Scientometrics
https://doi.org/10.1007/s11192-021-04190-9

ITGInsight–discovering and visualizing research fronts in the scientific literature

Xuefeng Wang[1] · **Shuo Zhang**[1] · **Yuqin liu**[2]

Received: 3 April 2021 / Accepted: 15 October 2021

Abstract
Nowadays, most organizations face the challenge of having to track the latest technological developments so as to discover new technology opportunities and to identify threats in their competitive environment. The capacity to do this relies heavily on the ability to recognize scientific innovation. Hence, monitoring emerging research directions in the scientific literature has become an important task for both researchers and policy makers. Yet the best method of doing so is still a topic of controversy. Our goal is to develop a generic computational framework that can describe a research domain in terms of its research fronts and further track the evolution trends of the knowledge structures behind each research front for the purposes of identifying knowledge innovation. The results show the evolution trends of knowledge structures could lead up to pioneering research. Implemented in ITGInsight, a C# application, the modelling and visualization process incorporates a topic clustering model and a topic evolution model to reveal knowledge structures and their evolution trends. Using the framework in a case study on synthetic biology, we verified the results it produced by consulting the literature and a panel of domain experts. The tool proves to be powerful font of insightful information that would be difficult and time-consuming for researchers and policy makers to gather on their own. Anyone involved in R&D planning, research funds allocation, and technology opportunity analysis will find the framework useful.

Keywords Research front · Research trend · Knowledge structure · Topic clustering model · Topic evolution model · Information visualization

Introduction

Nowadays, most organizations face the challenge of having to track the latest technological developments so as to discover new technology opportunities and to identify threats in their competitive environment, especially given the exponential growth of accessible

✉ Yuqin liu
liuyuqin2004@126.com

[1] School of Management and Economics, Beijing Institute of Technology, Beijing 100081, China

[2] School of Journalism and Publication, Beijing Institute of Graphic Communication, Beijing 102600, China

Published online: 01 December 2021

information (Lee et al., 2014, 2017; Porter & Cunningham, 2005). Further, facing such technique revolution wave tide of the fast fierce development, timely and relevant information on new technology and its industrialization in both internal and external environments will not only has a direct impact on the competitiveness and development of organizations (Liao et al., 2003; Yoon et al., 2015), but also become one of the important forces to boost the economy growth, as well as the main index to decide the integrative competition of a country or territory (Lee et al., 2017; Yoon et al., 2014; Zhu & Porter, 2002). In recent years, numerous studies have indicated that scientific activity is the seed of technological innovation. And that innovation occurs by combining independent pieces of scientific output, especially from the hard sciences like chemistry, biomedicine, and electronics (Shibata et al., 2008; Tijssen, 2002; Wang et al., 2021b). As such, the capacity to capture and assimilate new knowledge related to technology development relies heavily on one's ability to recognize scientific innovation. To this end, monitoring emerging research directions in the literature has become an important task for both researchers and policy makers (Wang & Chen, 2019). Yet, among the strengths and weaknesses of the many technology opportunity analysis frameworks there are out there, there is no one definitive crystal ball.

Currently, the main output of science activities lies in a number of scientific publications or news reports (Bowen & Casadevall, 2015; Casadevall & Fang, 2014; Shibata et al., 2008), and much scholarly effort has been expended in developing approaches to demarcate the frontiers of research in a stream (Wang et al., 2021b). Generally, there are two approaches for discovering research fronts (Huang & Chang, 2014; Shibata et al., 2008). One straightforward manner is the qualitative method, e.g., Delphi, expert interview, scenario planning, which relies on the knowledge of domain experts. This method is time-consuming and subjective and becoming more so on both counts in the current information-flooded era. Another is the quantitative method, e.g., bibliometric analysis, patent citation analysis or their combination, which is more time- and cost- efficient and often used to complement the expert-based approaches. Strotmann and Zhao (2014) combined author co-citation and bibliographic coupling analysis to unveil the knowledge base and research fronts of information science (IS). Their experimental results show that the main research fronts in IS are knowledge domain analysis via quantitative studies of science and technology, and information retrieval and representation. Using the Web of Science and the Airiti Library, Fang and Lee (2021) explored sub-fields, characteristics, research fronts and trends in research on technology education through bibliometric and co-citation analysis. Their insights into the evolution of research topics and research fronts provided valuable information for technology educators and policymakers.

Timeliness is key to these insights. To make the most of this knowledge, researchers and policymakers must learn of research fronts and key technology developments in advance (Morris et al., 2014). Thus, exploring the development trends of research fronts needs to be paid sufficient attention, and particular focus needs to be paid to revealing the evolutionary trends of technological advancements at a micro level. Few studies deal with this aspect of the analysis. In academia, researchers often use citation analysis, dividing the sample into citation windows of different time frames to track the development of research fronts in a timely and accurate fashion. Shibata et al. (2008) detected emerging research fronts in a huge number of academic papers related to regenerative medicine–a field of radically innovative research. Analyzing the clustering results according to the average published year and parent–child relationship of each cluster, they derived many future development trends in regenerative medicine. Small (2006) and Upham and Small (2010) collected data from three overlapping 6-year periods and studied twenty-two broad disciplines. They tracked the emergence and growth of research fronts in these disciplines from co-citation

cluster strings. Upham and Small (2010) went so far as to classify the research fronts into five categories according to the maintenance time of each. The five categories are: emerging fronts, growing fronts, stable fronts, shrinking fronts, and exiting fronts. Huang and Chang (2014) applied bibliographic coupling and a sliding window to explore the research fronts of organic light-emitting diodes (OLED) from 2000 to 2009. They identified eighteen research fronts that match those predicted by subject experts related to OLED materials. Closer observation of the evolution showed that among the eighteen fronts, four were emerging, two were growing, eleven were stable, and one was shrinking. Nevertheless, it is crucial to acknowledge that the above methods cannot describe the evolution trends of research fronts at the micro level and they rely heavily on expert opinions.

Given these shortcomings, we apply research fronts as a way to represent a research domain's state-of-the-art thinking (Chen, 2006; Price, 1965). Our goal is to develop a generic computational framework to describe a research domain in terms of its research fronts and further track the evolution trends of the knowledge structures behind each research front for the purposes of identifying knowledge innovation. Thus, this paper focuses on answering the following research questions:

(i) What are the research fronts of a research domain? What are the knowledge structures of each research front?
(ii) What are the trends and directions of each knowledge structure? How do they support the development of a research front?

The remainder of this paper consists of four sections. We first contain a brief literature review to introduce some related works. Next, we describe the generic computational framework, which uses a topic clustering model and a topic evolution model to describe the research domain. The goal is to enable researchers or policy makers to identify and understand the research domain more clearly. Then, we illustrate the new computational framework using a case study on synthetic biology. Finally, we present remarks and directions for further study.

Related work

Research fronts discovery

The concept of a research front was originally introduced by Price (1965) to characterize the transient nature of a research domain. They can be generally divided into five categories: emerging fronts, growing fronts, stable fronts, shrinking fronts, and exiting fronts (Upham & Small, 2010). Pinpointing the research fronts of a specific field can not only provide insights into current focuses, but also serve as an important indicator of government blueprints for technological development and policy decision making. Studies on research fronts are generally conducted using either qualitative or quantitative methods. Although not mutually exclusive, quantitative methods, with their combination of citation, bibliographic coupling, co-citation, co-word, etc. are more appropriate for studying research fronts (Huang & Chang, 2014). This is because quantitative methods not only provide realistic descriptions of current research development but are also more time- and cost-efficient. Undoubtedly, co-citation analysis and bibliographic coupling analysis are the two dominant methods for discovering research fronts of research fields and have been

frequently studied, used, and improved upon since their introduction (Strotmann & Zhao, 2014). Shibata et al. (2009) conducted a comparative study to investigate different methods of detecting emerging research fronts, including co-citation, bibliographic coupling, and direct citations. Their experiments show that the direct citation method detected larger and younger emerging clusters earlier, achieving the overall best performance at identifying research fronts. Further, bibliographic coupling performed better than co-citation analysis. In a case study on biomedicine, Boyack and Klavans (2010) compared the accuracy of co-citation analysis, bibliographic coupling, direct citation, and a hybrid approach that combined bibliographic coupling with citation analysis. They find that bibliographic coupling slightly outperformed co-citation analysis using both accuracy measures, and direct citation was the least accurate mapping approach by far. However, the hybrid approach improved upon bibliographic coupling in all respects. Ever since, bibliographic coupling has attracted extensive attention. For example, Huang and Chang (2014) used bibliographic coupling and a sliding window to explore research fronts in organic light-emitting diodes (OLED) from 2000 to 2009. Their experiments show that bibliographic coupling with a sliding window is an effective tool for discovering the evolution of research fronts. Pieiro-Chousa et al. (2019) aimed to provide researchers and policy makers with a valuable source of knowledge that could allow them to define research lines, business strategies and policy measures. Therefore, they analyzed the current research fronts at the intersection of the three concepts in business research, through performance metrics, bibliographic coupling, and word co-occurrence analysis.

