Technological Forecasting & Social Change 92 (2015) 182–195

Contents lists available at ScienceDirect

Technological Forecasting & Social Change

Exploring technological opportunities by mining the gaps between science and technology: Microalgal biofuels

Ming-Yeu Wang [a,*], Shih-Chieh Fang [b,c], Yu-Hsuan Chang [d]

[a] Department of BioBusiness Management, National Chiayi University, Chiayi, Taiwan
[b] Department of Business Administration and Institute of International Business, National Cheng Kung University, Tainan, Taiwan
[c] Center for Energy Technology and Strategy, National Cheng Kung University, Tainan, Taiwan
[d] Gainia Intellectual Asset Services, Inc., Hsinchu, Taiwan

ARTICLE INFO

Article history:
Received 3 November 2012
Received in revised form 1 June 2014
Accepted 14 July 2014
Available online 30 August 2014

Keywords:
Science and technology
Technological opportunity
Text mining
Microalgae
Biofuel
High-dimensional data

ABSTRACT

The interaction between scientific and technological knowledge facilitates exploration of new technological opportunities; however, gaps between them typically impede exploration of these opportunities. Scientific papers and technological patents record modern and advanced knowledge in scientific discovery and technological development; therefore, comparing their statuses can identify the gaps and explore potential technological opportunities. Because microalgal biofuels are a promising alternative energy resource devoid of territorial land use problems, this study applies text mining and an algorithm that can cluster objects of high-dimensional data to microalgal biofuel papers and patents, and explores their potential technological opportunities. The results demonstrate that a text-based clustering approach is appropriate for identifying scientific and technological applications for microalgal biofuels. The results indicate that microalgal photosynthesis and light utilization have abundant scientific outcomes for technological engineers to potentially apply. Technological opportunities exist in synthesis, harvesting, extraction, and lipid conversion. Scientific knowledge underlying biofuels accompanying high-value co-products of production require sustained exploration and reporting through research. These needs represent potential technological opportunities.

1. Introduction

Technological opportunity is the potential for technological progress in general or within a particular field (Klevorick et al., 1995; Olsson, 2005; Frenz and Prevezer, 2012) that affects the overall industry and individual enterprises. At the industrial level, technological opportunities determine technological development (Olsson, 2005), affect industry R&D intensity, and lead to heterogeneous R&D productivity in different industries (Klevorick et al., 1995; Olsson, 2005). At the enterprise level, technological opportunities affect R&D costs and innovation activities, resulting in different R&D productivity and operating performance among enterprises (Frenz and Prevezer, 2012; Nieto and Quevedo, 2005; Cohen et al., 1987; Cohen and Levinthal, 1989). In addition to affecting current enterprises, technological opportunities can facilitate the startup of new businesses and successful commercialization of entrepreneurs' inventions (Shane, 2001). Because technological opportunities profoundly influence industries and enterprises, systematic methods of exploring potential technological opportunities undoubtedly benefit industries and enterprises.

Klevorick et al. (1995) identify three sources that contribute to an industry's technological opportunities: advances in scientific knowledge, technological advances originating outside the industry, and feedback from industrial technology. Among those, advances in scientific knowledge are most powerful. It provides an expanding pool of theory, technique,

* Corresponding author. Tel.: +886 5 2732883; fax: +886 5 2732874.
E-mail addresses: mingyeu.wang@gmail.com, mywang@mail.ncyu.edu.tw (M.-Y. Wang), fangsc@mail.ncku.edu.tw (S.-C. Fang), shikodoofmy@gmail.com (Y.-H. Chang).

http://dx.doi.org/10.1016/j.techfore.2014.07.008
0040-1625/

and problem solving capability that industrial R&D uses, and unlocks new technological capabilities (Klevorick et al., 1995; Narin et al., 1997; Meyer, 2000). The interaction between scientific and technological knowledge nourishes exploration of new technological opportunities. Scientific knowledge lays the foundation for technological knowledge and provides better solutions for product commercialization. Feedback stimulus from technological knowledge can promote continuous exploration and study of scientific knowledge (Glänzel and Meyer, 2003; Ziman, 1988; Meyer, 2002).

Scientific and technological knowledge are complementary, but the gap between science and technology hinders the development of technological opportunities. Previous studies on exploration of technological opportunities have, therefore, used a single knowledge source. For example, Yoon and Park (2005) and Yoon and Kim (2012) explore technological opportunities based on technological knowledge alone. In contrast, Shibata et al. (2010) successfully combine scientific and technological knowledge sources to analyze the gap between them. Their method employs a citation-based clustering approach that identifies citation networks and further compares scientific clusters with technological clusters to identify commercialization gaps. The extracted commercialization gap is equivalent to the potential of technological opportunities. However, Shibata et al. (2010) analyze only the largest component in the citation networks and remove isolated nodes and components not connected with the largest one. This study, however, observes that citations among literature materials in some emerging technological fields occur infrequently. Applying their methods may omit a large portion of literature data, losing important information. To solve this problem, a text-based clustering approach that analyzes nonstructured text through text mining, which replaces the citation network for identifying scientific and technological clusters, and extracts commercial gaps and potential technological opportunities, is adopted.

This approach extracts representative words and terms–features–from documents. Because the literature typically contains hundreds to thousands of features, the matrices representing the relationship between documents and features are high-dimensional spaces; thus, clustering these objects poses challenges presented by the "curse of dimensionality" (or "distance concentration effect"), whereby similarity measures cannot discriminate between the nearest and farthest neighbors for a given object (Beyer et al., 1999). The vector space extracted by text mining contains more or less irrelevant features, which may conceal relevant features and confuse the clustering process (Kriegel et al., 2009). Therefore, a suitable clustering algorithm is selected for our empirical technology.

Due to serious global climate change, the development of clean, environmentally-friendly, and renewable energies has become an important strategy in many countries (Rosenberg, 1982). Biofuel is the fourth largest energy source, after petroleum, coal, and natural gas. Although biofuel can reduce CO_2 emission, planting biofuel crops requires large land areas and may reduce space for grain crops. Microalgae are a type of amphibious plants, and thus, are unaffected by land area restrictions. The oil content of microalgal cells is higher than that of any land plant, producing higher biofuel yields; moreover, it has the advantage of rapid growth (Smith et al., 2010). However, microalgal biofuel production cost remains higher than that of fossil fuel (Davis et al., 2011). Exploring the gaps between scientific research and technological development provides useful clues to technological opportunities for overcoming the challenge in microalgal biofuel production systems. The citation network relationships among scientific and technological literature are not frequent because microalgal biofuels remain a newly emerging field. Therefore, text-based clustering is more appropriate than citation-based clustering for exploring technological opportunities in microalgal biofuels.

Therefore, on the basis of scientific and technological literature related to microalgal biofuels and by employing text-based clustering for analysis, this study examines current major scientific and technological research fields, and then explores future potential technological opportunities by exploring the gaps between them. Finally, this study suggests future R&D directions for microalgal biofuels.

2. Related research

2.1. Technological opportunity

Scientific technological knowledge and technological knowledge within and beyond industries are important sources of technological opportunities (Klevorick et al., 1995). The interaction between science and technology has been studied since Price's (1965) research. Price finds that science and technology are independent, and accumulate their own knowledge structure; in special cases, the two interact. Since then researchers have found that interactions between science and technology have become increasingly active. Rich scientific research fields can stimulate innovation and technological development (e.g., Rosenberg, 1982); conversely, technologies with less scientific exploration may also inspire important scientific breakthroughs. Additionally, basic science advancement requires support from more advanced technologies (e.g., Nelson, 1982). Thus, science and technology are interdependent (Meyer, 2002; Petrescu, 2009).

Despite this interdependence, a gap exists between them. Scientists may be unaware of the application of their discoveries, while manufacturers are often unaware which scientific discoveries can improve their technological development and commercialization (Hellmann, 2007). Scientific papers and technological patents present the results of scientific discovery and technological development (Martino, 2003; Robinson et al., 2013; Kostoff, 2006); thus, an analysis and comparison of scientific papers and technological patents can determine the gap and identify technological opportunities (Shibata et al., 2010).

