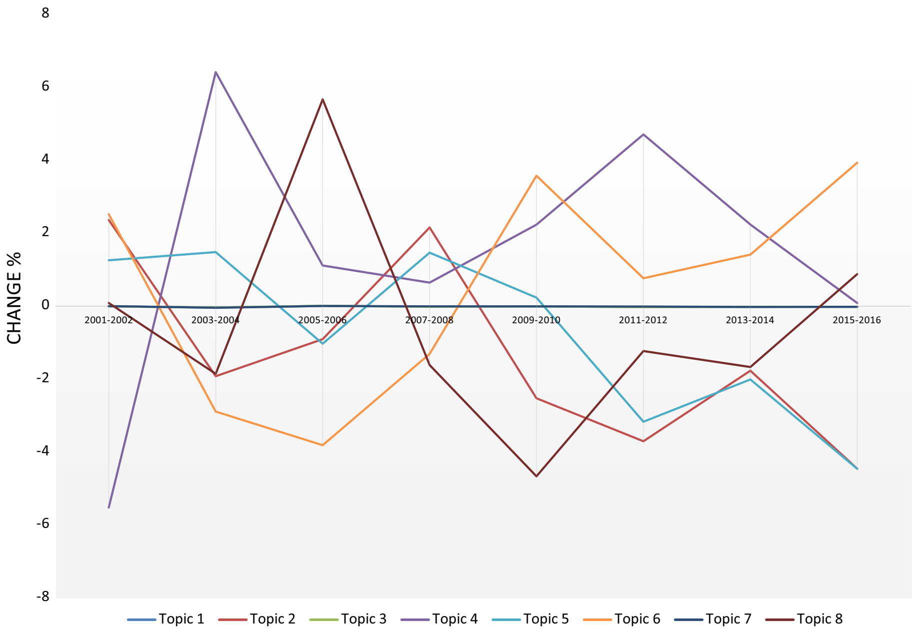
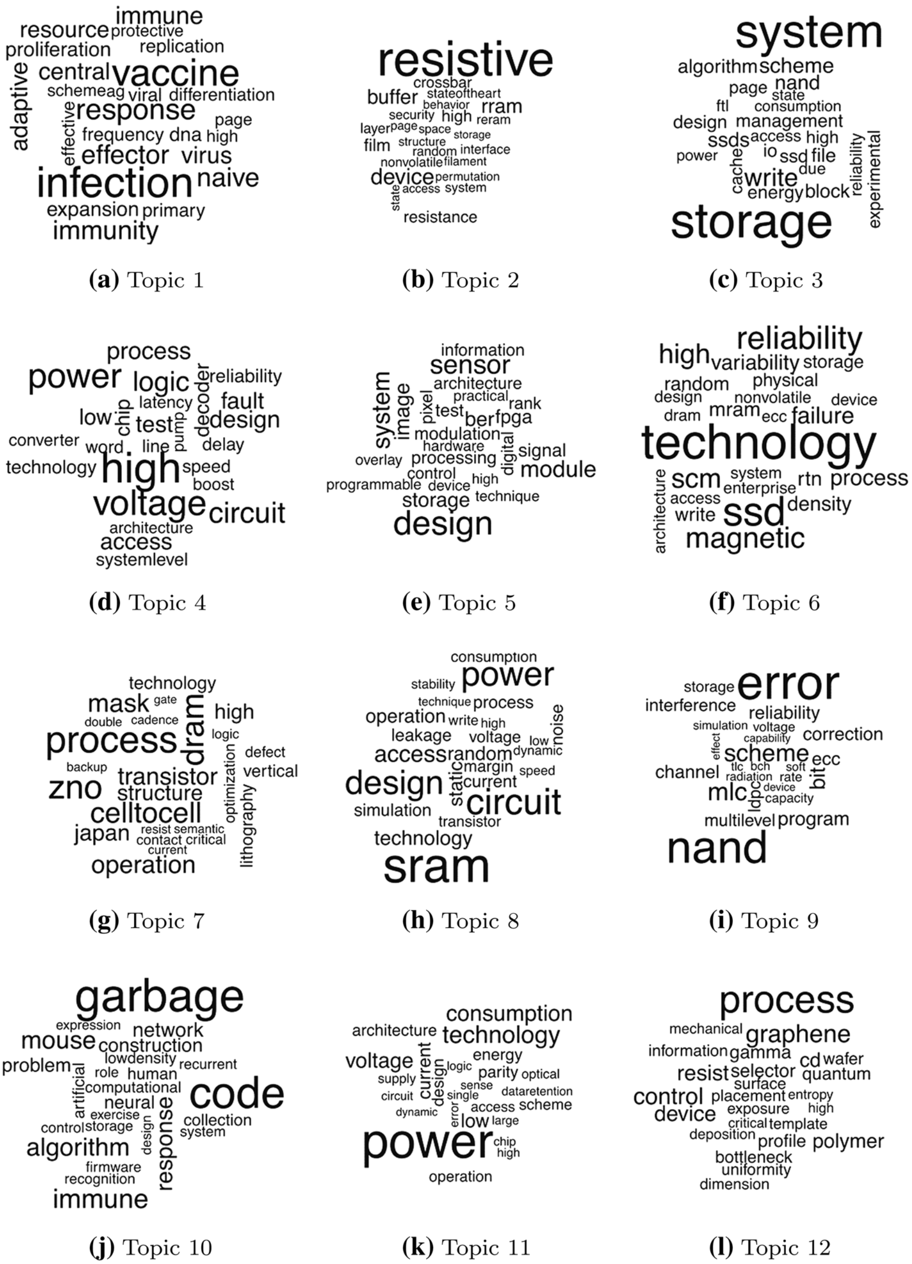


Fig. 6 Topical change of light-emitting diode (LED) technology, 2001–2016

For flash memory, the KL divergence value suggested that 15 latent topics could be derived from the dataset. When modeled using LDA, we identified clear patterns, as shown in Fig. 7. In the analysis, although concerted efforts were made to exclude medical research from the dataset, Topic 1 is an outlier relating mostly to medical technology. Topic 2 emphasizes the term "resistive, rram," which refers to the new generation of memory technology known as resistive random access memory (RRAM). RRAM is known as a breakthrough NVM technology and the most promising candidate for the next generation of memory (Chang et al. 2016). RRAM has a very low operation voltage, while it performs faster than previous generations, with higher reliability (Chang et al. 2016). It has significant potential for commercialization in the future (Chang et al. 2016). Topic 15, with the top words of "change, resistance, phase," is related to Phase-change random access memory (PCRAM) technology, which may also have a promising future (Loke et al. 2012). Topic 3 is related to the area of SSD configuration and power management. SSD is a storage device that includes an integrated circuit assembly that stores the data. The top words in Topics 4, 11, and 14 are "power, high, voltage, consumption, charge, device." These topics cover the research about the settings of voltage thresholds and power management, which play an important role in flash performance (Cai et al. 2013). In fact, the voltage threshold approach can influence the performance and reliability of flash memory.