To take a closer look at the research fronts, the words obtained from articles are used to label research fronts. For example, in the research of Small and Griffith (1974), the research fronts were labeled by word profiles derived from citing articles. Persson and Olle (1994) and Morris et al. (2003) both manually examined title words to label research fronts. Liu et al., (2015a, 2015b) and Ma and Liu (2016) believe the keywords of the articles could help pinpoint its research focus, while Chen (2006) contends that researchers are interested in not only the most commonly used terms but also the terms that can lead to profound changes. Therefore, an ideal labeling method should distinguish emerging trends and rapid changes in the foreground from more persistent themes in the background. In his research, a current research front is identified based on such burst terms extracted from titles, abstracts, and descriptors (Chen, 2006). Since then, this method has been widely used in academia (Chen & Yang, 2021; Ping, 2015; Yi & Di, 2016). However, according to the concept of research front, it is more appropriate to use terms that can present the essence or knowledge of a research front, instead of merely focusing on the burst of single words.

Emerging trends detection

Emerging research trends are those topics that have recently attracted people's interest and are being discussed by more and more (Galitsky et al., 2004). Thus, detecting emerging research trends is a process of discovering what information is currently 'hot' in a specific field and actively signaling the fact a new development is detected (Cobo et al., 2012; Wang et al., 2021b; Zhang et al., 2016). Prior studies have proposed two ways to analyze the emerging trends. One is to collect and read piles of literature, review the articles, and summarize the trends and directions for further research. Another is bibliometric methods, which involves conducting statistical analyses of the research outputs of countries, research institutes, journals, research fields, and so on

(Keiser & Utzinger, 2005; Xie et al., 2008). Common bibliometric techniques include word frequency analysis, citation analysis, and co-word analysis. However, citation analysis has its drawbacks. First, significant publication delay causes citation delay, which can seriously skew the results of current research trends analysis. Second, as is well known, impact does not necessarily show up as citations. As an example, Mendelian genetics is a widely accepted phenomenon but seldom cited (Wang et al., 2013). In addition, word-based methods of analysis are sensitive to the choice of words. Determining a word frequency threshold is a subjective matter and one that requires manual intervention in the process of detecting emerging trends. Thus, researchers pay more attention to the combined application of these methods.

Taking number of downloads, keywords, and articles into consideration, Wang et al. (2013) designed a method of detecting emerging research trends. They find that, in scientometrics, new indices are being developed to quantify scientific productivity, such as the g-index, and that researchers are focusing on new and emerging fields like, webometrics, semantics, text mining, and open access. Zhang et al. (2013) constructed a hybrid model for composing technology roadmaps by integrating bibliometrics with qualitative methodologies and visualization techniques. The mapping could array details on the technology evolution process and macro-technology development status. Zhang et al. (2016) developed a series of technology roadmapping models that could balance qualitative and quantitative methods to get from the historical data-based technology roadmapping to forecast future development trajectories. Huang et al. (2016) combined topical analysis, patent citation analysis, and term clumping analysis to capture technology evolution pathways in detail, which achieved a balance between data-driven and expert-influenced conclusions. To some extent, the hybrid methods overcome some of the shortcomings of the above methods. However, it is crucial to acknowledge that the above methods cannot describe the evolution trends at the micro level and they all rely on expert opinions to some extent.

Our proposal is for a generic computational framework to describe a research domain based on its research fronts and further track the evolution trends of the knowledge structures behind each research front for the purposes of identifying knowledge innovation. In summary, the main contributions of this works are as follows:

(1) A novel approach for labeling research fronts in a research domain;
(2) A topic clustering model for discovering the knowledge structures of a research front;
(3) A topic evolution model for tracking the evolution trends of the knowledge structures of a research front.

ITGInsight

We next discuss the domain term extraction method, the topic clustering model, and the topic evolution model and demonstrate how they are used to profile a research domain within ITGInsight. The conceptual outline of the framework is given in Fig. 1. As illustrated, the research fronts are labeled via domain term extraction. The topic clustering model identifies the knowledge structures of each research front, and the topic evolution model traces changes in those knowledge structures, providing the trend analysis component of the results.

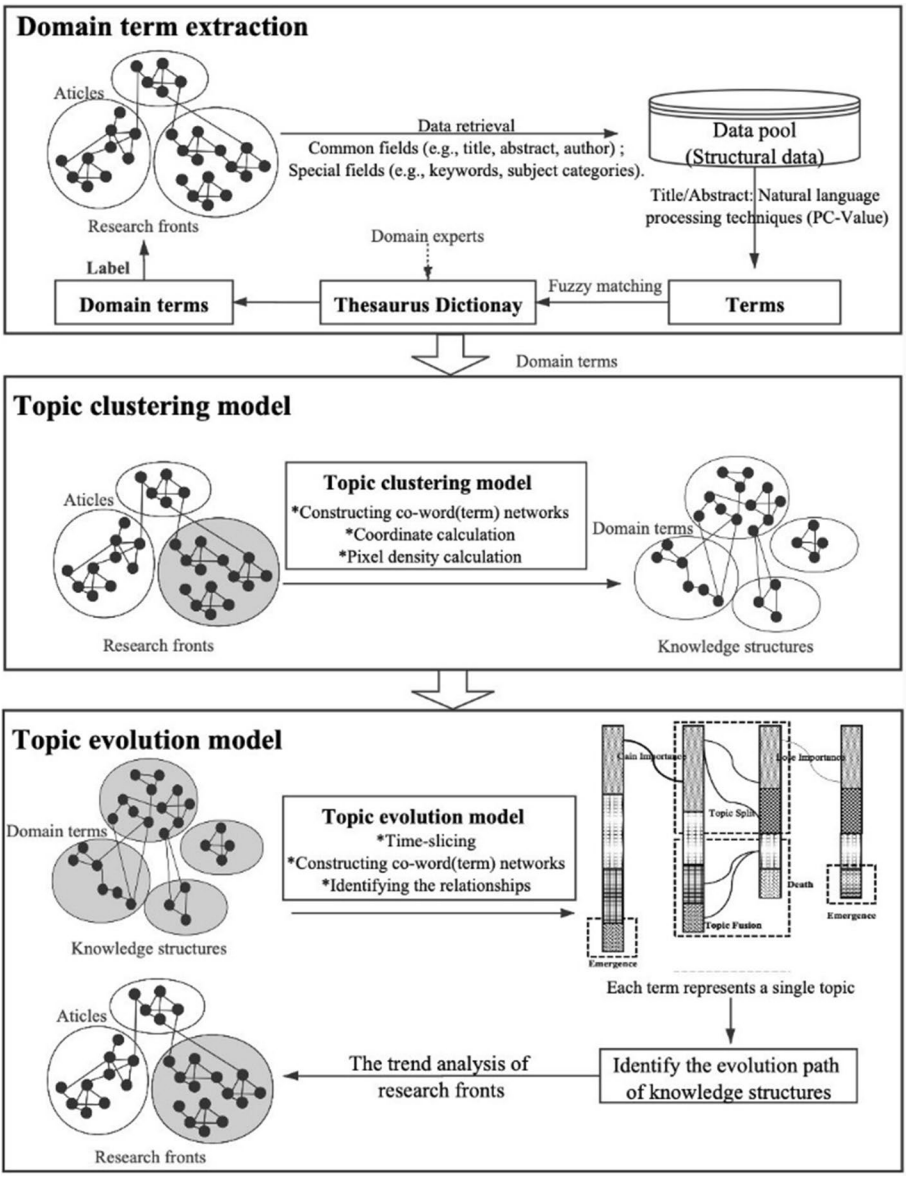


Fig. 1 The conceptual model of ITGInsight *Terms embody concepts, and concepts are knowledge units (Tian et al., 2012; Ye et al., 2012). Thus, knowledge structures can be represented by their core terms and their co-occurrence in a network.*

Domain term extraction

The purpose of this step is to retrieve relevant information with which to label the research fronts. The process for doing so is based on the assumption that, if a word in a paper appears many times, it is more likely to be a domain term, especially if it is a

long word. Therefore, our chosen method is PC-value (Han et al., 2011), a domain term extraction method originally based on C-value method (Frantzi et al., 2000). PC-value considers the frequency statistics of words in documents, as Wang et al., (2014) state, terms extracted by PC-value method represent the essence or knowledge of a research domain. Further, experiments show that this method has higher accuracy than C-value (Han et al., 2011). The formula for calculating PC-value is as follows:

$$PC-value(a)=\begin{cases}\log_2^{|a|}\cdot f(a)+2^{|a|-2}\cdot g(a) & a\ is\ not\ nested\\ \log_2^{|a|}\cdot (f(a)-\frac{1}{|T_a|}\sum_{b\in T_a} f(b))+2^{|a|-2}\cdot g(a), & otherwise\end{cases} \quad (1)$$

where a is the candidate term, $|a|$ is the length of a, $f(a)$ is its frequency of occurrence in the corpus, and $g(a)$ is the document frequency of a. b represents an extracted candidate term that contains a, and $f(b)$ is the total frequency of b in the corpus. T_a is the set of extracted candidate terms that contain a, and $|T_a|$ is their number.

The PC-Value procedure works as follows: Words are segmented and parts of speech are tagged. Meaningless and extreme words are then deleted, and words of more than 95% similarity are combined. The remaining words are then sorted by their PC-value to form the preliminary results. The most important aspect of this step is to establish a thesaurus based on the preliminary results. (It is worth noting that users can often improve on the thesaurus with the help of domain experts). The thesaurus is used to reduce noise, consolidate related terms, and provide more refined terms. Table 1 summarizes the procedure.