Some previous studies exploring technological opportunities are based on technological patents alone, whereas others combine scientific papers and technological patents. Among the patent studies, Yoon and Park (2005) use text mining to extract patents' keywords, through which they establish a technological dictionary and morphology. The undeveloped morphological combination indicates potential technological opportunities. The method of exploring technological opportunities proposed by Shibata et al. (2010) combines science and technology. They use social network analysis to establish citation networks of science and technology based on citation relationships in paper and patent references. After extracting the papers and patents in the largest components from both network diagrams, they identify scientific and technological

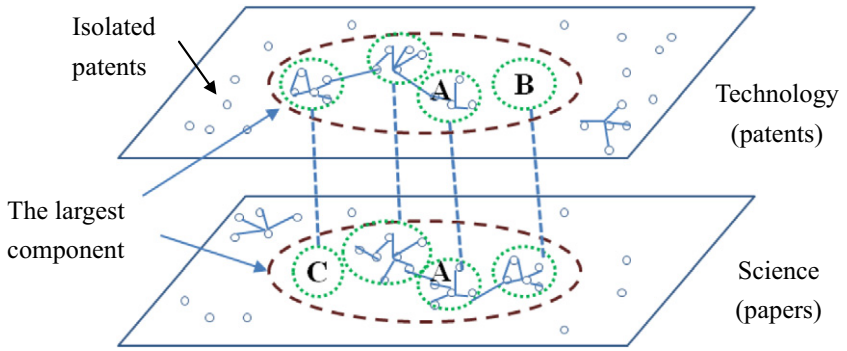

Fig. 1. Relationships between science and technology.
Source: Modified from Shibata et al. (2010).

clusters in both components. By comparing them, they derived three relationships as shown in Fig. 1.

In Fig. 1, Cluster A is a field containing both technological and scientific literature, where science and technology interact and co-evolve. Cluster B is an existing scientific research field without technological development, presenting a gap between science and technology that indicates potential technological development or commercialization opportunities. Cluster C is a field in which technological developments appear but lack scientific research, also presenting a gap. This cluster has no scientific work and requires advanced technological support; therefore, this gap indicates opportunities for scientific research advancement. In addition, Shibata et al. (2010) describe a field that both scientific research and technological development failed to explore, not shown in Fig. 1.

The method proposed by Shibata et al. (2010) analyzes only the largest components in the citation network, and remove isolated nodes and components not connected to the largest ones. As shown in Fig. 1, scientific papers and technology patents beyond the circle consisting of the largest components are omitted. Some emerging technologies may have had insufficient time to accumulate knowledge and form a representative component; therefore, the removal process may lose potentially important information.

Scientific papers and technological patents include structured and unstructured items. Structured items contain uniform fields and formats, such as publication date and patent number, whereas papers' and patents' context information is an example of unstructured items. Text mining can analyze the natural language of unstructured text to extract useful technological information (Kostoff, 2008). This study uses a method proposed by Shibata et al. (2010) as its base and employs text mining rather than social network analysis in identifying scientific and technological clusters.

Text mining integrates approaches of data mining, machine learning, natural language processing, information retrieval, and knowledge management. It extracts effective, non-trivial, hidden, previously unknown, and potentially useful knowledge from non-structured or semi-structured texts (Feldman and Sanger, 2007; Weiss et al., 2005). The approach can handle a large volume of unstructured text and mine important hidden information from a set of documents, called a corpus. For paper and patent corpuses, text mining has many potential applications. Previous studies have used text mining or combined text mining with other methods to mine different knowledge and information from academic papers or patent documents. For example, Wang et al. (2010) combine text mining and TRIZ (Theory of Inventive Problem Solving) to investigate technological evolutionary trends. Combining text mining, network analysis, and citation analysis, Lee et al. (2009) visualize patent information that can support enterprises in discovering business opportunities. Wu et al. (2011) combine text mining and bibliometric analysis to explore technology trends. To overcome the problem that invention, traditionally, requires scientists' spontaneous creativity or countless attempts, Kostoff (2008) proposes a literature-related discovery approach, where text mining analyzes literature to suggest potential discoveries.

2.2. Text mining and clustering for high-dimensional data

Feldman and Sanger (2007) explain that text mining procedures include the preprocessing of document collections, storage of intermediate representations, techniques to analyze these intermediate representations, and visualization of the results. Preprocessing of document collections involves an attempt to convert unstructured texts into structured textual data for computer processing. Therefore, preprocessing involves stemming, stopword removal, and extraction of representative words or terms, which are called the features of the document. The relationships between documents and feature extractions are often presented by the vector space model (VSM) proposed by Salton et al. (1975). In the VSM model, the relationships between documents and features are projected in a multi-dimensional Euclidean space. A document is presented as a vector containing weights of several features; a corpus is a vector space composed of several documental vectors. This vector space is the structured textual data that can be analyzed by computers.

Cluster analysis reveals groups of similar documents. Cluster detection is based on similarity between documents, typically determined using measures of the dimensions in vector space. When the number of dimensions increases, all documents in the high-dimensional vector space must be nearly equidistant from each other (Parsons et al., 2004). Thus, conventional clustering algorithms cannot detect meaningful clusters. Beyer et al. (1999) called this the “curse of dimensionality.” Vector space transformed from a corpus typically has numerous features, so the space is a high-dimensional space. Therefore, conventional clustering algorithms may not identify meaningful clusters. Aside from the problem of similarity measures,

vector space generated from a corpus is typically very sparse, featuring many zero values in the matrix (Kriegel et al., 2009). Some features may be irrelevant to the themes, which may confuse clustering algorithms (Parsons et al., 2004).

Previous studies typically used feature transformation and selection techniques to cluster objects in high-dimensional space. Feature transformation techniques attempt to summarize a dataset by creating combinations of the original features: principle component analysis and singular value decomposition are two well-known techniques. However, transformations generally preserve the original relative distances between documents; thus, information from irrelevant dimensions may mask meaningful clusters if the data contain many irrelevant dimensions. Feature selection discovers and retains the features of a dataset that are most relevant to research task (Parsons et al., 2004); however, it may not always be feasible to prune excessive dimensions without information loss when some document features are differently correlated (Aggarwal and Yu, 2000). Therefore, feature selection techniques may be ineffective when dimensions have locally varying relevance for different clusters of documents. Fig. 2 illustrates the concepts that documents have locally varying relevance and possible clusters in a vector space. Cluster 3 represents a traditional cluster in global space, whereas clusters 1, 2, and 4 appear only in a subset of relevant dimensions, which present a specific theme. Note that D8 may be assigned to more than one cluster (Müller et al., 2009).

Feature transformation and selection techniques handle clustering problems in a global space by computing only one subspace of the original data space, wherein the clustering can then be performed (Kriegel et al., 2009; Müller et al., 2009). They perform "ineffectively" to detect clusters because each cluster may exist in a different subspace. In clustering documents, it is common to find that each cluster exists in a different subspace. Related documents only have a similar word weights within a subset of terms that are likely to be different for different groups of thematically relevant documents (Kriegel et al., 2009).

Recently, scholars have taken the concept of feature selection one step further by selecting relevant subspaces for each cluster separately and developing clustering algorithms for high-dimensional space. Numerous possible subspaces can be selected. To obtain high-quality clusters, given reasonable computation time, the first class of algorithms focuses only on clusters in axis-parallel subspaces. These algorithms are called projected clustering, or subspace clustering algorithms. The second class searches for subspace solutions where clusters may exist in any arbitrarily oriented subspace, and are usually called generalized subspace/projected clustering, or correlation clustering algorithms. The third class lies between the above two and are referred to as pattern-based clustering algorithms (Kriegel et al., 2009).