Topic 5 (see Fig. 7), with top terms like "sensor, design, architecture, system," can be associated with the flash memory storage system used in modern wireless sensor devices. Flash memory has become a prevalent storage medium for sensor devices because of its beneficial features, such as non-volatility, small size, light weight, fast access speed, shock resistance, high reliability, and low power consumption (Rizvi and Chung 2010). Topic 7 highlights terms like "process, zno, DRAM," which relate to a new generation of NVM



(a) Topic 1 **(b)** Topic 2 **(c)** Topic 3

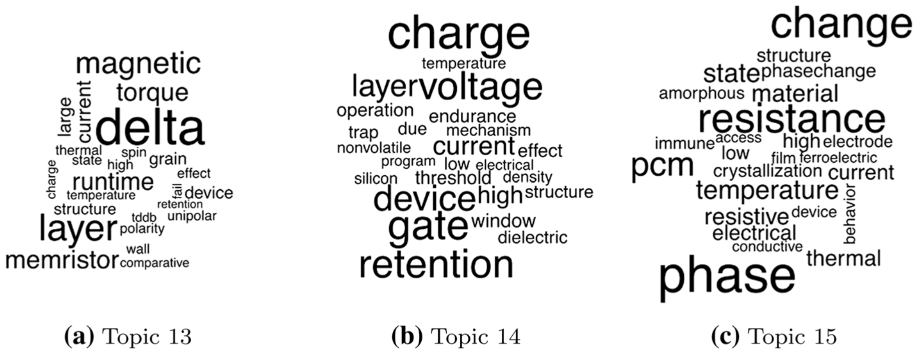
(d) Topic 4 **(e)** Topic 5 **(f)** Topic 6

(g) Topic 7 **(h)** Topic 8 **(i)** Topic 9

(j) Topic 10 **(k)** Topic 11 **(l)** Topic 12

Fig. 7 Word clouds of 15 on flash memory topics created using LDA. Topic 1–12

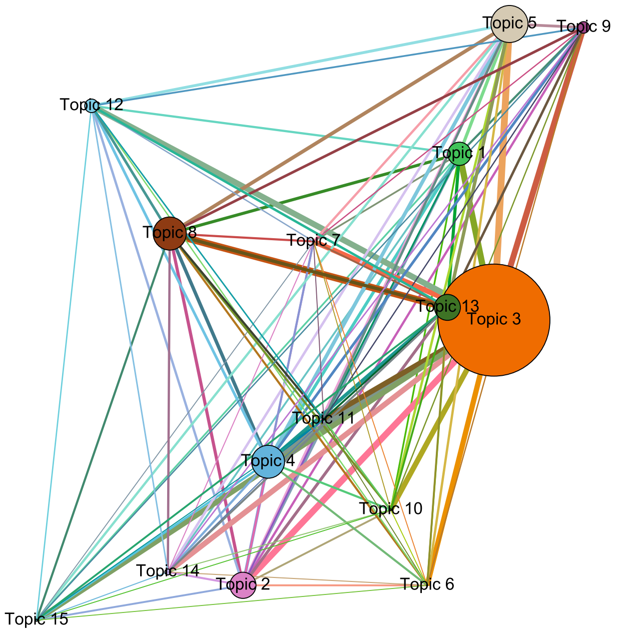
known as zinc-oxide (ZnO) charge trapping memory cell flash technology (Simanjuntak et al. 2016). Topic 8 is about static random-access memory (SRAM), which is a type of a memory used in a computer's cache memory. Topic 9 corresponds to NAND flash memory, while Topic 10 is about the garbage collection (GC) process in NAND flash memory,



(a) Topic 13 **(b)** Topic 14 **(c)** Topic 15

Fig. 8 Word clouds of 15 on flash memory topics created using LDA. Topic 13–15

Fig. 9 Bipartite graph showing the relationships between the flash memory topics. The graph consists of 10,734 nodes and 160,785 edges connecting document nodes to topic nodes



which secures the free space prior to writing new data. Topic 12 demonstrates recent research on graphene. In terms of flash memory, graphene possesses unique properties, such as a high density of states, high work function, and low dimensionality, which can enhance flash memory performance (Hong et al. 2011).

Topic 13 (see Fig. 8) is represented with top terms like “delta, magnetic, layer, memristor,” representing a mixture of major components or operation processes in flash memory (e.g., oxide layer) and memristors. The word memristor is often used as a synonym for resistive RAM (ReRAM or RRAM). Research (Zidan et al. 2013) suggests that memristors are expected to replace the NAND flash memory in future.

The flash memory topic bipartite network was transformed into a one-mode network to illustrate the relationship between the topics. The network graphs in Fig. 9 show one large