Table 1 The term extraction procedure

Step	Description
Word segmentation process	
1	*Raw dataset*- apply a hidden Markov model for word segmentation and compute their PC-Values
2	*Data cleaning*-remove common/meaningless words (e.g., a/an, the, what, detailed description, some time, method) and extreme words (e.g., occurrence in only one record, word length less than 2)
3	*Word segmentation outputs*-sort the words according to their PC-value to form the preliminary results
Term extraction	
4	*Fuzzy matching*-combine words with similar structures based on pattern commonality, such as stemming and text similarity to form a thesaurus
5	*Raw dataset*-combine the hidden Markov model with the thesaurus for term extraction
6	*Term extraction*-Obtain the domain terms according to the PC-value

(*Source*: Adapted and modified from Wang et al., 2021a)

ITGInsight supports various user-defined dictionaries, and they can be used to intervene in natural language processing without directly modifying the original data.

Topic clustering model

The analysis of the knowledge structures of research fronts is important for knowledge diffusion of the emerging field, promoting the interdisciplinary and more researchers to participate in the discussion and research in related fields. Traditionally in bibliometrics, co-citation analysis, co-word analysis and bibliographic coupling are used to analyze the knowledge structures of a research front (Liu et al., 2015a, 2015b; Yan et al., 2015). However, although the concept of knowledge structure is widely used, there is no consensus over its meaning or how they should be applied. Some researchers believe that the knowledge structures in a research front have multiple dimensions, and they can be partially revealed through different analysis perspectives (Li & Chu, 2016). Tian et al. (2012) and others argue that terms are the embodiment of the concept, and that concepts are the units of knowledge (see also Ye et al., 2012). Therefore, the knowledge structures of a specific field can be represented by their core terms and their co-occurrences in a network map of the domain. Following this philosophy, ITGInsight applies a clustering algorithm to co-word (term) network, and the resulting clusters are regarded as the knowledge structures.

We chose a version of LinLog (Noack, 2004), modified with a pixel density function (Liu et al., 2017) as our topic clustering model. The LinLog algorithm constructs a network graph of uniform density, showing separated clusters with interpretable distances between the clusters. The pixel density function then is used to color the nodes and the final results could be presented as a topographic map, where the contour lines clearly reveal the relationship between different domain terms (Wang et al., 2021a). Details of the specific steps are outlined below.

Step1: *Construct a co-term matrix.*

A co-occurrence matrix is constructed according to the strength of the relationships between domain terms as follows:

$$Corr_{n*n} = \begin{bmatrix} Doma_1 & Domainterm_1 & Domainterm_2 & \cdots & Domainterm_1 & \cdots & Domainterm_1 \\ Domainterm_1 & r_{11} & r_{12} & \cdots & r_{1i} & \cdots & r_{in} \\ Domainterm_1 & r_{21} & r_{22} & \cdots & r_{2i} & \cdots & r_{2n} \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ Domainterm_1 & r_{i1} & r_{i2} & \cdots & r_{ii} & \cdots & r_{in} \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ Domainterm_n & r_{n1} & r_{n2} & \cdots & r_{ni} & \cdots & r_{nn} \end{bmatrix} \tag{1}$$

Relationship strength r_{ij} depends on the frequency of co-occurrence. At the end of this step, we have a set of nodes V (domain terms) and a set of edges E (the relationships between domain terms).

Step2: *Coordinate calculation.*

The domain terms are clustered with the LinLog algorithm (Noack, 2004), which determines the position coordinates of domain terms in a plane. The two main components of LinLog are an energy model and an energy minimization algorithm. In the energy model, repulsion is based on edges repelling each other rather than node repulsion. Formally, this is defined as

$$U_{LinLog}(p) = \sum_{(u,v)\in E} P_u - P_v - \sum_{(u,v)\in V^{(2)}} \deg(u)\deg(v) P_u - P_v \tag{2}$$

where $P_u - P_v$ represents the distance between node u and node v,$P_u\left(P_v\right)$ is a vector of the node positions, $V^{(2)}$ is the set of all subsets of V which have exactly two elements, and *deg*

(u)(*deg(v)*) is the node degree. The first part of the subtraction in the above equation is a calculation of the attraction between adjacent nodes. The second part is the calculation for the repulsion between the edges. As such, each node ends up having an influence on the drawing, which is proportional to the number of edges connected to them (the degree). This property of the node can thus be reflected on the graph being drawn by making the size of the node proportional to its degree.

The energy minimization algorithm used in LinLog was proposed by (Barnes & Hut,1986). It is based on the idea that the combined effect of a group of nodes can be represented by the effect of the center of mass of that group. The algorithm consists of three components: building an octree for a 3-D space or a quadtree for a 2-D space; computing the center of mass of all the cells; calculating the forces on each point. At the end of this step, we have the position coordinates of each node.

Step3: *Plane pixel density function.*

After obtaining the position coordinates of each node, a pixel density function (Liu et al., 2017) determines the color of the nodes. The density function formula is

$$Density(x,y) = \sum_{i=1}^{n} f(Number_i)_e - \alpha\left(\frac{\sqrt{(x-x_i)^2+(y-y_i)^2}}{Distance}\right)^{\beta}, \alpha > 0, \beta > 0 \quad (3)$$

where (x_i, y_i), $i = 1 \ldots N$ is the position coordinates of each node, noting that the distance between two different nodes depends on the average two-dimensional Euclidean distance, and $f(Number_i)$ is the standardized value for the number of nodes. (x, y) represents the pixel coordinates on the computer screen. After standardizing, these are mapped to an RGB color code. By default, ITGInsight uses a blue-yellow–red color scheme, where blue corresponds to the highest item density and red corresponds to the lowest item density. Changes in node color form different contour lines, and the final results are presented as a topographic map.

Topic evolution model

The topic evolution model traces the evolutionary changes in the research front's knowledge structures. Here, each term represents a single topic. By tracking the evolutionary trends of different terms in each knowledge structure, we can infer trends in R&D and further unveil knowledge innovation. The specific methods are outlined below.

With the thesaurus generated in Sect. 3.1 in hand, the bibliographic records are divided into time periods $D_t = \{d_{t1}, d_{t2}, d_{t3}, \ldots, d_{tm}\}$(*d*: document; *m*: document number; *t:* time). Each document corresponds to many domain terms, $d = \{w_1, w_2, w_3, \ldots, w_z\}$(*w*: term), and each term represents a single topic. The next step is to construct a co-term network for each time period based on term (topic) co-occurrence. More specifically, if Topic *a* and Topic *b* appear in the same document, they are deemed to have a co-occurrence relationship. Further, the number of times *a* and *b* co-occur in any document is considered to be a reflection of the strength of the connection and is weighted according to the number of co-occurrences. The differences between each co-term network for each time period reveals how topics have evolved. (Wang et al., 2021a). For instance, some topics emerge in a time period; some disappear. Some gain or lose importance; others merge or split. Figure 2 illustrates how the following evolutionary relationships are identified with some examples. In addition, in the actual process of ITGInsight, each topic has a different color, and the edges

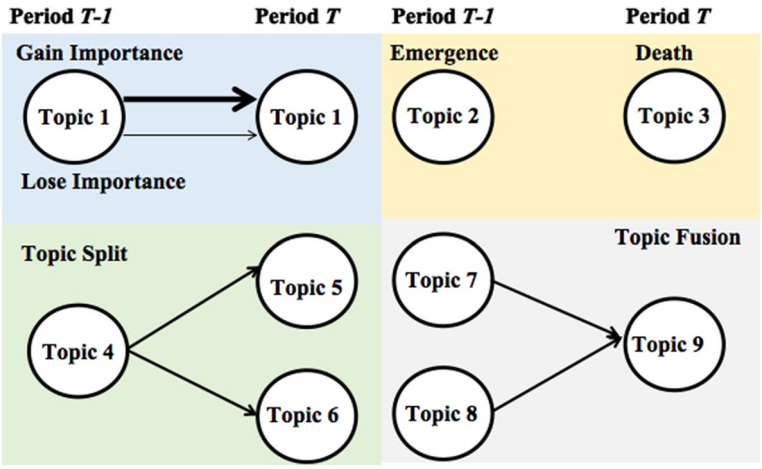


Fig.2 The evolution relationships

are colored according to the node relationships. Therefore, the final results are presented as a topic evolution map.

Type1-Gain/Lose importance: the frequency with which a topic is mentioned has been steadily increasing/decreasing over several consecutive periods.
Type2-Emergence: a new topic is appears suddenly within a period; *Death*: a topic does not appear in this or any subsequent periods.
Type3-Topic split: a new topic is generated from an existing topic (In time period *T-1*, two topics appear in the same document; in time period *T*, they begin to appear in different documents).
Type4-Topic fusion: an existing topic fuses with another existing topic (In time period *T-1*, two topics appear in different documents; in time period *T*, they begin to appear in the same document).