From the application viewpoint, the second class is generalized, concise, and meaningful compared with the other two. ORCLUS (arbitrarily Oriented projected CLUSter generation), is the first proposed generalized projected clustering algorithm (Aggarwal and Yu, 2000). This algorithm arose from the observation that many datasets contain inter-feature correlations, and is a *k*-means-like approach. It includes three steps: assigning clusters, finding subspaces, and merging clusters. During cluster assignment, the algorithm iteratively assigns each object to its closest seed. The distance between two points is measured in subspace *E*, where *E* is a set of orthonormal vectors in some *d*-dimensional space. Finding the subspaces redefines the subspace *E* associated with each cluster by calculating the covariance matrix for a cluster and selecting the orthonormal eigenvectors with the smallest eigenvalues. The selected eigenvectors correspond to the projected subspace, where the clustered objects exhibit high density, and hence can exclude the most noisy subspaces. Calculation and selection are iteratively adapted to the current state of the updated cluster. As a result, each successive iteration continues to strip noisy subspaces from different clusters. The third step merges nearby clusters that have similar directions of high density. The number of clusters and size of the subspace dimensionality must be specified by the researcher. The cluster sparsity coefficient can statistically evaluate the choice of subspace dimensionality (Kriegel et al., 2009; Parsons et al., 2004; Aggarwal and Yu, 2000). If the value approaches 1, then the chosen subspace dimension may be too large. A value close to 0 can be interpreted as a hint that a strong cluster structure has been found (Szepannek, 2013).

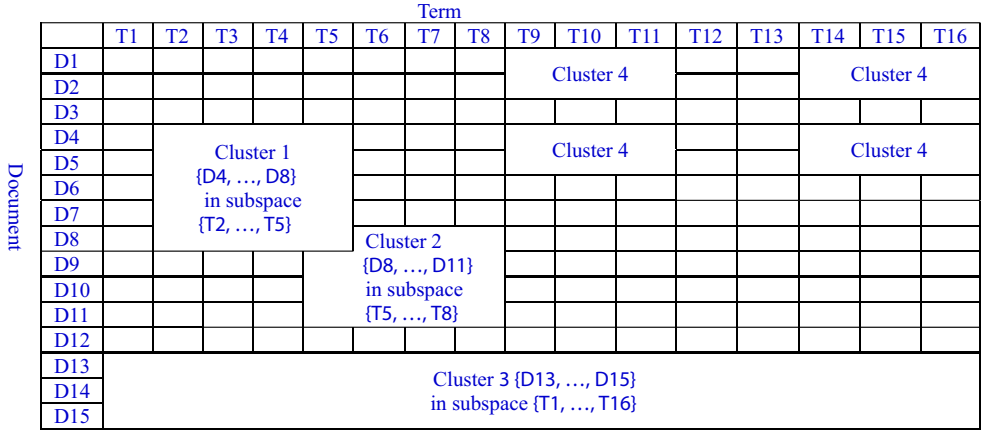

Fig. 2. Conceptualization of subspace clustering.
Source: Modified from Müller et al. (2009).

2.3. Microalgal biofuel

Among several renewable energies, biofuels possess several advantages, such as reducing carbon emissions and lower waste content. However, the problem of requiring substantial land for cultivating most biofuel sources impedes the development and usage of biofuels as a feasible option for renewable energies. Microalgae are amphibious plants, presenting no land problem. They grow rapidly and contain more oil than any terrestrial biofuel source (Smith et al., 2010). Microalgae can be produced all year-round, yielding higher oil production than that of the best oilseed crops. For example, microalgal biodiesel yield containing only 30% oil by weight is 58,700 liters per hectare (L/ha), which is superior to 1190 L/ha for rapeseed, 1892 L/ha for Jatropha, and 2590 L/ha of Karanj (Pongamia pinnata) (Singh and Gu, 2010). This extracted oil can be used to produce many high energy density transport fuels, its biomass residues can be converted into biofuel through either the biochemical or thermochemical pathway (Smith et al., 2010). The primary pathway of biomass produced in microalgae is photosynthesis. Its reactants are carbon dioxide and water through light absorption, and the products are oxygen and glucose, the latter being used for ethanol or other biofuel production. Two systems enable microalgae cultivation: open pond and closed tubular photobioreactor (PBR) systems. Comparing the two systems, the open pond system includes lower capital investment and available technology, while its disadvantages are higher downstream processing cost, higher water usage, and lower flexibility to strain selection. The PBR system, in contrast, offers higher flexibility to strain selection and lower downstream processing cost and water usage (Davis et al., 2011).

3. Research process

The research process includes four stages: document collection, text mining, clustering for research and technological fields, and identification of technological opportunities. Fig. 3 presents the research process; the paragraphs below describe these stages in detail.

Step 1, document collection, includes scientific and technological documents. The Science Citation Index and patent databases are typical sources for basic research and technological development, respectively; therefore, this study collects scientific papers related to microalgal biofuels from the Science Citation Index-Expanded (SCIE) database and technological documents from the United States Patent and Trademark Office (USPTO).

The search strategy for retrieving documents related to microalgal biofuels was the submission of keywords to paper and patent databases (Konur, 2011). The first phase of submitted keywords comprised of several microalgae-related keywords, such as “microalgae,” “alga,” and the names of 32 microalgae species that yield biofuel. The identified microalgae-related documents to biofuels were then narrowed down by using a Boolean operator AND, with the second phase of keywords comprising of biofuel-related words, such as “bio-energy,” “bio-fuel,” “bio-diesel,” and “bio-hydrogen.” Table 1 displays detailed queries to retrieve paper documents.

When retrieving patent documents, the biofuel-related keywords in the second phase are modified to account for the particular wording and terminology used in patent applications being different from those found in other types of documents. Academic authors typically use precise wording in articles, whereas patent applicants are accustomed to submitting genus claims that embrace a class of entities characterized by a common property (Lefstin, 2008), to broaden the scope of patent protection. Taking bio-butanol from the biofuel-related keywords as an example, the word “bio” means that the butanol feedstock is biomass rather than fossil-based, and the bio-butanol has the same chemical properties as butanol, revealing that the word “butanol” is a general (genus) term for “bio-butanol.” Therefore, this study adopts general terms for each biofuel in the second phase of the keyword search, and general terms are obtained by removing “bio” from the lists of biofuel-related words in Table 1. Considering that broad keywords may include patents unrelated to microalgae biofuels, this study removes them by referring to the main international patent classifications (IPCs) designated to each patent after the patents are downloaded.

The searched fields in the patent database included title, abstract, and claim, while those in paper database were title (TI) and topic (TS). Documents published from 1990 to the end of 2013 were searched. Year 1990 was selected because the

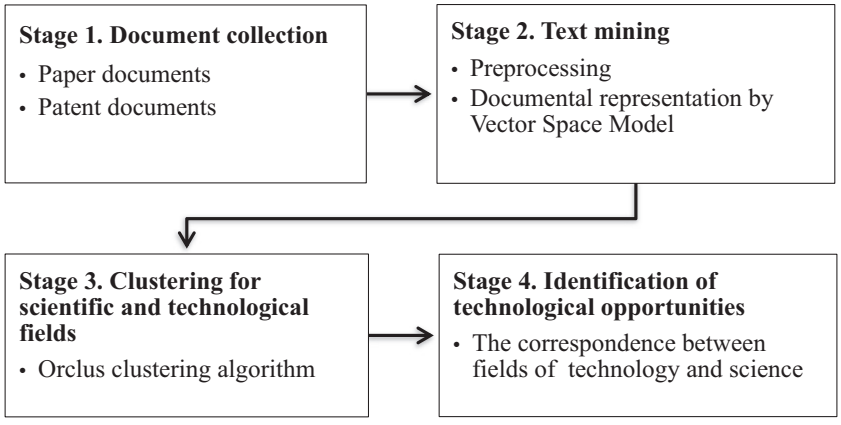

Fig. 3. Research process.

Table 1
Retrieval queries for microalgal biofuels.