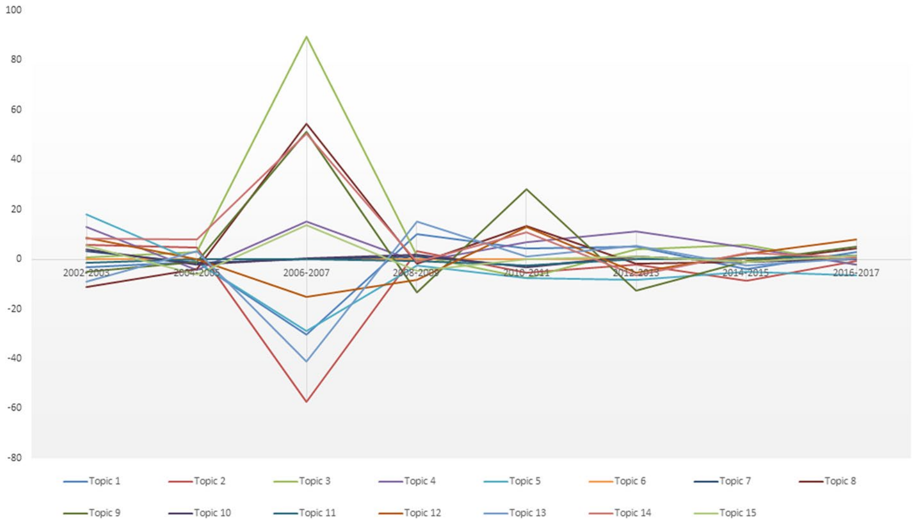


Fig. 10 Topical change of flash memory technology, 2002–2017

node, Topic 3, with 14 more equally sized nodes. The strong topical linkage between Topics 3, 4, 11 and 13 is due to the similarity in their themes. The focuses of these topics are flash memory components, modes of operation, power management, memory architecture, or data storage configuration.

Highlighting emergence in Fig. 10, we see several rapidly changing topics. For flash memory, the importance of specific topics has been more dynamic. Of particular interest is the period of 2006–2007, in which several topics increased significantly, while others decreased in importance. For example, during this period, Topic 3, the largest topic by size , increased in importance, with significant reductions in importance of new-generation technologies, such as RRAM (Topic 2) or novel sensor applications (Topic 5). Turning our focus to recent emergent applications, we highlight Topic 12, which shows consistent, increasing growth in the two latest time slices. Topic 12 highlights graphene research, and especially, the application of graphene to form graphene flash memory. This new material option has shown great promise in increasing the performance of flash memory (e.g., Hong et al. 2011).

Comparison of results from selected methods

Table 4 shows the novel concepts and emerging terms in flash memory and LED technology detected by selected three methods (EScore, LDA, count based). For flash memory, EScore and LDA identified emerging PCRAM and RRAM field of research. In addition, LDA was able to show three more active research area about application of flash memory in wireless sensor devices, zinc oxide charge trapping memory cell and utilization of graphene for development of new flash memories. The terms detected by TF-IDF method are at detailed layers related to technical terminologies of the fabrication materials for enhancing the flash memory performance. For instance TCNQ, Nanocrystal based, SiNx and Fullerene. The occurrence of these terminologies is associated to a certain time period

Table 4 The comparison of detected emerging concepts and research topics by four selected methods (E-score, LDA, TF-IDF and Term count) within flash memory and LED research area

Emerging concepts	E-score	LDA	TF-IDF	Term count
Flash memory technology				
1	Phase change random-access memory (PCRM)	Phase change random-access memory (PCRM)	Novel material options for development of memory cells: tetracyanoquinodimethane (TCNQ), nanocrystal based memory cell, SiNx, Fullerene	–
2	Resistive random-access memory (RRAM, memristors)	Resistive random-access memory (RRAM, memristors)	Spin transfer torque memories	–
3		Zinc-oxide charge trapping memory cell	SONOS memory cell	–
4		Graphene (for development of flash memory)		–
5		Novel application area for flash memory in wireless sensor device		–
LED technology				
1	Phosphorescent organic (OLED)	Phosphorescent organic (OLED)	Novel material for enhancing LED performance: ZnTe, quinacridone, organoborn	–
2	Graphene	Graphene (appeared in topic 8)	Components: microball lens	–
3	Visible light communication (VLC)	Visible light communication (VLC) (appeared in topic 1)		–
4	Organometal halide perovstike	Novel material for LED fabrication: ZnO, GaN (appeared in topic 8)		–

presented in Table 5. Moreover, spin transfer torque and SONOS memory cells were identified by TF-IDF as novel technological alternatives for flash memory in their corresponding time. The concepts and keywords detected by term count method were rather generic and not useful.

In case of LED technology (see Table 4), the LDA method seems to cover more recent emerging fields comparing to EScore and TF-IDF. EScore and LDA methods detected research development in phosphorescent organic OLED, VLC as a new application area for LED and usage of graphene for light tuning purposes. Only LDA algorithm identified the ongoing field of research in ZnO and GaN. In addition, EScore results shows research interested in organometal halide perovstike, which is an element that can be used for fabrication of LED for large scale displays. TF-IDF method presented particular terms related to novel materials or components for manufacturing of LED such as; ZnTe, quanacridone, organobron and microball lens.

Discussion

The major objective of this paper is to demonstrate and compare different approaches in detecting ETs within a scientific publication dataset. We examine the EScore indicator, unsupervised machine learning algorithms known as LDA and term count based methods. The EScore can be considered a two-stage indicator that involves several parameters (country, author information, keyword or abstract, year). It scores terms on their degree of emergence. Then, secondary indicators are composed to reflect the extent to which countries, organizations, or authors are publishing most actively using the emergent terms in their abstract records. The EScore algorithm's functionality is based on the recent theoretical definition of ET and its related attributes (Rotolo et al. 2015). The unsupervised machine learning method is a generative model, and it uncovers latent patterns in the textual data. Relying on the frequency of keywords occurring each years, the term count based methods aim to identify technological novelty and ostensivity. The interpretation of the captured pattern to be translated to an ET depends on expert validation.

The main finding of the study is that all three methods used, with the exception of term counting without TF-IDF, are able to track emergent aspects in the two case technologies. The straight term counting was limited by high frequency terms in the corpus, making a strong argument for the use of TF-IDF to understand the dynamism of technological vocabulary across time. Comparing the other methods, it is difficult to argue the superiority of one of these methods over the others. Our results highlight that measuring superiority would probably be an impractical endeavour. The approaches are able to track emergence, but are suitable for different use cases. Although our analysis did not look at statistical overlap between clusters, as in for example (Velden et al. 2017; Waltman et al. 2019), the in depth case studies allow us to propose different use cases for the methods. First, the EScore algorithm uses a metadata enriched approach to identify emergence. It links actors to the emergent terms allowing for the results to be informed on the actors, not only terms. However, we did notice that the results were crowded with general terms making identification of emergent terms require expert opinion. The TF-IDF term counting operates significantly different, highlighting emergent terms at an extremely detailed level. The results suggests that TF-IDF based term detection could be operationalized to inform an emergent area of research, but would require a knowledgeable expert to take full use of the detailed information or additional measures to inform the analysis of context. The LDA approach produces meaningful, large-scale topical changes and highlights the thematic concepts of the dataset. It also shows the topical change at the document level, as well as topical linkage. These results differ from the EScore or term based approach in that the results are on a totally different abstraction level, offering a strategic level view with broad conceptual changes in the technological landscape. With the word clouds, network visualizations and time series, LDA analysis offers a practical view to theme importance allowing for example being informed on recent trends, but lacking the ability to act on the details.