Procedure

The overall procedure for using ITGInsight is described in the following steps.

a. Identify a research domain using the broadest possible words. This is to ensure that the subsequent analysis covers all the major components of the research domain.
b. Data collection: The words chosen in the first step are used to retrieve bibliographic records from a database, including titles, abstracts, authors, institutions, publications, cited references, etc. ITGInsight has many filters that allow users to import data from almost any data source, e.g., Web of Science, PubMed, Derwent Innovation Index. Further, it includes user-defined field extraction, duplicate record detection, and several other useful functions.
c. Threshold selection: Users first specify the range of the study period. They can then set a threshold for any field, such as word occurrence frequency, word length, bibliographic coupling counts, citation counts, etc.

d. Extract domain terms: ITGInsight then extracts relevant keywords from the titles and abstracts of the documents in the dataset using PC-Value method.
e. Calculation: Topic clustering and topic evolution are then calculated and the results are displayed visually. There are functions for zooming, scrolling, searching, and redisplaying the graph as a user-defined type to help users examine large maps.
f. Verify the results: The results can be verified by asking domain experts or examining the relevant literature. (Further, ITGInsight can interpret the analytical results automatically on the basis of a widely used framework.)

For a more extensive discussion of the functionality of ITGInsight, the ITGInsight manual is available at http://cn.itginsight.com/Files/download/itginsight_manual.pdf.

Case study

To illustrate how ITGInsight works in practice, we chose synthetic biology as a domain and used it to answer the following questions: (i) What are the research fronts of this domain? What are the knowledge structures in each research front? (ii) What are the trends and directions of each knowledge structure? How do they promote the development of the identified research fronts?

Data collection

As defined by three scientific committees of the European Commission, synthetic biology is "the application of science, technology and engineering to facilitate and accelerate the design, manufacture and/or modification of genetic materials in living organisms" (Breitling & Takano, 2015; Breitling et al., 2015). The key features of this technology domain–e.g., "the de novo synthesis of genetic material and an engineering-based approach to develop components, organism a product"–have attracted tremendous scientific and industrial attention, and it will further give birth to the next biotechnology revolution.

We applied the following consolidated search strategy proposed by Philip et al. (2017) to publications recorded in Web of Science (WoS) for the period 2000–2020 in Science Citation Index Expanded (SCI-Expanded): "*(((TS=("synthetic biolog*" OR "synthetic dna" OR "synthetic genom*" OR "synthetic *nucleotide" OR "synthetic promoter" OR "synthetic gene* cluster") NOT TS=("photosynthe*")) OR (TS=("synthetic mammalian gene*" AND "mammalian cell") NOT TS="photosynthe*") OR (TS="synthetic gene*" NOT TS=("synthetic gener*" OR "photosynthe*")) OR (TS=("artificial gene* network" OR ("artificial gene* circuit*" AND "biological system")) NOT TS="gener*") OR (TS=("artificial cell") NOT TS=("cell* telephone" OR "cell* phone" OR "cell* culture" OR "logic cell*" or "fuel cell*" or "battery cell*" or "load-cell*" or "geo-synthetic cell*" or "memory cell*" or "cellular network" or "ram cell*" or "rom cell*" or "maximum cell*" OR "electrochemical cell*" OR "solar cell*")) OR (TS=("synthetic cell") NOT TS=("cell* telephone" OR "cell* phone" OR "cell* culture" OR "logic cell*" or "fuel cell*" or "battery cell*" or "load-cell*" or "geo-synthetic cell*" or "memory cell*" or "cellular network" or "ram cell*" or "rom cell*" or "maximum cell*" OR "electrochemical cell*" OR "solar cell*" OR "photosynthe*")) OR (TS=("artificial nucleic acid*" OR "artificial *nucleotide")) OR (TS=("bio brick" or "biobrick" or "bio-brick"))))*". From this search (on 9th May 2020), we retrieved 12,725 records. ITGInsight was then used for record cleaning. After removing

duplicate records, including the papers published with the same title, abstract and authors, our synthetic biology publication dataset comprised 12,525 articles.

Domain term extraction

The PC-Value method discussed in Sect. 3.1 was applied to extract the domain terms. To verify the efficacy of the PC-Value method, we compared it with burst term detection. The burst detection procedure was as follows: After word segmentation and part of speech tagging, meaningless and extreme words were deleted, and words with similar stems were combined. We then selected the top 50 terms according to strength for each method. The two lists are shown in Table 2.

The results combined with the word segmentation outputs were then e-mailed to three PhD candidates in a related field. The goal of the evaluation was to verify whether the PC-Value method is more appropriate to present the essence or knowledge of a research front. Their responses can be summarized into two points:

First, both the two methods are appropriate to complement expert-based approaches for characterizing the current research fronts as they are more time- and cost- efficient. Second, the burst detection method is more suitable for detecting abrupt changes, as this list reflects changes in the researchers' focal points. For example, we can see "machine learning" appears in the results of burst terms detection. Nowadays, new metabolic models have been created in synthetic biology field to expand our understanding of metabolism beyond what can be extracted from experimental data. These models can deal with large-scale data processing in a way that humans cannot. These genome-scale models have been equipped with various optimization algorithms, which endeavor to characterize the complex metabolic networks and regulations, predict intracellular metabolite dynamics. Particularly, machine learning is becoming a powerful emerging tool for deciphering complex biological phenomena, recognizing hidden physiological patterns, and learning the rules that govern how life was created. However, "machine learning" cannot reflect the essential characteristics of a research domain as a domain term. From this standpoint, the results of PC-Value are therefore more appropriate.

Research fronts analysis and knowledge structures discovery

According the previous studies, a subcluster of three or more articles is substantial enough to represent a meaningful outcome (Huang & Chang, 2014). Therefore, subclusters of at least three articles qualify as the basis of a research front. Among these subclusters, ITGInsight identified five research fronts between 2000 and 2020 through bibliographic coupling analysis and the LinLog algorithm (Fig. 3).

The five research fronts span 162 articles. However, from simply reading the titles and abstracts manually, it is difficult for researchers to understand the contents of each research front. This is where the ITGInsight method helps to shine a brighter light on the actual knowledge contained in each research front–the goal being to identify the specific knowledge structures. An example is useful here. Taking research front 1 in the upper left, this front contains 43 core documents and the domain terms: mammalian cell [6], gene circuit [5], synthetic biology method [3], gene expression [3], mathematical model [3], synthetic gene circuit [3], phase separation [3], live cell [3], logic gate [3], human health [2], Crispr system [2], novel function [2], protein engineering [2], immune response [2], quorum sensing [2], metabolic processes [2], molecular computation [2], RNA-only delivery [2],

Table 2 The results of term extraction

No	Domain terms	PC-value	Burst terms	Strength
1	Gene expression	1528.62	Synthetic gene	102.68
2	Artificial gene synthesis	949.31	Synthetic DNA	29.54
3	Synthetic cell	949.00	Synthetic oligonucleotide	27.96
4	Metabolic engineering	929.44	Synthetic promoter	24.52
5	Synthetic promoter	498.00	Codon usage	18.97
6	Cell free system	466.58	Fusion protein	18.40
7	Essential gene	456.00	Synthetic gene network	17.91
8	Gene circuit	439.75	Genetic interactions	17.08
9	Biological system	364.00	Synthetic genetic interactions	15.25
10	Natural product	360.00	Recombinant protein	15.20
11	Mammalian cell	348.00	Gene therapy	14.96
12	Synthetic genetic network	347.07	Base pairs	13.73
13	Synthetic gene circuit	336.61	Systems biology	13.55
14	System biology	301.00	Synthetic genetic array	13.46
15	Recombinant protein	282.33	Transgene expression	13.36
16	Nucleic acid	271.00	Bottom-up synthetic biology	13.36
17	Fusion protein	262.29	Biological network	13.32
18	Transcription factor	235.62	Gene network	13.18
19	C. glutamicum	232.00	Biological system	12.79
20	DNA sequence	220.43	Inclusion bodies	12.56
21	Secondary metabolite production	208.76	Regulatory network	12.32
22	Transgenic plant	192.00	Microbial cell factories	11.52
23	CPG ODN	182.33	Molecular systems biology	11.49
24	System metabolic engineering	166.65	Machine learning	11.40
25	Genetic code	164.83	Plasmid DNA	11.27
26	Transgene expression	164.00	t cells	11.19
27	DNA polymerase	161.57	Gene delivery	11.00
28	Small molecule	157.00	Amino acid residues	10.99
29	Microbial cell factory	153.63	High affinity	10.79
30	Essential protein	152.00	Escherichia coli	10.72

Comparison of C-Value and PC-Value has already been published by one co-author (Han, H. Q., Zhu, D. H., & Wang, X. F. (2011). Technical Term Extraction Method for Patent Document. *Journal of The China Society for Scientific and Technical Information 30* (12), 1280–1285). Therefore, it will not be reiterated here.

probiotic strain [2], etc. (The contents in brackets represent domain term frequency). From this we can infer that Research Front 1 concerns "synthetic mammalian gene circuits".