Retrieval queries
((microalga* or micro-alga* or macroalga* or macro-alga* or alga or algal or algae) or (names of 32 microalgae species)) and (bio-energy or fuel* or biofuel* or bio-fuel* or biomethan* or bio-methan* or biodiesel* or bio-diesel* or biohydrogen or bio-hydrogen or bioethanol* or bio-ethanol* or biooil* or bio-oil* or bio-gas or bio-gas or biorefiner* or bio-refiner* or bioreactor* or bio-reactor*)

Note: The 32 microalgae species are *Ankistrodesmus* sp., *Beijerinck*, *Botryococcus braunii*, *Chaetoceros*, *Chlamydomonas reinhardtii*, *Chlorella*, *Chlorococcum*, *Crypthecodinium cohnii*, *Cylindrotheca* sp., *Diatoms*, *Dunaliella*, *Ellipsoidion* sp., *Euglena gracilis*, *Eustigmatophytes*, *Haematococcus pluvaris*, *Isochrysis*, *Monallanthus salina*, *Monodus subterraneus*, *Nannochloris* sp., *Nannochloropsis*, *Neochloris oleabundans*, *Nitzschia* sp., *Oocystis pusilla*, *Pavlova*, *Phaeodactylum tricornutum*, *Prymnesiophytes*, *Scenedemus*, *Schizochytrium* sp., *Skeletonema*, *Spirulina*, *Tetraselmis*, and *Thalassioria pseudonana*.

Intergovernmental Panel on Climate Change (IPCC), an organization established by the Union Nations, published the First Assessment Report on Climate Change in 1990 (IPCC, 1992), likely to promote micro-algal biofuel research and development.

The second stage conducted preprocessing and documental representation by VSM. The mined text was in the fields of title, abstract, and keywords for paper documents, and in title, abstract and the first independent claim for patent documents. In preprocessing, this study erased word suffixes to retrieve their radicals; removed stopwords, number, and punctuation; eliminated whitespace; and converted characters to lower case. Extracted terms called sparse terms did not appear in most documents. Because their usefulness is low, this study excluded them to reduce noise.

The purpose of documental representation by VSM is to organize a document-by-term matrix, where each cell records the term's importance in a given document. This study adopts the widely used frequency-inverse document frequency (TF-IDF) weighting measures (Feldman and Sanger, 2007).

3.1. Stage 3. Clustering for scientific and technological fields

The extracted document-by-term matrices in stage 2 are two high-dimensional spaces. The extracted terms could be inter-correlated. ORCLUS, a generalized projected clustering algorithm, is used to cluster the paper and patent documents. To implement ORCLUS, we prespecify four parameters: final number of clusters (k), dimensionality of subspaces where the final clusters are concentrated (l), initial number of clusters (k_0), and factor for the cluster number reduction in each iteration of the algorithm (a and $a < 1$). The field of microalgal biofuels comprises different underlying knowledge streams and technologies, revealing that microalgal biofuels are involved in several scientific and technological fields; thus, this study specifies the final number of clusters k ranging from 5 to 10 for each corpus in question when implementing ORCLUS. Aggarwal and Yu (2000) experiments determined that ORCLUS performs well when the specified dimensionalities are between 2 and 8, provided the synthetic data is generated from five clusters. The optimal value of dimensionality l is 6. This study widens the range and tries dimensionalities from 2 to 12. k_0 is chosen to be large (here, the greatest value that computers can handle) to enable the iterations to begin with a larger number of seeds, and therefore improving the likelihood that each cluster will be covered by at least one seed (Aggarwal and Yu, 2000). For sensitivity analysis, several values of k_0 near the largest are additionally specified. We set a as 0.75, which slowly reduces the number of clusters in the merging iterations.

A broad range of parameter settings for paper and patent corpuses separately are evaluated and appropriate settings both by referring the cluster sparsity coefficients provided by ORCLUS and by inviting experts to assess the performance of clustering results are determined. Based on the best clustering results, the papers in a cluster are a scientific field with a similar concept, and the patents in a cluster are a technological field with a similar concept.

This study uses *tm*, *Snowball*, and *ORCLUS* packages in R language for preprocessing, documental representation, and ORCLUS algorithm implementation.

3.2. Stage 4. Identification of technological opportunities

Based on mined scientific fields and technological fields, this study identifies fields with scientific activities but no technological applications, and those with technological applications but no scientific activities. The former fields provide the potential for new technological opportunities because advanced scientific knowledge is the strongest driver of new technological opportunities. The latter fields can also enrich technological opportunities because industrial applications suggest the need for the creation of new knowledge.

4. Results

4.1. Step 1: Document collection

Using the search strategy described in Section 3, this study obtained 6332 paper documents and 1358 patent documents. Paper documents without a digital object identifier (DOI) always returned an empty abstract, and most without publication year or author address. Therefore, 1030 paper documents without a DOI were excluded. Thirty-four paper documents with empty abstracts were also removed. Documents retained included "article," "book chapter," or "proceedings paper," documents removed included "biographical-item," "review," "note," or "editorial material." The final number of retained documents was 4680.

This study obtains 1358 patents based on the search strategy, and then refers to the IPCs designated to each patent to remove those unrelated to microalgal biofuels. According to the World Intellectual Property Organization (WIPO), classifications of A01N "pertain to preservation of bodies of humans or animals or plants or parts thereof." The algae-related patents with A01N typically treat algae as organisms that are threats to human resources, e.g., food; thus, algae-related patents often involve methods or chemical processes to prevent algal growth. After removing patents with A01N as the primary IPC, the retained number of patents is 1173.

As Fig. 4 illustrates, the annual number of published papers on microalgal biofuels slowly grew from 1990 to 2001 and notably increased after 2006. The slope of the annual number of granted patents is relatively flat, revealing that technological development and commercialization of microalgal biofuels is less aggressive than related scientific research.

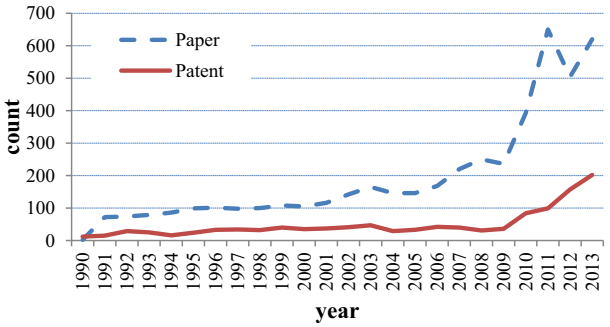

Fig. 4. Annual numbers of collected papers and patents.

This study calculates the cumulative frequencies by country of reprint author of papers and first applicants of patents, and then identifies the top 10 countries by cumulative frequency. As Fig. 5 demonstrates, the U.S. produced the highest number of both papers and patents, outperforming other nations in both scientific research and technological development for microalgal biofuels. Germany ranked third in papers and second in patents, and Japan ranked fourth in papers and third in patents, indicating that both countries dedicate resources toward researching and developing of microalgal biofuels. Although China ranked second in papers published, the number of patents granted is not ranked in the top ten, revealing that a gap exists between scientific research and technological development in China.

4.2. Stage 2: text mining

After preprocessing text, such as unifying synonyms, removing stopwords, and stemming, 20,143 and 7511 words from the paper and patent corpuses, respectively, were obtained. To reduce noise, terms with high zero-entries (high sparsity) in the derived document-by-term matrix were excluded. For the paper corpus, the sparsity is set at 95%, meaning that terms with more than 95% zero-entry in the derived document-by-term matrix are excluded. A 95% sparsity-level setting retained 256 terms. Patent corpus sparsity is set at 96%. The settings produce document-by-term matrices of 4680 × 256, and a matrix of 1173 × 216 for paper and patent corpuses. For each corpus, the matrix cells record the term weights for the corresponding documents, applying TF-IDF weighting measures.

4.3. Stage 3: clustering for scientific and technological fields

On the basis of the derived document-by-term matrices, the ORCLUS algorithm with the broad range of parameter settings separately introduced in Section 3 for paper and patent corpuses is executed; hence, clustering the documents.