Discussing the pros and cons of the methods in more detail, one of the prominent advantage of LDA compared to EScore or count based methods is its focus on the context rather than keyword counts. LDA results provided the central topics of research for LEDs and flash memory technology. Each identified topic created is based on a distribution of keywords. The interpretation for the appearance of each term within the topic can be done based on its neighboring keywords. For instance, the “Graphene” research topic has been detected by both EScore and LDA within LED technology datase. Based on the EScore table it is difficult to understand how Graphene is related to LED research. LDA makes interpretation easy by providing context for “Graphene” in Topic word probabilities. Topic 8 describes novel materials and chemical compounds used for fabrication of LED, such as ZnO, GaN and also Graphene. LDA detects the central research topics and, in detail, what is the role of Graphene in LED research. This can be produced directly from by the approach, without extra input from the user.

EScore’s strengths come from its ability to limit the impact of outliers. Authors are not always consistent in using similar terminologies or different terms can be used to describe the same phenomena. The development of terminological consensus can take time. While the EScore is flexible in assigning a term frequency threshold, one of the main ideas behind its design is focusing on persistent concepts rather than a one-time outburst. According to Kuhn (1970) , in the development process of any new science, there is progress on the emergence of esoteric vocabulary and skills, and a refinement of concepts that increasingly lessens their resemblance to their usual common-sense prototypes. The methods shown here have very different capabilities in identifying novelties. The TD-IDF term count approach is sensitive to small shifts. Results seem unstable as the highest Δ terms seem to appear and disappear on a yearly basis. The control parameters offered by the EScore approach allow for stability without limiting the detection of novelties. The drawback of stability is that we are unable to detect the small changes and details, the strength of TF-IDF term counting.

The topical clusters and terms generated by LDA are purely based on documents’ language, and they are independent of an article’s authors, country, and citation information. This independence has pros and cons. The advantage of being independent of the citation and connection information among authors is that the phenomena of emergence or niche research will not be limited to a certain circle of the scientific community. Drawing from Kuhn’s work (1970), scientific knowledge is represented by vocabularies; thus, tracing the change within the language used by authors may lead to detecting shifts in their ideas. LDA is a tool that considers the language that represents old or new ideas. At the same time, the machine learning method used in this study overlooked the role of the research community, which is defined as one of the major attributes of ET (Rotolo et al. 2015).

In summary, the selection of a specific method hinges on the intended use of the results. If the user is interested in detecting a niche and ongoing area of research that has not yet become dominant knowledge in the scientific community. The sensitivity of the machine learning algorithm can be adjusted, such that it could be able to detect any topical outburst with smaller frequencies. For instance, if the model keeps the keywords with minimum occurrences in the preprocessing phase, the final results will include rare concepts/terms. The flexibility of setting preprocessing rules can expand the spectrum of the results by showing both new high-frequency terms and new rare terms. However, our results suggest that the extremely simple count based calculation is already extremely sensitive in tracking changing terminology.

This study is not without limitations. First, we only examined three methods in two technology cases. Adding more methods and case studies will surely provide more insight. Possible methodological addition could include, for example, one-class SVM, which has been used widely for outlier detection in computer science. For the case approach, LED and flash memory have the attributes of an interesting case study, manufacturing a controlled test dataset could serve a purpose. We think that a synthetic dataset could yield a more controlled sample where the future development is still unknown.

Acknowledgements Open access funding provided by Technical Research Centre of Finland (VTT). The Academy of Finland research supported this work with the Grant (288609) "Modeling Science and Technology Systems Through Massive Data Collections". We acknowledge support from the US National Science Foundation (Award #1759960—"Indicators of Technological Emergence") to Search Technology, Inc., and Georgia Tech. The findings and observations contained in this paper are those of the authors and do not necessarily reflect the views of the National Science Foundation.

Appendix: Results of term count based methods

See Tables 5, 6, 7 and 8

Table 5 Ten highest delta values for flash using TF-IDF weights between year 2006 and 2016

	1st	2nd	3rd	4th	5th	6th	7th	8th	9th	10th
2006–2007	hydrogen (0.52)	mev (0.49)	defective (0.41)	oxynitride (0.41)	tcnq (0.4)	tip (0.4)	xray (0.37)	trap (0.37)	onchip (0.35)	amorphization (0.35)
2007–2008	intermittent (0.63)	dipole (0.62)	lambda (0.51)	bch (0.5)	delta (0.5)	speech (0.49)	mlcs (0.48)	tcells (0.46)	cmol (0.44)	alloptical (0.44)
2008–2009	polymer (0.58)	movement (0.54)	vsasfg (0.53)	cobalt (0.52)	rtn (0.52)	multigate (0.52)	wsn (0.47)	health (0.47)	dfm (0.47)	laser (0.46)
2009–2010	memristor (0.61)	zro (0.6)	coefficient (0.54)	sttram (0.54)	register (0.51)	deployment (0.5)	pathogen (0.5)	mbus (0.49)	injection (0.49)	status (0.48)
2010–2011	amorphous-crystalline (0.59)	tnf (0.56)	delta (0.54)	fea (0.52)	round (0.51)	swap (0.5)	victim (0.49)	mlc (0.48)	multilayer (0.47)	ti (0.46)
2011–2012	sb (0.65)	gamma (0.63)	fringe (0.61)	trim (0.6)	lcmv (0.56)	bidirectio-nal (0.55)	call (0.55)	ptype (0.54)	pzt (0.53)	nanostructure (0.53)
2012–2013	treg (0.72)	imd (0.7)	transaction (0.62)	nanolami-nate (0.6)	nucleation (0.58)	sneak (0.58)	nanocrystalbased (0.56)	transge-nic (0.55)	dcsf (0.54)	logger (0.54)
2013–2014	overlay (0.75)	ga (0.6)	unreliable (0.6)	coset (0.57)	sinx (0.55)	uv (0.54)	mobile (0.54)	nps (0.54)	seus (0.53)	fullerene (0.53)
2014–2015	nk (0.66)	fg (0.64)	backup (0.59)	sonos (0.57)	sorting (0.56)	sibased (0.55)	highrate (0.51)	cm (0.51)	approxima-te (0.51)	ncap (0.5)
2015–2016	stuckat (0.66)	sort (0.59)	uncertainty (0.58)	segment (0.58)	pico (0.57)	early (0.55)	video (0.55)	pwm (0.54)	spintronic (0.54)	tsv (0.54)
2016–2017	scheduler (0.58)	seus (0.55)	behaviour (0.53)	rtn (0.52)	crystallization (0.45)	mlc (0.45)	reram (0.45)	entropy (0.45)	accelerated (0.44)	gas (0.44)