The literature tells us that the knowledge structures of synthetic biology fall roughly into four categories: parts, devices, systems, and applications (Lucentini, 2006). When we go a step further to produce a topographic map (Fig. 4) of each research front through the topic clustering model, the map reveals that two knowledge structures appear in the same cluster: "Applications 1—Biomedicine (Immune response)" (inflammatory intestinal disease, gut microbiome, therapeutic substance, immune response, human health) and "Devices 1—Synthetic gene circuit" (synthetic gene circuit, novel

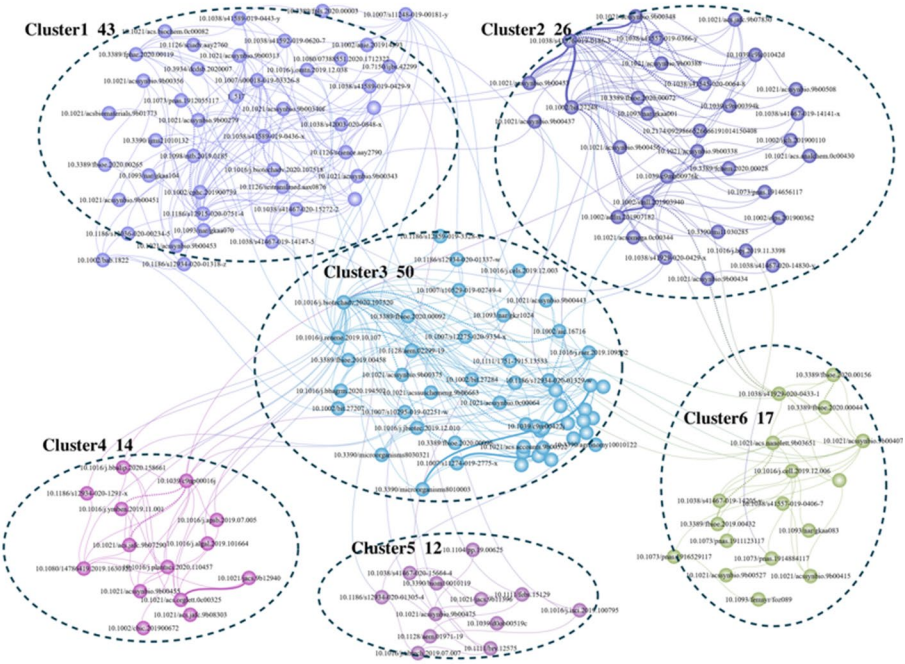


Fig. 3 A 396-node network of bibliographic coupling for the field of synthetic biology (2000–2020) *(Note: The coupling strength threshold = 2)*

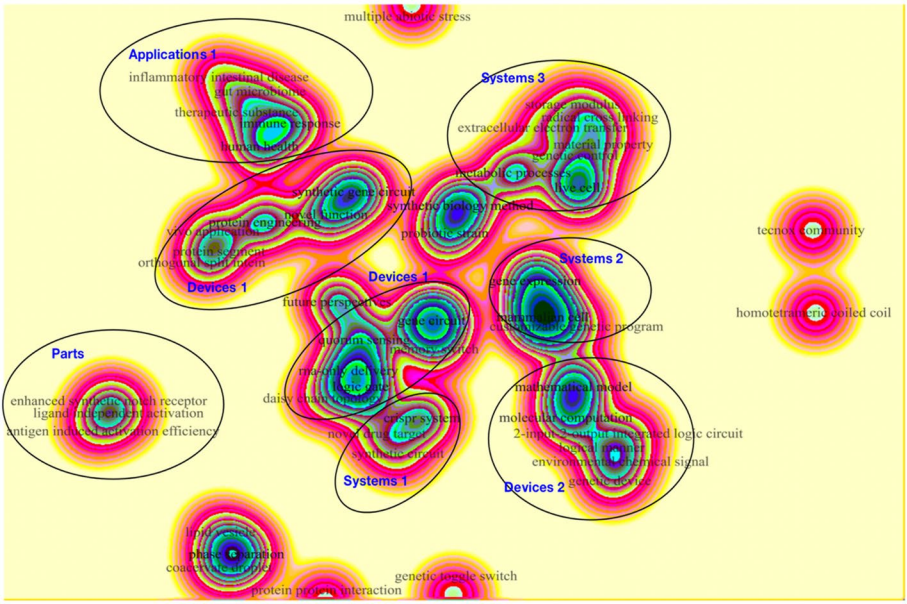


Fig. 4 Topographic map

function, protein engineering, vivo application, protein segment, orthogonal split intein-protein). This shows that a close line exists between these two knowledge structures. In particular, synthetic gene circuit and immune response with highest item density play an important role in this phenomenon. Again consulting the literature, we find that a proof-of-concept immunomodulatory gene circuit platform that enables tumor-specific expression of immunostimulators in the research of Nissim et al. (2017). In their words,

> *The design comprised de novo synthetic cancer-specific promoters and, to enhance specificity, an RNA-based AND gate that generates combinatorial immunomodulatory outputs only when both promoters are mutually active. These outputs included an immunogenic cell-surface protein, a cytokine, a chemokine, and a checkpoint inhibitor antibody. In in vivo efficacy assays, lentiviral circuit delivery mediated significant tumor reduction and prolonged mouse survival.*

The experiment results show that their design could be adapted "to drive additional immunomodulators, sense other cancers, and potentially treat other diseases that require precise immunological programming" (e.g., rheumatoid arthritis, inflammatory intestinal diseases and other autoimmune diseases) (Nissim et al. (2017).

Figure 4 also reveals that "Devices 1—Synthetic gene circuit" (gene circuit, logic gate, quorum sensing, memory switch, RNA-only delivery, etc.) and "Systems 1—CRISPR systems" (CRISPR systems, novel drug target, synthetic circuit) appeared in the same cluster. This shows that a close line exists between these two knowledge structures. In particular, gene circuit, logic gate and CRISPR system, with the highest item density, play an important role in this phenomenon. The literature confirms this case. Li et al. (2020) believe quorum sensing (QS) is a bacterial cell-to-cell communication system. Bacteria sense the density of bacterial population by secreting diffuse small molecular signals, thus causing the coordinated expression of a group of specific genes at the transcriptional level (Garg et al., 2014; Miller et al., 2001). In recent years, genetic circuits containing components of the bacterial QS system have been constructed through synthetic biology to achieve intra-species and inter-species artificial communication, and these genetic circuits based on QS have great application potential in biotechnology and biomedicine (García-Aljaro et al., 2012). By combining the logic switch (logical gate) with the quorum sensing system, researchers have tried to construct a series of new gene lines to transform and realize interspecific communications between bacterial populations (Karafyllidis, 2012). Swofford et al. (2015) cloned the lux quorum-sensing (QS) system and a GFP reporter into non-pathogenic Salmonella. Fluorescence and bacterial density were measured in culture and in a tumor-on-a-chip device to determine the critical density necessary to initiate expression. The experiment results show that QS Salmonella is a promising tool for cancer treatment that will target drugs at tumors, while preventing damage to healthy tissue. We also find that "CRISPR system" has a strong relationship with "novel drug target". In the research of Behan et al. (2019),

> *they performed genome-scale CRISPR–Cas9 screens in 324 humancancer cell lines from 30 cancer types and developed a data-driven framework to prioritizecandidates for cancer therapeutics. They further integrated cell fitness effects withgenomic biomarkers and target tractability for drug development to systematically prioritizenew targets in defined tissues and genotypes. Their analysis provides a resource ofcancer dependencies, generates a framework to prioritize cancer drug targets and suggestsspecific new targets.*

The principles described in this study can inform the initial stages of drug development by contributing to a new, diverse and more effective portfolio of cancer drug targets, thus confirming the relationship between "CRISPR system" and "novel drug target".

"Systems 2—Gene expression" (gene expression, mammalian cell, customizable genetic program) and "Devices 2—Genetic devices (Mathematic model)" (mathematic model, molecular computation, genetic devices, etic) also appear in the same cluster, showing another close line that exists between these two knowledge structures. In particular, gene expression, mammalian cell, customizable genetic program, and mathematic model with highest item density play an important role in this phenomenon. To verify the results, we look up the related literature.

> *Nowadays,* genetically *engineering cells to perform customizable functions is an emerging frontier with numerous technological and translational applications. However, it remains challenging to systematically engineer mammalian cells to execute complex functions. To address this need,* Joseph et al. (2021) *developed a method enabling accurate genetic program design using high-performing genetic parts and predictive computational models. Specifically, they built multifunctional proteins integrating both transcriptional and posttranslational control, validated models for describing these mechanisms, implemented digital and analog processing, and effectively linked genetic circuits with sensors for multi-input evaluations. The functional modularity and compositional versatility of these parts enable one to satisfy a given design objective via multiple synonymous programs. Their approach empowers bioengineers to predictively design mammalian cellular functions that perform as expected even at high levels of biological complexity.*

Additional knowledge structures in Research Front 1 include "System 3 – Metabolic processes" (metabolic processes, live cell, storage modulus, extracellular electron transfer, genetic control, material property, radical cross linking) and "Parts – Synthetic notch receptor". Metabolic processes, live cell and synthetic notch receptor have the highest item densities, and these play an important role in both System 3 and Parts (Roybal et al., 2016; Chen et al., 2018).

Emerging trends detection

Detecting emerging trends in various fields of research is currently an area of great interest in academia and industry. It is also a critical component of resource allocation decisions in research laboratories, government institutions, and corporations. According to Mane and Borner (2004), new areas of science continually evolve, while others gain or lose importance, merge, or split. Retaining an overview of the structure of a specific research field can be difficult due to these highly dynamic changes. Thus, ITGInsight includes a topic evolution model that detects emerging and evolving trends.