The cluster sparsity coefficients provided by ORCLUS to preliminarily screen out appropriate parameter settings are used. In Aggarwal and Yu's study (Aggarwal and Yu, 2000), the smallest sparsity coefficient is 0.002. Here, given that the specified final clusters k range between 5 and 10, the specified subspace dimensionalities l are not less than $(k - 2)$, and the specified initial number of clusters k_0 are, or approach the greatest value that computers can handle, most sparsity coefficients are below 0.001. When implementing clustering, Aggarwal and Yu (2000) suggested defining a minimum threshold for sparsity coefficients and choosing the largest value of l, where the cluster sparsity coefficient is less than the threshold. Accordingly, this study sets a low threshold of 0.001, and parameter settings whose cluster sparsity coefficients are below the threshold. However, this study finds that a trade-off between values of l and k_0 occurs frequently under the limitation of computer memory, meaning that to increase one

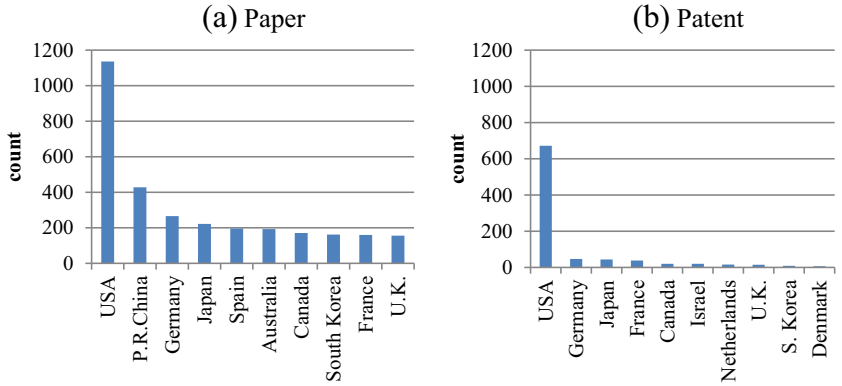

Fig. 5. Cumulative frequencies of papers and patents, by country.

the other must be reduced to derive a converged clustering solution. As a result, the highest specified value of l as 12 to ensure k_0 at a high level is maintained.

Next, this study outputs the cluster centroids and cluster members for each parameter setting candidate. Clusters are identified after identifying terms with high centroids and reading sample articles. Invited experts then read sample papers and patents for assessment. Considering the greatest value of k_0 the computer can handle in this study may not be large enough, which was approximately 35 and 45 for the paper corpus and patent, respectively; ORCLUS was further implemented three times for each parameter setting whose result is identified by experts as being relatively realistic. The iterations introduce stability of the recommended parameter setting, even though the low initial seeds are not very large. Stable clustering may imply that a large number of initial seeds were used. After several runs, the most realistic and stable clustering results corresponded to a parameter setting of cluster numbers k being 9 and subspace dimensionalities l being 11 (paper corpus), and a setting of k being 7 and l being 10 (patent corpus). The corresponding cluster sparsity coefficients are 0.0006 and 0.0002. The generated cluster sizes for scientific fields are 962, 1442, 245, 258, 248, 235, 274, 753, and 263; the sizes for technological fields are 72, 121, 120, 161, 129, 192, and 378. In successive analyses, nine scientific fields and seven technological fields are denoted as SF1 to SF9 and TF1 to TF7, respectively.

Terms with high centroids to interpret the characteristics of the scientific and technological fields and lists representative terms in Tables 2 and 3 were identified. In addition, research categories of papers and the IPCs in patent documents to preliminarily identify these field characteristics were referred to. The following two subsections describe the characteristics of nine scientific and seven technological fields.

4.3.1. Characteristics of scientific fields

According to the research areas defined by ISI, SF1 papers are in the categories of ecology, marine and freshwater biology, and fisheries. According to SF1 terms, this field relates to interactions between algae and ecosystems, such as the role of algae in the food web of certain fish species or regions, effect of environmental factors on algae communities, and algal growth in different conditions. Representative terms include food, diet, fish, feed, ecosystem, community, benthic, and species. Therefore, SF1 is labeled as "algal ecosystem."

Most SF2 papers are in plant sciences, biochemistry and molecular biology, and biotechnology and applied microbiology, and together comprise the largest scientific field paper count. Terms with high centroid value include light, fluorescence, photosystem, electron, cell, excitation, and absorption. These terms are related to electron transporter chains in photosynthetic reactions; thus, SF2 includes the terms chlorophyll, complex, pigment, and structure, which reveals that SF2 papers investigate algal structures and proteins, photosynthetic processes, and light utilization in photosystems. Microalgae are photosynthetic organisms; therefore, their photosynthetic efficiency is a crucial determinate in their productivity. Therefore, SF2 is labeled as "photosynthesis and light utilization."

The main categories of SF3 are biotechnology and applied microbiology, and chemistry. The term hydrogen reveals that some SF3 papers investigate bio-hydrogen production through light-driven pathways, although more terms are associated with gene expression, enzymes, and metabolic pathways. Terms such as gene, express, ferment, protein, metabolic, and pathway are found, particularly in studies that investigate the genetic and cellular processes involved in synthesis and regulation in algal strains, and hence identify the most useful. Therefore, SF3 is labeled as "strain screening by genetic approaches."

Table 2
Scientific fields in paper documents.

SF	Count	Mean year	Main categories	Representative terms (in decreasing order by centroid)	Field naming
1	962	2004.4	Ecology; Marine & Freshwater Biology	Food, diet, fish, feed, ecosystem, community, sea, benthic, species, nutrient, water, organism, carbon, consumption, phytoplankton, source, fed, marine, dynamic, ecology	Algal ecosystem
2	1442	2003.5	Plant Sciences	Light, fluorescence, photosystem, chlorophyll, complex, photosynthesis, electron, protein, cell, excitation, pigment, model, structure, absorption, temperature, react, grow	Photosynthesis and light utilization
3	245	2008.1	Biotechnology & Applied Microbiology	Hydrogen, gene, express, ferment, protein, metabolic, pathway, photosynthesis, light, green algae, strain, cell, enzyme, plant, evolution, biological, condition, molecular	Strain screening by genetic approaches
4	258	2008.3	Biotechnology & Applied Microbiology	Extract, photobioreactor, temperature, lipid, biomass, reactor, light, culture, gas, degree, cultivation, grow, bioreactor cell, solar, photosynthesis, carbon dioxide, oxygen	Microalgal cultivation and growth conditions
5	248	2006.4	Biotechnology & Applied Microbiology	Cell, culture, grow, bioreactor, medium, concentration, batch, cultivation, biomass, strain, ferment, carbohydrate, uptake, carbon, nutrient	Heterotrophic or mixotrophic cultivation
6	235	2008.1	Biotechnology & Applied Microbiology	Model, membrane, kinetic, remove, solution, concentration, enzyme, time, metabolic, chemical, biomass, organism, compound, treatment, separate, system, biological, material, parameter	Kinetic modeling and chemical parameter investigation
7	274	2009.2	Agricultural Engineering; Environmental Engineering	Wastewater, digest, remove, anaerobic, treatment, nutrient, diet, biomass, phosphorus, lipid, grow, day, nitrogen, cultivation, rate, waste, biofuel, fed, chlorella, biodiesel, feed, system, culture	Integration of algae cultivation with wastewater treatment
8	753	2010.8	Agricultural Engineering; Biotechnology & Applied Microbiology	Lipid, biodiesel, biofuel, oil, biomass, fuel, cultivation, acid, fatty, feedstock, content, process, nitrogen, grow, extract, culture, strain, economic, cost, optimized, yield, water, system, harvest, life	Synthesis, harvesting, extraction, and conversion of lipids
9	263	2010.2	Biotechnology & Applied Microbiology	Oil, acid, fatty, biodiesel, lipid, composition, fuel, yield, biomass, react, feedstock, properties, biofuel, strain, culture, gas, chain, conversion	Conversion of algal biomass to fuels

Table 3
Technological fields in patent documents.