Table 6 Ten highest delta values for flash memory using counts between year 2006 and 2016

	1st	2nd	3rd	4th	5th	6th	7th	8th	9th	10th
2006–2007	memory (786)	flash (400)	cell (317)	paper (178)	system (149)	technology (144)	performance (138)	device (135)	storage (127)	design (124)
2007–2008	memory (182)	cell (89)	flash (76)	energy (59)	read (57)	low (55)	nand (52)	structure (51)	simulation (50)	immune (43)
2008–2009	memory (552)	flash (352)	performance (179)	cell (177)	paper (138)	technology (123)	storage (111)	device (111)	write (101)	charge (99)
2009–2010	surface (21)	policy (19)	effector (18)	mouse (13)	ftl (12)	parallel (11)	overhead (11)	address (10)	mw (10)	protection (9)
2010–2011	memory (745)	flash (395)	cell (257)	device (162)	performance (141)	nand (117)	paper (114)	current (106)	layer (106)	storage (104)
2011– 2012	cell (129)	gate (81)	paper (66)	process (57)	error (52)	retention (51)	high (49)	voltage (48)	nand (45)	sram (39)
2012–2013	performance (62)	power (57)	logic (54)	storage (44)	ssds (41)	flash (38)	cmos (38)	access (35)	scheme (34)	ssd (34)
2013–2014	memory (298)	cell (176)	paper (87)	performance (80)	sram (73)	access (72)	write (66)	magnetic (58)	ssd (56)	storage (55)
2014–2015	flash (92)	rate (63)	technique (58)	error (51)	performance (43)	data (43)	ecc (40)	disk (39)	write (37)	memristor (36)
2015–2016	memory (596)	cell (326)	flash (181)	performance (170)	read (143)	device (128)	resistive (125)	channel (117)	operation (116)	power (108)

Table 7 Ten highest delta values for LED using TF-IDF weights between year 2006 and 2016

	1st	2nd	3rd	4th	5th	6th	7th	8th	9th	10th
2006–2007	hydrogen (0.52)	mev (0.49)	defective (0.41)	oxynitride (0.41)	tcnq (0.4)	tip (0.4)	xray (0.37)	trap (0.37)	onchip (0.35)	amorphization (0.35)
2007–2008	intermittent (0.63)	dipole (0.62)	lambda (0.51)	bch (0.5)	delta (0.5)	speech (0.49)	mlcs (0.48)	tcells (0.46)	cmol (0.44)	alloptical (0.44)
2008–2009	polymer (0.58)	movement (0.54)	vsasfg (0.53)	cobalt (0.52)	rtn (0.52)	multigate (0.52)	wsn (0.47)	health (0.47)	dfm (0.47)	laser (0.46)
2009–2010	memristor (0.61)	zro (0.6)	coefficient (0.54)	sttram (0.54)	register (0.51)	deployment (0.5)	pathogen (0.5)	mbus (0.49)	injection (0.49)	status (0.48)
2010–2011	amorphous crystalline (0.59)	tnf (0.56)	delta (0.54)	fea (0.52)	round (0.51)	swap (0.5)	victim (0.49)	mlc (0.48)	multilayer (0.47)	ti (0.46)
2011–2012	sb (0.65)	gamma (0.63)	fringe (0.61)	trim (0.6)	lcmv (0.56)	bidirectional (0.55)	call (0.55)	ptype (0.54)	pzt (0.53)	nano structure (0.53)
2012–2013	treg (0.72)	imd (0.7)	transaction (0.62)	nano laminate (0.6)	nucleation (0.58)	sneak (0.58)	nano crystal-based (0.56)	transgenic (0.55)	dcsf (0.54)	logger (0.54)
2013–2014	overlay (0.75)	ga (0.6)	unreliable (0.6)	coset (0.57)	sinx (0.55)	uv (0.54)	mobile (0.54)	nps (0.54)	seus (0.53)	fullerene (0.53)
2014–2015	nk (0.66)	fg (0.64)	backup (0.59)	sonos (0.57)	sorting (0.56)	sibased (0.55)	highrate (0.51)	cm (0.51)	approximate (0.51)	ncap (0.5)
2015–2016	stuckat (0.66)	sort (0.59)	uncertainty (0.58)	segment (0.58)	pico (0.57)	early (0.55)	video (0.55)	pwm (0.54)	spintronic (0.54)	tsv (0.54)

Table 8 Ten highest delta values for LED using counts between year 2006 and 2016

	1st	2nd	3rd	4th	5th	6th	7th	8th	9th	10th
2006–2007	high (296)	light (290)	current (275)	efficiency (274)	leds (199)	optical (195)	quantum (191)	power (142)	method (141)	organic (140)
2007–2008	light (385)	emission (262)	leds (245)	organic (223)	efficiency (198)	lightemitting (176)	surface (163)	spectrum (158)	high (146)	energy (144)
2008–2009	temperature (146)	structure (130)	phosphor (122)	emission (114)	polym (92)	excitation (91)	design (82)	transition (80)	applied (79)	physics (78)
2009–2010	light (336)	high (255)	phosphor (163)	layer (159)	organic (158)	optical (157)	leds (123)	current (122)	efficiency (120)	device (118)
2010–2011	light (534)	efficiency (342)	emission (325)	high (255)	energy (235)	blue (229)	spectrum (223)	method (221)	electron (203)	surface (194)
2011–2012	light (312)	optical (283)	power (245)	high (241)	efficiency (224)	emission (204)	leds (203)	temperature (191)	phosphor (190)	graphene (189)
2012–2013	light (407)	leds (307)	lightemitting (252)	high (221)	layer (217)	publishing (216)	energy (199)	diode (197)	method (182)	phosphor (175)
2013–2014	emission (416)	energy (331)	light (315)	blue (254)	device (214)	electron (182)	group (178)	phosphor (170)	structure (169)	potential (165)
2014–2015	performance (315)	fluorescence (252)	quantum (244)	high (223)	emission (200)	blue (191)	color (159)	flexible (157)	molecular (141)	study (141)
2015–2016	inactivation (13)	milk (11)	fabryperot (10)	ledpdt (9)	dinuclear (9)	persulfate (8)	agzn (8)	photoinduced (7)	guided (7)	tcta (7)

References

Adner, R., & Levinthal, D. A. (2002). The emergence of emerging technologies. *California Management Review*, *45*(1), 50.