To begin assembling a picture of how interest in these topics has emerged and evolved, we first took the scientific papers and divided them into years. We then selected the top 30 most frequently mentioned terms for each year and performed co-word analysis with a text mining technique. The differences between each co-word network for each time period reveal how the topics have evolved (Wang et al., 2021a), Fig. 5 illustrates how the topics have gained or lost importance, merged, or split over the period of study. Each topic has a different color, and the thickness of the connection represents the strength between topics. This analysis reveals insights on two levels–first, it shows

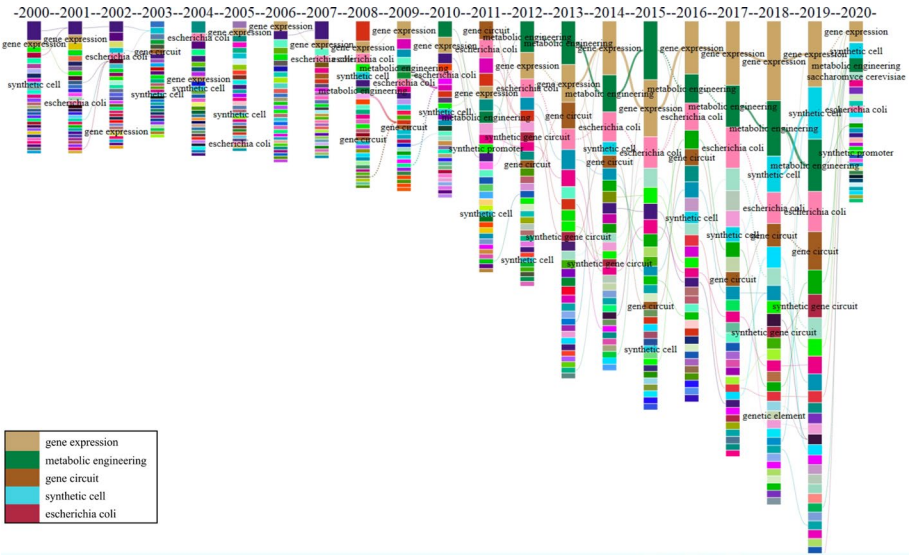
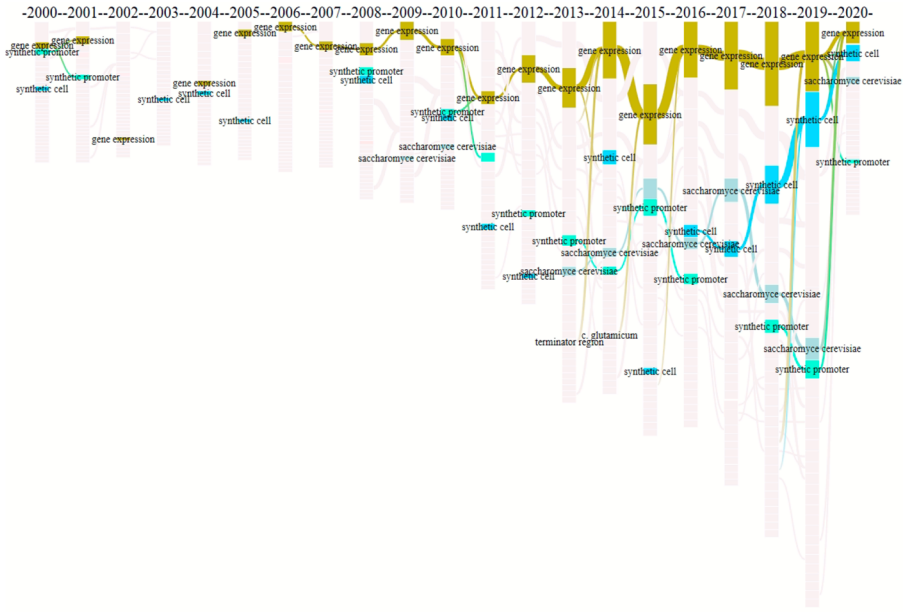


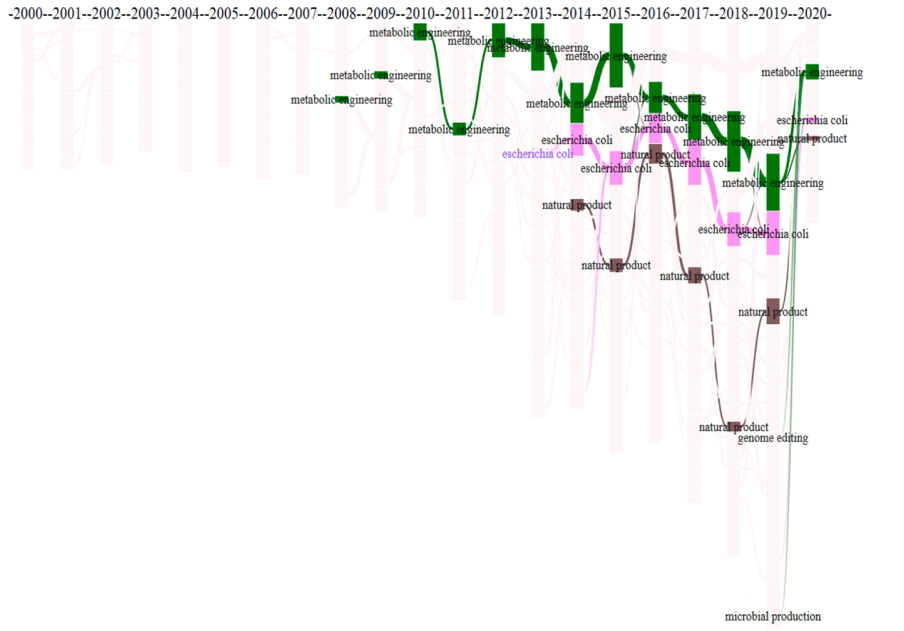
Fig. 5 Macro evolution in Research Front 1

some broad trends in the field and, second, it reveals a host of micro-level translations. At the macro level, scholars have paid attention to "gene expression" (Systems 2), "metabolic engineering" (Systems 3), and "gene circuit/synthetic gene circuit" (Devices 1). These are the terms of the highest item density. We can therefore infer that "Systems 2—Gene expression", "Systems 3—Metabolic engineering" and "Devices 1- Gene circuit" are not only important knowledge structures at present but will also continue to receive attention in the future. Further, "synthetic cell", "Escherichia coli", and "Saccharomyces cerevisiae" have gained more prominence in recent years. These also represent future development trends.

From Fig. 4, we find a close line exists between "Gene expression" (Systems 2) and "Genetic devices /Mathematic model (Devices 2)". This is because researchers regulate gene expression with the help genetic devices that use mathematic models. For some researchers, it is clear that enthusiasm for gene expression research is increasing. However, at the micro level, we find that other researchers are beginning to pay more attention to the relationship between "gene expression", "synthetic cell", "Saccharomyces cerevisiae", and "synthetic promoter" (Fig. 6a). Turning to the literature for verification, we find Saccharomyces cerevisiae can be used as chassis cells (Devices) in synthetic biology (Zhou et al., 2021). To date, the majority of natural products are extracted directly from organism, but this practice has gradually begun to be replaced by alternative approaches from synthetic biology. The reason being that natural extraction is inefficient and it consumes biological resources. One of these alternative approaches is modifying the host's metabolic pathways by adding heterometabolic pathways of Saccharomyces cerevisiae. Nowadays, the direct synthesis of target metabolites has become the research hot spot in synthetic biology, as has the metabolic engineering of Saccharomyces cerevisiae by optimizing and transforming the exogenous gene promoter. This works to regulate the expression level of the exogenous gene in the host and coordinate the metabolic pathways of multiple hosts. The terms identified span the structures and types of promoters as well the methods of optimizing the



(a) The evolution of gene expression (Systems 2)



(b) The evolution of "metabolic engineering" (Systems 3)

Fig. 6 Micro-level trends in Research Front 1

expressions and the processes of synthetizing natural products from the constructed Saccharomyces cerevisiae.

Figure 6 (b) shows us another emerging trend: "metabolic engineering" (Systems 3) and its strong relationship to "natural product" and "Escherichia coli". Escherichia coli is an excellent production host of natural products through metabolic engineering methods that can be used as chassis cells (Devices). The pyran ring is a very common structural unit of many natural, bioactive molecules that is widely found in plants, bacteria, and fungi. Yang et al. (2014) isolated a series of benzopyrans, named xiamenmycin, from Streptomyces Xiamen 318 for the first time. They believe xiamenmycin can be used as a drug candidate for anti fibrosis activity. Further, a series of "unnatural precursors" of benzopyran were obtained by combinatorial biosynthesis of isopentenyl transferase (ximb) in the biosynthesis pathway of xiamenmycin. In addition, heterologous biosynthesis of xiamenmycin in Escherichia coli was realized through metabolic pathway reconstruction. This work not only enriched the chemical structure diversity of benzopyran compounds, but also laid a foundation for the construction of efficient biosynthesis system of benzopyran active natural products (He et al., 2018).