TF	Count	Mean year	Main IPC	Representative terms (in decreasing order by centroid)	Field naming
1	72	2007.9	C12N	Sequence, nucleic, acid, encode, molecule, gene, isolate, enzyme, express, fatty, cell, group, protein, organism, oil	Genetic engineering on algae
2	121	2004.5	C02F	Fraction, water, alkyl, biomass, liquid, algae, treatment, solid, ester, waste, carbon, lipid, separate, surface, treat, solvent, material, oxygen, protein, remove, organism, system, nutrient	Recycling CO2 and wastes by algae
3	120	2006.3	B01D, C12M	Tank, gas, algae, water, light, wet, culture, liquid, stream, flow, solvent, biomass, system, wall, cell, lipid, separate, fraction, harvest, remove, oxygen, photosynthesis, dioxide, cellulose, carbon, reactor, filter, dried, filtrate	Culturing, harvesting, and dewatering technologies
4	161	2006.3	B01D, C02F	oil, filter, stream, extract, algae, water, system, liquid, separate, solvent, fraction, flow, biomass, species, chamber, solid, gas, protein, component, remove, lipid, grow, process, fluid	Dewatering and extraction technologies
5	129	2008.0	C12P, A61K	Biomass, fatty, acid, omega, oil, fuel, material, extract, lipid, animal, plant, chain, transport, feedstock, ferment, ethanol, harvest, microbial, sugar, heat, wet, conversion	Conversion of algal extracts to ethanol and nutritional products
6	192	2006.4	C07C	Cell, fluid, react, ester, feedstock, hydrocarbon, hydrogen, fatty, catalyst, alcohol, fuel, production, triglyceride, process, diesel, pressure, temperature, biodiesel, alkyl	Biofuel conversion technologies
7	378	2004.5	A61K, C02F, C12P	Composition, organism, material, compound, acid, medium, extract, mix, polymer, process, product, microorganism, grow, metal, substance, oxidant, method, component, bacteria, marine	Applications and potential co-products

The main categories of SF4 to SF6 are biotechnology and applied microbiology. Both SF4 and SF5 include the terms culture, grow, cultivation, and biomass, revealing that two SFs relate to microalgal cultivation. The terms photobioreactor, temperature, photosynthesis, solar, carbon dioxide, and oxygen are those that SF4 contains, but not SF5. These terms relate to environmental conditions for cultivating microalgae. Sunlight, gas exchange (both carbon dioxide and oxygen), and appropriate temperature are required for microalgal growth; therefore, SF4 is labeled as "microalgal cultivation and growth conditions."

Those terms with the highest centroids in SF5 are: cell, culture, and grow. Terms of medium, batch, strain, ferment, carbohydrate, uptake, nutrient, and nitrogen are those that SF5 contains, but not SF4. These terms reveal that SF5 includes the study of cellular growth mechanisms in a growth medium. SF5 also includes studies where microalgae are fed carbon sources, such as carbohydrate or nitrogen, to generate highly concentrated biomass, which indicates heterotrophic, or mixotrophic cultivation. This approach utilizes mature industrial fermentation technology. Therefore, SF5 is labeled as "heterotrophic or mixotrophic cultivation."

SF6 terms show specializations in modeling-related research, such as kinetic modeling of heavy metal absorption by microalgae in the aqueous phase. This field also investigates chemical parameters in the separation and harvesting of algae.

The main categories of SF7 are agricultural engineering and environmental engineering. The representative terms include wastewater, remove, anaerobic, digest, treatment, nutrient, diet, biomass, phosphorus, nitrogen, cultivation, and waste. SF7 focuses on integrating algae production and cultivation with wastewater treatment; that is, nutrient-rich wastewater can be used simultaneously for algae production and wastewater treatment.

SF8 terms indicate a focus on synthesis, harvesting, extraction, and applications of algal oils and lipids (including fatty acids and triglycerides), which are used in biodiesel production. Terms such as lipid, oil, biomass, cultivation, acid, fatty, content, nitrogen, and grow present that SF8 includes studies that investigate lipid accumulation and oil content of algae. Algal cultures are mainly grown in aqueous conditions; hence, harvesting and dewatering are required processes for converting algae into liquid fuels. The terms optimized, yield, water, harvest, and life indicate that SF8 projects include these steps. Furthermore, the terms lipid, biodiesel, economic, cost, extract, and system present that SF8 projects involve cost-efficient system design for extracting lipids from algal biomass and applying the extracted lipids to biodiesel production. Therefore, SF8 is labeled as "synthesis, harvesting, extraction, and conversion of lipids."

SF9 relates converting algal biomass to fuels, emphasizing the conversion of algal oil extracts to biodiesel. The conversion processes from algal biomass to final fuel specifications (e.g., biodiesel and biogas) are highly interdependent, so SF9 also covers studies on feedstock characterizations, such as lipid and fatty-acid compositions and thermochemical processes.

Based on the mean published year of papers in each field listed Table 2, papers in SF2 were published earliest, with a mean published year of mid-2003, demonstrating early exploration of microalgal photosynthesis and light utilization processes. The field of algal ecosystem (SF1) also attracted earlier investigations, with a mean published year of mid-2004. The means of SF8 and SF9 are 2011 and 2010, respectively, revealing the recent expansion of science and technology in the areas of synthesis, harvesting, extraction, and conversion of biomass to fuels.

4.3.2. Characteristics of technological fields

Table 3 shows the main IPCs, representative terms, patent counts, and naming for each technological field. The main IPC of TF1 is C12N. WIPO defines C12N as "microorganisms, enzymes, genetic engineering, and culture media." TF1 terms, such as sequence, nucleic, acid, encode, isolate, gene, control, carrier, enzyme, and express, demonstrate that TF1 relates to algal genetic engineering, which modifies algal strains and enables the selection of strains possessing desirable properties for biofuel production. Therefore, TF1 is labeled as "genetic engineering on algae."

The main IPC of TF2 is C02F. By definition, C02F involves water and sewage treatment. In Table 3, terms such as fraction, water, treatment, waste, carbon, remove, organism, and nutrient indicate that TF2 covers processes or technologies using nutrients (e.g., nitrogen and carbon dioxide) in wastewater or other waste material for the growth of algal biomass.

The main IPCs of TF3 are B01D and C12M, where B01D relates to separation technologies and methods, and C12M relates to apparatuses for enzymology or microbiology. Terms such as gas, water, light, culture, biomass, photosynthesis, and carbon dioxide are associated with algal cultivation. Specifically, the terms tank and reactor indicate the cultivation of algae in bio-reactors. Terms such as water, cell, wall, lipid, separate, fraction, harvest, collect, and dried reveal that TF3 further covers technologies, apparatuses or systems of removing water from algal cells. Therefore, the label of TF3 is "culturing, harvesting and dewatering technologies."

In TF4, the terms oil, extract, water, protein, component, and lipid indicate that TF4 focuses on the steam extraction of lipids and proteins from algal oil or biomass. Terms such as water, separate, biomass, and remove reveal that TF4 comprises dewatering steps prior to the extraction process. Therefore, TF4 is labeled "dewatering and extraction technologies."

TF5 involves extracting fatty acids from algal biomass, which can then be converted into biodiesel and biofuels, long chain polyunsaturated fatty acids (e.g., omega-3 supplements), and biomass residues for animal feed and plant fertilizers. The terms biomass, fatty, acid, omega, oil, extract, lipid, chain, and conversion present the technological content. TF5 also includes the enzymatic extraction of ethanol; the terms ferment, ethanol, and sugar present the technological content. Therefore, TF5 is labeled "conversion of algal extracts to ethanol and nutritional products."