Arun, R., Suresh, V., Madhavan, C. V., & Murthy, M. N. (2010). In *Pacific-Asia conference on knowledge discovery and data mining* (pp. 391–402). Berlin: Springer.

Ayres, R. (1969). *Technological forecasting and long-range planning*. New York: McGraw-Hill.

Bez, R., Camerlenghi, E., Modelli, A., & Visconti, A. (2003). Introduction to flash memory. *Proceedings of the IEEE*, *91*(4), 489.

Bhadade, A., Mehta, P., Kanade, S., & Nataraj, G. (2015). Utility of light-emitting diode microscopy for the diagnosis of pulmonary tuberculosis in HIV infected patients. *International Journal of Mycobacteriology*, *4*(1), 31.

Bhatti, S., Sbiaa, R., Hirohata, A., Ohno, H., Fukami, S., & Piramanayagam, S. (2017). Spintronics based random access memory: A review. *Materials Today*, *20*, 530–548.

Blei, D. M., Ng, A. Y., Jordan, M. I., & Lafferty, J. (2003). Latent dirichlet allocation. *Journal of Machine Learning Research*, *3*, 2003.

Boyack, K. W., Newman, D., Duhon, R. J., Klavans, R., Patek, M., Biberstine, J. R., et al. (2011). Clustering more than two million biomedical publications: Comparing the accuracies of nine text-based similarity approaches. *PLoS ONE*, *6*(3), e18029.

Breitzman, A., & Thomas, P. (2015). The emerging clusters model: A tool for identifying emerging technologies across multiple patent systems. *Research Policy*, *44*(1), 195.

Cai, Y., Haratsch, E. F., Mutlu, O., & Mai, K. (2013). Threshold voltage distribution in MLC NAND flash memory: Characterization, analysis, and modeling. *Design automation test in Europe conference exhibition* (pp. 1285–1290).

Carley, S., Newman, N., Porter, A., & Garner, J. (2018). An indicator of technical emergence. *Scientometrics*, *115*, 35.

Chang, J., Gerrish, S., & Wang, C. (2009). Reading tea leaves: How humans interpret topic models. *Advances in neural information processing systems*.

Chang, T. C., Chang, K. C., Tsai, T. M., Chu, T. J., & Sze, S. M. (2016). Resistance random access memory. *Materials Today*, *19*(5), 254.

Chang, T. C., Jian, F. Y., Chen, S. C., & Tsai, Y. T. (2011). Developments in nanocrystal memory. *Materials Today*, *14*(12), 608.

Cole, B. (1988). Flash-theres more than one road to dense nonvolatile memory. *Electronics*, *61*(18), 108.

Craford, M. G. (1997). Overview of device issues in high-brightness light-emitting diodes. *Semiconductors and Semimetals*, *48*, 47.

Criscuolo, P., & Verspagen, B. (2008). Does it matter where patent citations come from? Inventor vs. examiner citations in European patents. *Research Policy*, *37*(10), 1892.

Day, G. S., & Schoemaker, P. J. H. (2000). Avoiding the pitfalls of emerging technologies. *California Management Review*, *42*(2), 8.

Deerwester, S., Dumais, S. T., & Harshman, R. (1990). Indexing by latent semantic analysis. *Journal of the American Society for Information Science*, *41*(6), 391.

Farshchi, R., Ramsteiner, M., Herfort, J., Tahraoui, A., & Grahn, H. (2011). Optical communication of spin information between light emitting diodes. *Applied Physics Letters*, *98*(16), 162508.

Feenberg, A. (2010). Ten paradoxes of technology. *Techné: Research in Philosophy and Technology*, *14*(1), 3.

Feldman, R., & Sanger, J. (2006). *Text mining handbook: Advanced approaches in analyzing unstructured data*. Cambridge: Cambridge University Press.

Fukuda, K., Shimizu, Y., Amemiya, K., Kamoshida, M., & Hu, C. (2007). In *IEEE international electron devices meeting, IEDM* (pp. 169–172). IEEE.

Garner, J., Carley, S., Porter, A. L., & Newman, N. C. (2017). Technological emergence indicators using emergence scoring. In *Proceedings of PICMET'17: Technology management for interconnected world.*

Gerken, J. M., & Moehrle, M. G. (2012). A new instrument for technology monitoring: Novelty in patents measured by semantic patent analysis. *Scientometrics*, *91*(3), 645.

Goldstein, J. (1999). Emergence as a construct: History and issues. *Emergence*, *1*, 49–72.

Guo, Y., Xu, C., Huang, L., & Porter, A. (2012). Empirically informing a technology delivery system model for an emerging technology: Illustrated for dye-sensitized solar cells. *R&D Management*, *42*(2), 133.

Gustafsson, R., Kuusi, O., & Meyer, M. (2015). Examining open-endedness of expectations in emerging technological fields: The case of cellulosic ethanol. *Technological Forecasting and Social Change*, *91*, 179.

Haitz, R., & Tsao, J. Y. (2011). Solid-state lighting: ‘The case’10 years after and future prospects. *Physica Status Solidi (a)*, *208*(1), 17.

Hofmann, T. (1999). In *Proceedings of the 22nd annual international ACM SIGIR conference on research and development in information retrieval-SIGIR ’99* (pp. 50–57). New York: ACM Press.

Holland, J. H. (2000). *Emergence: From chaos to order*. Oxford: Oxford University Press.