In general, in Research Front 1, "Systems 2—Gene expression", "Systems 3—Metabolic engineering" and "Devices 1—Gene circuit/Synthetic gene circuit" have attracted continuous attention. However, the trend analysis at the knowledge innovation level also reflects changes in the focal points of researchers. For example, "Systems 2—Gene expression" and "Devices 2 – Genetic devices (Mathematic model)", as two main knowledge structures, have a strong relationship in Fig. 4. But with the deepening of research, researchers began to focus on how to use Saccharomyces cerevisiae as chassis cells and regulate genes in synthetic cells at the expression level by optimizing and transforming an exogenous gene promoter. Meanwhile, the research of Systems 3—Metabolic engineering is growing in prominence. Researchers are beginning to use Escherichia coli as an excellent production host of natural products through metabolic engineering methods. These directions represent the focal points of synthetic mammalian gene circuit in the future and are confirmed by the research of Yu et al. (2021) .

Conclusions

ITGInsight is an advanced text mining and visualization tool that can be used by a wide range of researchers from different disciplines to discover and visualize research fronts and knowledge structures from the scientific literatures. In terms of method innovation, ITGInsight offers a generic computational framework that relies on a topic clustering model and a topic evolution model to profile a research domain in terms of its research fronts and further track the evolution trends of the knowledge structures behind each research front for the purposes of identifying knowledge innovation.

The results of the case study are encouraging as we have demonstrated that ITGInsight is capable of answering the following questions: (i) What are the research fronts of a research domain? What are the knowledge structures of each research front? (ii) What are the trends and directions of each knowledge structure? and How do they promote the development of a research front? Further, ITGInsight has three practical advantages, as follows. 1) Compared to burst terms or subject words, the domain terms obtained by the PC-Value method more appropriately represent the essence or knowledge of a research front. 2) The topic clustering model, which combines the LinLog algorithm (Noack, 2004) with a pixel

density function (Li u et al., 2017) for discovering the knowledge structures in a research front not only provides a network graph of more uniform density and clearly delineated clusters, but it also displays the results in topographic map form. 3) The topic evolution model helps researchers to trace the emerging and evolutionary trends in research focuses over time. It reveals insights on two levels – first, some broad trends in the field and, second, a host of micro-level interpretations.

From a practical standpoint, most organizations today are facing the challenge of tracking the latest technological developments. Recognizing technological opportunities and identifying competitive threats is a part of daily business. Maintaining knowledge regarding key technology developments and predicting key technologies is critical (Morris et al., 2014). In addition to opportunities to expand horizontally, a number of application domains are particularly in need of trend detection and topic-tracking tools, such as high tech fields and areas related to national security like intelligent analysis. A tool-like ITGInsight could enable researchers to perform quantitative and qualitative studies of scientific-subject domain more easily and assist policy makers to draw up government development blueprint planning.

Despite these meaningful contributions, ITGInsight can be further developed in a number of ways. First, domain knowledge will affect the amount and specificity of information ITGInsight provides. By drawing on the strengths from multiple disciplines, ITGInsight could become a melting pot that turns otherwise isolated techniques into an integrated ecosystem. Secondly, technology-oriented analysis is evolving as part of ongoing strategy or policymaking. It is a dynamic process, and this software allows for regular updating and evaluation as part of what is for most organizations an ongoing demand. Our further work will further explore this link to ongoing future-oriented analyses.

Acknowledgements This work is partly supported by the General Program of the National Natural Science Foundation of China (Grant No.72074020, 71774012). The previous version of this work is published on Artificial Intelligence + Informetrics (AII) 2021 Workshop (Wang et al., 2021a), and the findings and observations present in this paper are those of the authors and do not necessarily reflect the views of the supporters or the sponsors. We are grateful to many scholars and software enthusiasts who provide their valuable opinions and suggestions in the process of ITGInsight design and development. Users could download and install the latest version of ITGInsight from http://en.itginsight.com/download/.

Author's contribution XW conceived and designed the research framework and partially developed the software. YL conceived and designed the analysis and developed the software. SZ wrote the paper and performed the analysis.

References

Barnes, J., & Hut, P. (1986). A hierarchical O(N log N) force-calculation algorithm. *Nature, 324*(6096), 446–449.

Behan, F. M., Iorio, F., Picco, G., Gonçalves, E., Beaver, C. M., Migliardi, G., Santos, R., Rao, Y., & Sassi, F. (2019). Prioritization of cancer therapeutic targets using CRISPR–Cas9 screens. *Nature, 568*(7753), 511–516.

Bowen, A., & Casadevall, A. (2015). Increasing disparities between resource inputs and outcomes, as measured by certain health deliverables, in biomedical research. *Proceedings of the National Academy of Sciences, 112*(36), 11335–11340.

Boyack, K. W., & Klavans, R. (2010). Co-citation analysis, bibliographic coupling, and direct citation: Which citation approach represents the research front most accurately? *Journal of the American Society for Information Science and Technology, 61*(12), 2389–2404.

Breitling, R., & Takano, E. (2015). Synthetic biology advances for pharmaceutical production. *Current Opinion in Biotechnology, 35C*, 46–51.

Breitling, R., Takano, E., & Gardner, T. S. (2015). Judging synthetic biology risks. *Science, 347*(6218), 107.

Casadevall, A., & Fang, F. C. (2014). Causes for the persistence of impact factor mania. *Mbio, 5*(3), e00064–14.

Chen, C. (2006). CiteSpace II: Detecting and visualizing emerging trends and transient patterns in scientific literature. *Journal of the American Society for Information Science and Technology, 57*(3), 359–377.

Chen, X., & Liu, L. (2018). Gene Circuits for Dynamically Regulating Metabolism. *Trends in Biotechnology, 36*(8), 751–754.

Chen, J., & Yang, L. (2021). A Bibliometric Review of Volatility Spillovers in Financial Markets: Knowledge Bases and Research Fronts. *Emerging Markets Finance and Trade, 57*(5), 1358–1379.

Cobo, M. J., Lopez-Herrera, A. G., Herrera-Viedma, E., & Herrera, F. (2012). SciMAT: A New Science Mapping Analysis Software Tool. *Journal of the American Society for Information Science and Technology, 63*(8), 1609–1630.

Fang, Y. S., & Lee, L. S. (2021). Research front and evolution of technology education in Taiwan and abroad: Bibliometric co-citation analysis and maps. *International Journal of Technology and Design Education*, 1–32.

Frantzi, K., Ananiadou, S., & Mima, H. (2000). Automatic recognition of multi-word terms: The C-value/NC-value method. *International Journal on Digital Libraries, 3*(2), 115–130.

Galitsky, L. M., Pottenger, W. M., Roy, S., & Phelps, D. J. (2004). *A Survey of Emerging Trend Detection in Textual Data Mining*. Springer.

García-Aljaro, C., Melado-Rovira, S., Milton, D. L., & Blanch, A. R. (2012). Quorum-sensing regulates biofilm formation in Vibrio scophthalmi. *BMC Microbiology, 12*, 287.

Garg, N., Da Manchan, G., & Kumar, A. (2014). Bacterial quorum sensing: Circuits and applications. *Antonie Van Leeuwenhoek, 105*(2), 289–305.

Han, H. Q., Zhu, D. H., & Wang, X. F. (2011). Technical term extraction method for patent document. *Journal of the China Society for Scientific and Technical Information, 30*(12), 1280–1285 (**in Chinese**).

He, B. B., Bu, X. L., Zhou, T., Li, S. M., Xu, M. J., & Xu, J. (2018). Combinatory Biosynthesis of Prenylated 4-Hydroxybenzoate Derivatives by Overexpression of the Substrate-Promiscuous Prenyltransferase XimB in Engineered E coli. *ACS Synthetic Biology, 7*(9), 2094–2104.

Huang, M. H., & Chang, C. P. (2014). Detecting research fronts in OLED field using bibliographic coupling with sliding window. *Scientometrics, 98*(3), 1721–1744.

Huang, Y., Zhang, Y., Ma, J., Porter, A. L., Wang, X. F., & Guo, Y. (2016). Generating Competitive Technical Intelligence Using Topical Analysis, Patent Citation Analysis, and Term Clumping Analysis. In *Anticipating Future Innovation Pathways Through Large Data Analysis*, 153–172. Cham: Springer International.

Joseph, J. M., Viswajit, K., Mihe, H., Patrick, S. D., et al. (2021). Model-guided design of mammalian genetic programs. *Science Advances, 7*(8), eabe9375.

Karafyllidis, I. G. (2012). Quantum Gate Circuit Model of Signal Integration in Bacterial Quorum Sensing. *Transactions on Computational Biology & Bioinformatics IEEE/ACM, 9*(2), 571–579.

Keiser, J., & Utzinger, J. (2005). Trends in the core literature on tropical medicine: A bibliometric analysis from 1952–2002. *Scientometrics, 62*(3), 351–365.

Lee, Y., Kim, S. Y., Song, I., Park, Y., & Shin, J. (2014). Technology opportunity identification customized to the technological capability of SMEs through two-stage patent analysis. *Scientometrics, 100*(1), 227–244.

Lee, J., Kim, C., & Shin, J. (2017). Technology opportunity discovery to R&D planning: Key technological performance analysis. *Technological Forecasting and Social Change, 119*, 53–63.

Li, M., & Chu, Y. (2016). Explore the research front of a specific research theme based on a novel technique of enhanced co-word analysis. *Journal of Information Science, 43*(6), 725–741.