The main IPC of TF6 is C07C, which means acyclic or carbocyclic compounds. By definition, C07C contains several groups, with most being the preparation of hydrocarbons, where the "preparation" means "purification, separation, stabilization or use of additives." Together with terms such as ester, hydrocarbon, hydrogen, alcohol, fuel, triglyceride, diesel, and biodiesel, TF6 is found related to technologies that convert algal biomass into viable fuels, including hydrogen, alcohols, hydrocarbons, and biodiesel.

TF7 has the highest patent count among the seven TFs and covers several IPCs, primary among them A61K, C02F, and C12P. These three IPCs are for ablution systems or medical purposes, waste water or sewage treatment, and fermentation or enzymatic synthesis. Most terms in TF7 relate to the chemical compositions of algal extracts, such as composition, compound, acid, extract, polymer, and oxidant. A61K, the first part of chemical composition in TF7, relates to algal extracts for medical and cosmetic applications that are potential co-products of biofuel production. Based on C02F, the second part of chemical composition in TF7 is associated with water treatment for removing harmful substances by algae. The terms material, carbon, salt, and metal represent the technological content. Therefore, TF7 is labeled "applications and potential co-products."

As shown in Table 3, the seven patent fields, in chronological order by mean granted years, are TF2, TF7, TF3, TF4, TF6, TF1, and TF5. Early technology researchers engaged in using microalgae to recycle carbon dioxide and wastes (TF2) and extracting high-value products from microalgae (TF7). Conversion of microalgal extracts to ethanol and nutritional products (TF5) is the most recent field.

4.4. Identification of technological opportunities

A feasible production chain of microalgal biofuels, shown in Fig. 6, has been described. The chain begins with the microalgae ecosystem and biology. Microalgae comprise multiple strains, with diverse characteristics, and can be found in a variety of habitats, both aqueous and terrestrial. Knowledge of these ecosystems and of microalgae biology (e.g., photosynthetic processes and capabilities, light utilization, and growth conditions) enables researchers to identify and select potential strains for biofuel feedstocks. Our analytical results show that SF1 and SF2 focus on this knowledge discovery while SF3 and TF1 address approaches to improve strains through genetic engineering approaches. Then, the feedstocks of the selected microalgae strains require processing of cultivation, harvesting, dewatering, and biomass extraction to obtain fuel precursors such as lipids and carbohydrates. The corresponding fields include SF4, SF5, SF6, SF8, TF3, and TF4. After processing, the extracted biomass requires conversion technologies to convert biomass to biofuels (e.g., biodiesel and renewable hydrocarbon) and useful co-products (e.g., nutritional products or feed). SF9, TF5, TF6, and TF7 study the conversion technologies and end-products. After human uses, the electricity and heat energy yielded by biofuels, the wastewater, emitted carbon dioxide, and residuals can be recycled to further cultivate microalgae; furthermore, SF6, SF7, and TF2 investigate the integration of microalgae cultivation with wastewater treatment and recycling.

To compare annual scales of different stages in the chain, Fig. 7 displays paper and patent sizes for each field using two bubble charts, where the horizontal axis, vertical axis, and bubble area present year, field (nominal scale), and document count, respectively. The fields in the vertical axis correspond with the production chain of Fig. 6. As shown in Figs. 6 and 7, the upstream fields in the chain, algal ecosystem (SF1) and microalgal photosynthesis and light utilization (SF2), are fields where scientists focus attention, but lack technological literature. One possible reason is that these fields investigate natural principles, such as natural laws, natural phenomena, or naturally occurring relations, which are not patentable (USPTO, 2013). However, discoveries of promising microalgae strains and corresponding metabolic pathways and growth physiologies, undoubtedly benefit downstream biofuel production. Moreover, the large bubble areas in Fig. 7 show that the paper counts for the two fields are considerably larger relative to other fields, implying an abundance of scientific output for technological researchers and engineers to potentially apply.

Science and technology interact and co-evolve from the midstream to downstream fields in the production chain, which includes fields from genetic engineering applications to biofuel generation (SF3 to SF9, TF1 to TF6). The field of "synthesis, harvesting, extraction, and conversion of lipids (SF8)" has great and increasing bubbles by size from 2009, demonstrating recent attention and productive efforts. The corresponding technological field is TF4, with a relatively small bubble size. Considering the difference in sizes between

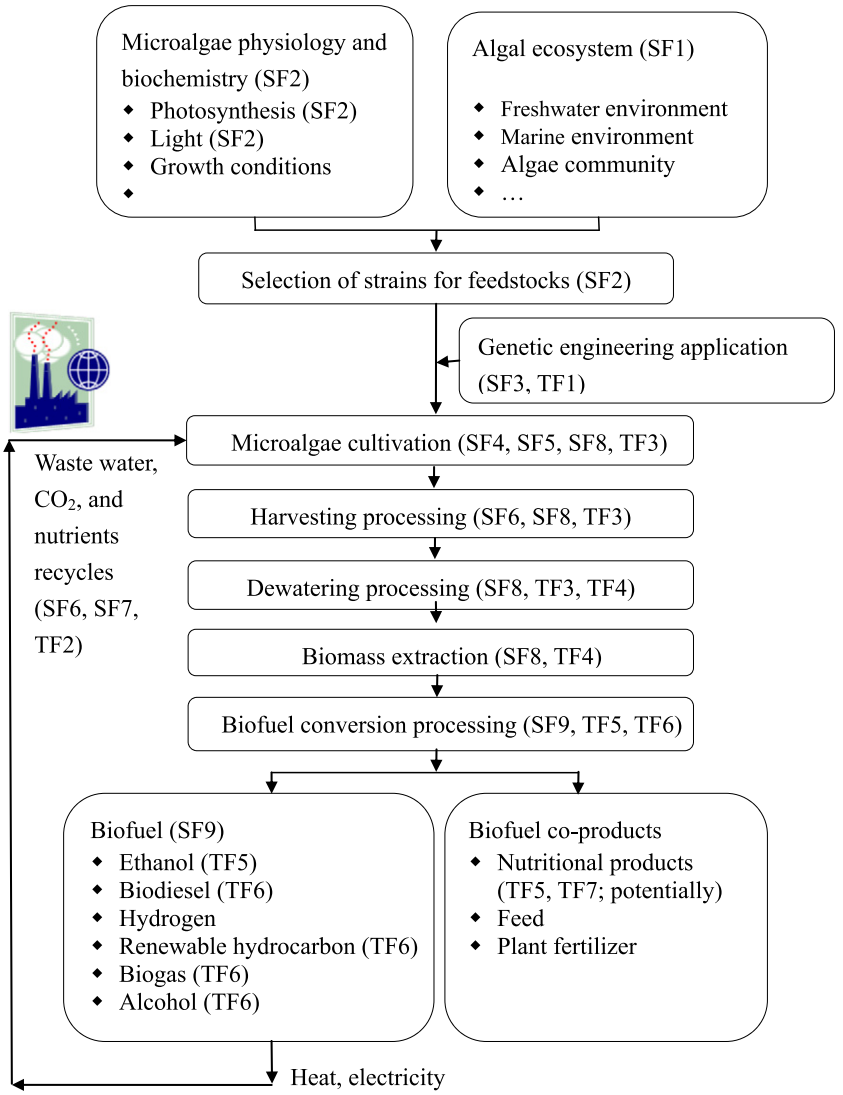

Fig. 6. Concordance of scientific and technological fields. Note: The SF and TF in parentheses are abbreviations for scientific and technological fields, respectively.