Holonyak, N, Jr., & Bevacqua, S. (1962). Coherent (visible) light emission from ga (As$_{1-x}$P$_x$) junctions. *Applied Physics Letters*, *1*(4), 82.

Hong, A. J., Song, E. B., Yu, H. S., Allen, M. J., Kim, J., Fowler, J. D., et al. (2011). Graphene flash memory. *ACS Nano*, *5*(10), 7812.

Hong, A. J., Song, E. B., Yu, H. S., Allen, M. J., Kim, J., Fowler, J. D., et al. (2011). Graphene flash memory. *ACS Nano*, *5*(10), 7812.

Hung, S., & Chu, Y. (2006). Stimulating new industries from emerging technologies: Challenges for the public sector. *Technovation*, *26*, 104–110.

Joung, H., An, Y., & Park, Y. (2015). A structured approach to explore knowledge flows through technology-based business methods by integrating patent citation analysis and text mining. *Technological Forecasting & Social Change*, *97*, 181.

Kim, H. S., Kim, C. K., & Jang, H. W. (2013). Fabrication of a microball lens array for OLEDs fabricated using a monolayer microsphere template. *Electronic Materials Letters*, *9*(1), 39.

Kim, J., & Lee, C. (2017). Novelty-focused weak signal detection in futuristic data: Assessing the rarity and paradigm unrelatedness of signals. *Technological Forecasting and Social Change*, *120*, 59.

Ko, Y. G., Hahm, S. G., Murata, K., Kim, Y. Y., Ree, B. J., Song, S., et al. (2014). New fullerene-based polymers and their electrical memory characteristics. *Macromolecules*, *47*(23), 8154.

Kuhn, T. S. (1970). *The structure of scientific revolutions* (Vol. II).

Lecz, R. C., & Lanford, H. (1973). Trend extrapolation: Workhorse of technological forecasting. *Industrial Marketing Management*, *3*(1), 57.

Lee, C., Kang, B., & Shin, J. (2015). Novelty-focused patent mapping for technology opportunity analysis. *Technological Forecasting and Social Change*, *90*, 355.

Lee, C., Kwon, O., Kim, M., & Kwon, D. (2018). Early identification of emerging technologies: A machine learning approach using multiple patent indicators. *Technological Forecasting and Social Change*, *127*, 291.

Lineback, J. (1988). High-density flash EEPROMS are about to burst on the memory market. *Electronics*, *61*(5), 47.

Li, M., Porter, A. L., & Suominen, A. (2017). Insights into relationships between disruptive technology/innovation and emerging technology: A bibliometric perspective. *Technological Forecasting and Social Change*, *129*, 285–296.

Liu, J., Gao, B., Cheng, Y., Xie, Z., Geng, Y., Wang, L., et al. (2008). Novel white electroluminescent single polymer derived from fluorene and quinacridone. *Macromolecules*, *41*(4), 1162.

Loke, D., Lee, T., Wang, W., Shi, L., Zhao, R., Yeo, Y., et al. (2012). Breaking the speed limits of phase-change memory. *Science*, *336*(6088), 1566.

Lu, C. Y., Hsieh, K. Y., & Liu, R. (2009). Future challenges of flash memory technologies. *Microelectronic Engineering*, *86*(3), 283.

Lu, K., & Wolfram, D. (2012). 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*, *63*, 1973–1986.

Magerman, T., Van Looy, B., & Song, X. (2010). Exploring the feasibility and accuracy of Latent Semantic Analysis based text mining techniques to detect similarity between patent documents and scientific publications. *Scientometrics*, *82*(2), 289.

Martin, B. (1995). Foresight in science and technology. *Technology Analysis & Strategic Management*, *7*(2), 139.

Masuoka, F., & Iizuka H. (1985). Semiconductor memory device and method for manufacturing the same. US Patent 4,531,203

Meyaard, D. S., Lin, G. B., Cho, J., Schubert, E. Fred, Shim, H., Han, S. H., et al. (2013). Identifying the cause of the efficiency droop in GaInN light-emitting diodes by correlating the onset of high injection with the onset of the efficiency droop. *Applied Physics Letters*, *102*(25), 251114.

Mittal, S., Vetter, J. S., & Li, D. (2015). A survey of architectural approaches for managing embedded dram and non-volatile on-chip caches. *IEEE Transactions on Parallel and Distributed Systems*, *26*(6), 1524.

Nagai, A., & Chujo, Y. (2010). Luminescent organoboron conjugated polymers. *Chemistry Letters*, *39*(5), 430.

Nakamura, S., Mukai, T., & Senoh, M. (1991). High-power GaN P-N junction blue-light-emitting diodes. *Japanese Journal of Applied Physics*, *30*(12A), L1998.

Nakamura, S., Senoh, M., & Mukai, T. (1993). High-power InGaN/GaN double-heterostructure violet light emitting diodes. *Applied Physics Letters*, *62*(19), 2390.

Nakamura, A., Yanagita, N., Murata, T., Hoshino, K., & Tadatomo, K. (2008). Effects of sapphire substrate misorientation on the GaN-based light emitting diode grown by metalorganic vapour phase epitaxy. *Physica Status Solidi (c)*, *5*(6), 2007.

Nanotechnology, N. (2015). Memory with a spin. *Nature Nanotechnology*, *10*, 185.

Pavan, P., Bez, R., Olivo, P., & Zanoni, E. (1997). Flash memory cells: An overview. *Proceedings of the IEEE*, *85*(8), 1248.

Porter, A., & Cunningham, S. (2005). *Tech mining: Exploiting new technologies for competitive advantage*. Hoboken: Wiley.

Porter, A. L., Cunningham, S. W., Banks, J., Roper, A. T., Mason, T. W., & Rossini, F. A. (2011). *Forecasting and management of technology*. Hoboken: Wiley.

Porter, A. L., Roessner, J. D., Jin, X. Y., & Newman, N. C. (2002). Measuring national 'emerging technology' capabilities. *Science and Public Policy*, *29*(3), 189.

Ranaei, S., Karvonen, M., Suominen, A., & Kassi, T. (2014). In *Portland international conference on management of engineering and technology* (pp. 2924–2937)

Rehurek, R., & Sojka, P. (2010). In *Proceedings of the LREC 2010 workshop on new challenges for NLP frameworks.*

Rizvi, S. S., & Chung, T. S. (2010). A survey of storage management in flash based data centric sensor devices in wireless sensor networks. In *Second international conference on communication systems, networks and applications* (Vol. 1). https://doi.org/10.1109/ICCSNA.2010.5743084.