Li, X., Jiang, W., Liang, Q., & Qi, Q. (2020). Application of bacterial quorum sensing system in intercellular communication and its progress in synthetic biology. *Synthetic Biology Journa, 1*(5), 42–57.

Liao, S. H., Sun, B. L., & Wang, R. Y. (2003). A knowledge-based architecture for planning military intelligence, surveillance, and reconnaissance. *Space Policy, 19*(3), 191–202.

Liu, J. S., Lu, L., & Lu, W. M. (2015a). Research Fronts in data envelopment analysis. *Omega, 58*, 33–45.

Liu, Z., Yin, Y., Liu, W., & Dunford, M. (2015b). Visualizing the intellectual structure and evolution of innovation systems research: A bibliometric analysis. *Scientometrics, 103*, 135–158.

Liu, Y. Q., Pang, J. H., Cui, Z. C., Wang, X. F., & Gui, J. (2017). An economic method of drawing a technology theme map. *Library and Information Service, 61*(13), 125–132 (**in Chinese**).

Lucentini, L. (2006). Just what is synthetic biology. *Scientist, 20*, 36.

Ma, V. C., & Liu, J. S. (2016). Exploring the research fronts and main paths of literature: A case study of shareholder activism research. *Scientometrics, 109*(1), 33–52.

Mane, K. K., & Borner, K. (2004). Mapping topics and topic bursts in PNAS. *Proceedings of the National Academy of Sciences of the United States of America, 101*(1), 5287–5290.

Miller, M. B., & Bassler, B. L. (2001). Quorum sensing in bacteria. *Annual Review of Microbiology, 55*, 165–199.

Morris, S. A., Yen, G., Wu, Z., & Asnake, B. (2003). Time line visualization of research fronts. *Journal of the American Society for Information Science and Technology, 54*(5), 413–422.

Morris, S. A., Yen, G., Zheng, W., & Asnake, B. (2014). Time line visualization of research fronts. *Journal of the Association for Information Science and Technology, 54*(5), 413–422.

Nissim, L., Wu, M. R., Pery, E., Binder-Nissim, A., Suzuki, H. I., Stupp, D., Wehrspaun, C., Tabach, Y., Sharp, P. A., & Lu, T. K. (2017). Synthetic RNA-Based Immunomodulatory Gene Circuits for Cancer Immunotherapy. *Cell, 171*(5), 1138-1150.e15.

Noack, A. (2004). An Energy Model for Visual Graph Clustering. *Proceedings of the 11th International symposium on Graph Drawing*, *29* (12), 425–436.

Persson, & Olle. (1994). The Intellectual Base and Research Fronts of JASIS 1986–1990. *Journal of the American Society for Information Science, 45*(1), 31–38.

Pieiro-Chousa, J., López-Cabarcos, M., Romero-Castro, N. M., & Pérez-Pico, A. (2019). Innovation, entrepreneurship and knowledge in the business scientific field: Mapping the research front. *Journal of Business Research, 115*, 475–485.

Ping, X. (2015). Study of international anticancer research trends via co-word and document co-citation visualization analysis. *Scientometrics, 105*(1), 611–622.

Porter, A. L., & Cunningham, S. W. (2005). *Tech mining : Exploiting new technologies for competitive advantage.* Hoboken, New Jersey: Wiley-Interscience.

Price, D. (1965). Networks Of Scientific Papers. *Science, 149*(3683), 510–515.

Roybal, K., Williams, J., Morsut, L., Rupp, L., & Lim, W. (2016). Engineering T Cells with Customized Therapeutic Response Programs Using Synthetic Notch Receptors. *Cell, 167*(2), 419-432.e416.

Shibata, N., Kajikawa, Y., Takeda, Y., & Matsushima, K. (2008). Detecting emerging research fronts based on topological measures in citation networks of scientific publications. *Technovation, 28*(11), 758–775.

Shibata, N., Kajikawa, Y., Takeda, Y., & Matsushima, K. (2009). Comparative study on methods of detecting research fronts using different types of citation. *Journal of the American Society for Information Science and Technology, 60*(3), 571–580.

Small, H. (2006). Tracking and predicting growth areas in science. *Scientometrics, 68*(3), 595–610.

Small, H., & Griffith, B. C. (1974). The Structure of Scientific Literatures I: Identifying and Graphing Specialties. *Social Studies of Science, 4*, 17–40.

Strotmann, A., & Zhao, D. (2014). The Knowledge Base and Research Front of Information Science 2006–2010: An Author Cocitation and Bibliographic Coupling Analysis. *Journal of the American Society for Information Science and Technology, 65*(5), 995–1006.

Swofford, C. A., De Ssel, N. V., & Forbes, N. S. (2015). Quorum-sensing Salmonella selectively trigger protein expression within tumors. *Proceedings of the National Academy of Sciences of the United States of America, 112*(11), 3457–3462.

Tian, Z., Wang, Z., Liu, Z., Xiang, H., Liu, J., & Zheng, Q. (2012). Learning to identify core term of knowledge unit from short text. *International Conference on Fuzzy Systems & Knowledge Discovery,* 1303–1308.

Tijssen, R. (2002). Science dependence of technologies: Evidence from inventions and their inventors. *Research Policy, 31*(4), 509–526.

Upham, S. P., & Small, H. (2010). Emerging research fronts in science and technology: Patterns of new knowledge development. *Scientometrics, 83*(1), 15–38.

Wang, J., & Chen, Y. J. (2019). A novelty detection patent mining approach for analyzing technological opportunities. *Advanced Engineering Informatics, 42*, 100941.

Wang, X. W., Wang, Z., & Xu, S. M. (2013). Tracing scientist's research trends realtimely. *Scientometrics, 95*(2), 717–729.

Wang, X. F., Li, R. R., Ren, S. M., Zhu, D. H., Huang, M., & Qiu, P. J. (2014). Collaboration network and pattern analysis: Case study of dye-sensitized solar cells. *Scientometrics, 98*(3), 1745–1762.

Wang, X. F., Zhang, S., & Liu, Y. Q. (2021a). ITGInsight - Discovering and Visualizing Science, Technology and Innovation Information for Generating Competitive Technological Intelligence. *Proceedings of the 1st Workshop on AI + Informetrics (AII2021) co-located with the iConference 2021*, 202–219.

Wang, X. F., Zhang, S., Liu, Y. Q., Du, J., & Huang, H. (2021). How pharmaceutical innovation evolves: The path from science to technological development to marketable drugs. *Technological Forecasting and Social Change, 167*(43), 120698.

Xie, S., Zhang, J., & Ho, Y. S. (2008). Assessment of world aerosol research trends by bibliometric analysis. *Scientometrics, 77*(1), 113–130.

Yan, B., Lee, T., & Lee, T. (2015). Mapping the intellectual structure of the Internet of Things (IoT) field (2000–2014): A co-word analysis. *Scientometrics, 105*(2), 1285–1300.

Yang, Y., Fu, L., Zhang, J., Hu, L., Xu, M., & Xu, J. (2014). Characterization of the Xiamenmycin Biosynthesis Gene Cluster in Streptomyces xiamenensis 318. *Plos One, 9*(6), e99537.

Ye, Y., Zhang, L., Zhao, X., & Ronald, R. (2012). An Experimental Study on Revealing Dom ain Knowledge Structure by Co-keyword Networks. *Journal of the China Society for Scientific and Technical Information, 31*(12), 1245–1251.

Yi, W., & Di, M. (2016). The research fronts and hotspots on nanotechnology based on journal of vacuum science & technology. *Open Journal of Social Sciences, 4*(3), 57–65.

Yoon, B., Park, I., & Coh, B.-Y. (2014). Exploring technological opportunities by linking technology and products: Application of morphology analysis and text mining. *Technological Forecasting & Social Change, 86*, 287–303.

Yoon, J., Park, H., Seo, W., Lee, J., Coh, B.-Y., & Kim, J. (2015). Technology opportunity discovery (TOD) from existing technologies and products: A function-based TOD framework. *Technological Forecasting and Social Change, 100*, 153–167.

Yu, Y., Zhu, X. N., Bi, C. H., & Zhang, X. L. (2021). Construction of Escherichia coli cell factories. *Chinese Journal of Biotechnology, 37*(5), 1564–1577 (**in Chinese**).

Zhang, Y., Guo, Y., Wang, X. F., Zhu, D. H., & Porter, A. L. (2013). A hybrid visualisation model for technology roadmapping: Bibliometrics, qualitative methodology and empirical study. *Technology Analysis & Strategic Management, 25*(6), 707–724.

Zhang, Y., Robinson, D., Porter, A. L., Zhu, D. H., Zhang, G. Q., & Lu, J. (2016). Technology roadmapping for competitive technical intelligence. *Technological Forecasting and Social Change, 110*, 175–186.

Zhou, M. Y., Bi, Y. H., Ding, M. Z., & Yuan, Y. J. (2021). One-Step Biosynthesis of Vitamin C in Saccharomyces cerevisiae. *Frontiers in Microbiology, 12*, 643472.

Zhu, D. H., & Porter, A. L. (2002). Automated extraction and visualization of information for technological intelligence. *Technological Forecasting and Social Change, 69*(5), 495–506.