Roche, I., Besagni, D., François, C., Hörlesberger, M., & Schiebel, E. (2010). Identification and charactерisation of technological topics in the field of molecular biology. *Scientometrics*, *82*(3), 663.

Rosen-Zvi, M., Griffiths, T., Steyvers, M., & Smyth, P. (2004). The author-topic model for authors and documents. In *Proceedings of the 20th conference on uncertainty in artificial intelligence* (pp. 487–494).

Rotolo, D., Hicks, D., & Martin, B. R. (2015). What is an emerging technology? *Research Policy*, *44*(10), 1827.

Schwartz, H., Nichols, D., & Johnston, A. (1997). Single-event upset in flash memories. *IEEE Transactions on Nuclear Science*, *44*(6), 2315.

Simanjuntak, F. M., Panda, D., Wei, K. H., & Tseng, T. Y. (2016). Status and prospects of ZnO-based resistive switching memory devices. *Nanoscale Research Letters*, *11*(1), 368.

Simpson, R., Fons, P., Kolobov, A., Fukaya, T., Krbal, M., Yagi, T., et al. (2011). Interfacial phase-change memory. *Nature Nanotechnology*, *6*(8), 501.

Small, H., Boyack, K., & Klavans, R. (2014). Identifying emerging topics in science and technology. *Research Policy*, *43*, 1450–1467.

Suominen, A., & Newman, N. (2017). A critical evaluation of the technological emergence concept. In *Proceedings of PICMET'17: Technology management for interconnected world.*

Suominen, A., Rilla, N., & Oksanen, J. (2016). Insights from social network analysis-case board interlocks in finnish game industry. In *49th Hawaii international conference on system sciences (HICSS).*

Suominen, A. (2013). Analysis of technological progression by quantitative measures: A comparison of two technologies. *Technology Analysis & Stratgic Management*, *25*(6, SI), 687.

Suominen, A., & Seppänen, M. (2014). Bibliometric data and actual development in technology life cycles: Flaws in assumptions. *Foresight*, *16*(1), 37.

Suominen, A., & Toivanen, H. (2015). Map of science with topic modeling: Comparison of unsupervised learning and human-assigned subject classification. *Journal of the Association for Information Science and Technology*, *67*, 2464–2476.

Tan, A. (1999). Text mining: The state of the art and the challenges. In *Proceedings of the PAKDD 1999 workshop.*

Tanaka, T., Saito, K., Nishio, M., Guo, Q., & Ogawa, H. (2009). Enhanced light output from ZnTe light emitting diodes by utilizing thin film structure. *Applied Physics Express*, *2*(12), 122101.

Tan, Z. K., Moghaddam, R. S., Lai, M. L., Docampo, P., Higler, R., Deschler, F., et al. (2014). Bright light-emitting diodes based on organometal halide perovskite. *Nature Nanotechnology*, *9*(9), 687.

Templeton, K., & Fleischmann, T.C. (2013). In *iConference 2013 proceedings.*

Tsao, J. Y., Han, J., Haitz, R. H., & Pattison, P. M. (2015). The blue led nobel prize: Historical context, current scientific understanding, human benefit. *Annalen der Physik*, *527*(5–6), A53–A61.

Tseng, Y. H., Lin, C. J., & Lin, Y. I. (2007). Text mining techniques for patent analysis. *Information Processing & Management*, *43*(5), 1216.

Velden, T., Boyack, K. W., Gläser, J., Koopman, R., Scharnhorst, A., & Wang, S. (2017). Comparison of topic extraction approaches and their results. *Scientometrics*, *111*(2), 1169.

Venugopalan, S., & Rai, V. (2015). Topic based classification and pattern identification in patents. *Technological Forecasting and Social Change*, *94*, 236.

Waltman, L., Boyack, K. W., Colavizza, G., & van Eck, N. J. (2019). A principled methodology for comparing relatedness measures for clustering publications. arXiv preprint arXiv:1901.06815.

Wang, X., Tian, H., Mohammad, M. A., Li, C., Wu, C., Yang, Y., et al. (2015). A spectrally tunable all-graphene-based flexible field-effect light-emitting device. *Nature Communications*, *6*, 7767.

Wells, H. (1999). *Anticipations of the reaction of mechanical and scientific progress upon human life and thought.* Courier Corporation.

Wold, S., Esbensen, K., & Geladi, P. (1987). Principal component analysis. *Chemometrics and Intelligent Laboratory*, *2*, 37–52.

Yam, F., & Hassan, Z. (2005). Innovative advances in led technology. *Microelectronics Journal*, *36*(2), 129.

Yau, C. K. C., Porter, A., Newman, N., & Suominen, A. (2014). Clustering scientific documents with topic modeling. *Scientometrics*, *100*(3), 767.

Zhang, Y., Duan, Z., Li, R., Ku, C. J., Reyes, P. I., Ashrafi, A., et al. (2013). Vertically integrated ZnO-based 1D1R structure for resistive switching. *Journal of Physics D: Applied Physics*, *46*(14), 145101.

Zhang, Y., Zhang, G., Chen, H., Porter, A. L., Zhu, D., & Lu, J. (2016). Topic analysis and forecasting for science, technology and innovation: Methodology with a case study focusing on big data research. *Technological Forecasting and Social Change*, *105*, 179.

Zhang, Z., Zhao, H., Matsushita, M. M., Awaga, K., & Dunbar, K. R. (2014). A new metal-organic hybrid material with intrinsic resistance-based bistability: Monitoring in situ room temperature switching behavior. *Journal of Materials Chemistry C*, *2*(2), 399.

Zheludev, N. (2007). The life and times of the led-a 100-year history. *Nature Photonics*, *1*(4), 189.

Zheng, W., Kankaanranta, J., & Suominen, A. (2012). Morphological analysis of technologies using multidimensional scaling. *Journal of Business Chemistry*, *9*, 147–160.

Zidan, M. A., Fahmy, H. A. H., Hussain, M. M., & Salama, K. N. (2013). Memristor-based memory: The sneak paths problem and solutions. *Microelectronics Journal*, *44*(2), 176.