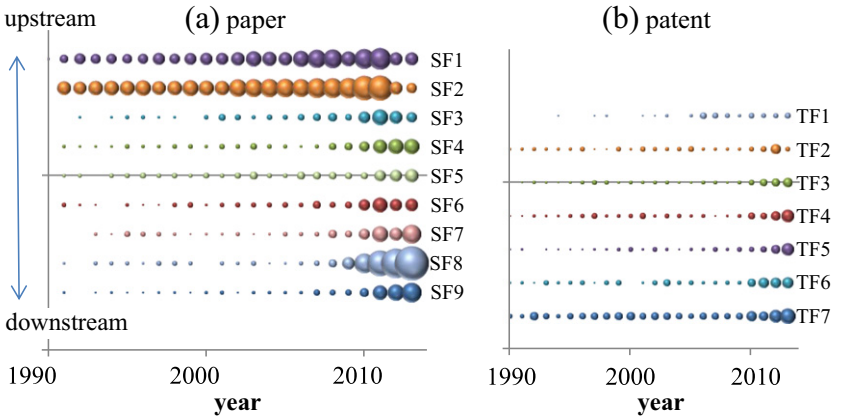

Fig. 7. Annual cluster sizes for each scientific and technological field. Note: Bubble area represents cluster size.

scientific and technological fields, the scientific field may have developed an abundance of basic knowledge for technological applications, thus presenting further opportunities for applications in the fields of synthesis, harvesting, extraction, and conversion of lipids.

In the downstream fields, two non-fuel end-products of microalgal conversion (TF5 and TF7), which are related to applications and potential co-products of biofuel productions, respectively, are discovered. The terms and IPCs of TF7 indicate that the majority of documents in the field are not directly related to biofuel productions; rather, they focus on developing non-fuel end-products such as chemical compositions for health food supplements. As described in Section 3, a broader query to retrieve technological literature than for scientific literature is adopted because patent applicants typically submit genus claims. As a result, fields relating to non-fuel end-products may only exist in the technological, and not scientific, literature. Based on analytical results from the scientific literature, which include purely microalgal biofuel studies, an obvious cluster that contains both the terms "biofuel" and "non-biofuel end-products" was not detected, which implies that studies that investigate the production of high-value non-fuel products, and are associated with microalgal biofuel, are not yet advanced enough to form a distinctive scientific field. Using every component of microalgal biomass is an effective approach to improving the economics of microalgal biofuel production (U.S. DOE, 2010); therefore, more scientific exploration is required to achieve new breakthroughs.

Our results are supported by the 2010 National Algal Biofuels Technology Roadmap, published by the U.S. DOE (2010). A major objective of the roadmap is to understand the status of algal biofuels in research, development, and deployment (RD&D) activities, and presents information that supports and guides RD&D investment. The roadmap summarizes three aspects of content for microalgal biofuel: technology, economy, and policy. The scientific and technological fields identified are consistent with the contents of technology described in the roadmap. The evolutionary history of algal biofuels introduced in the roadmap confirms the mean published year sequence derived for the scientific fields. Additionally, a report by International Energy Agency (IEA) stated that, "The production of liquid transportation fuels from algal biomass is technically feasible. However, there is a need for innovation in all elements of algal biofuels production to address technical inefficiencies" (Darzins et al., 2010). The technological fields identified contain activities at the upstream and downstream ends of the production chain, which could be reflected in the statement of "technically feasible" in the IEA report.

For each field, papers by the first author's affiliation and patents by the first applicant's are organized. The affiliations with largest paper counts from SF1 to SF9, respectively, are: Chinese Academy of Science (n = 15), Chinese Academy of Science (n = 29), Chinese Academy of Science (n = 12), University of Almeria (n = 8), Oregon State University (n = 7), Monash University (n = 7), University of Minnesota (n = 8), Chinese Academy of Science (n = 29), and University of Minnesota (n = 7). The applicants with the largest patent counts from TF1 to TF6, respectively, are: DuPont Company (n = 33), Heliae Development LLC (n = 13), Heliae Development LLC (n = 15), Heliae Development LLC (n = 12), Xyleco Inc. (n = 16), UOP LLC (n = 14), and Martek Biosciences Corporation (n = 9).

5. Discussion and conclusions

Text mining and the ORCLUS algorithm, which is capable of clustering highly dimensional data, to identify technological opportunities for microalgal biofuels is used. In conclusion, we acknowledge the limitations of this study and suggest future research directions.

The first aspect relates to a methodological extension. Shibata et al. (2010) propose extracting potential technological opportunities by comparing fields in science and technology literature. To identify scientific and technological fields, they adopt a citation-based clustering approach and analyze the largest components of the literature citation networks. This study notes that certain emerging technologies may be underdeveloped, with an insufficient basic knowledge base and citation network. Taking microalgae as an example, Fig. 8(a) illustrates the total annual number of papers and the annual number of papers in the largest component of the citation network. Fig. 8(b) displays the same data for patents. By citation-based clustering approach, totally 27% and 59% of papers and patents should be removed because the approach considers only the largest component. Therefore, this study uses a text-based clustering approach to identify fields in the entire

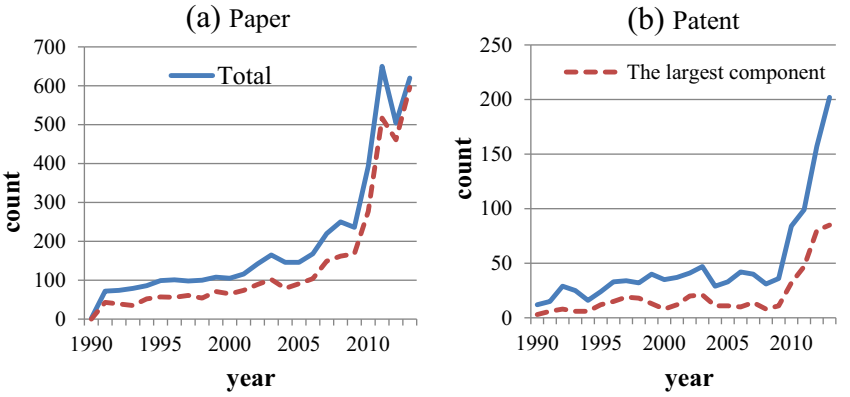

Fig. 8. Annual numbers of papers and patents.

literature domain. Moreover, up to two years could pass before a paper/patent receives citations from subsequent papers/patents (Shibata et al., 2010). As a result, the text-based clustering approach might have higher chance to identify important emerging fields than the citation-based one.

Compared with the citation-based clustering approach, our text-based clustering approach has the disadvantage that the validity of clustering results is severely affected by the preciseness of document wordings. In our experience of clustering scientific papers, the clustering results are quite consistent under different parameter settings, and cluster meanings are easy to identify because the wordings in papers are generally precise. In contrast, patent applicants typically use genus terms to enlarge their claims, and these wordings increase the challenges to our text-based clustering approaches. This study can detect meaningful clusters, under expert support, and referring to the IPC's definitions. Future research can develop approaches to integrating IPCs into text-based clustering approaches to enhance the clustering performance.

Previous studies analyzed the performance of text-based approaches, and some studies concluded that text-based approaches perform worse than citation-based ones (e.g. Shibata et al., 2011a). Regarding the results, our experience is that, if the number of extracted features is not large enough, then text-based approaches perform poorly because the matrices representing the corpuses prune excessive information. If the number of extracted features is large enough, to the best of our knowledge, none of the previous studies examined the nature of these data, which exist in high-dimensional spaces, with high sparsity. In these spaces, conventional similarity measures become invalid. This study reminds the necessity to adopt appropriate techniques for handling high-dimensional data which are generated through standard text mining processes.

As introduced in Section 2, high-dimensional data contains three classes of clustering: subspace, generalized subspace, and pattern-based. Previous studies have proposed several algorithms for each class, and each is characterized by its underlying cluster model and parameterization of the resulting cluster. It is important to first know which algorithm would be suitable for what kind of problems. So far, to the best of our knowledge, there is no empirical evaluation with respect to the effectiveness and efficiency for the algorithms in the second and third classes. Three papers have systematically evaluated state-of-the-art algorithms in the first class (Müller et al., 2009; Moise et al., 2009; Günnemann et al., 2011). Based on the new evaluation measures proposed by Müller et al. (2009) and Günnemann et al. (2011), PROCLUS outperforms others in several measures. When PROCLUS is applied to our empirical data, very unbalanced dimensional numbers of subspaces are obtained, meaning that most clusters exist in subspaces with only two to three dimensions, and one cluster occupies a space with high dimensionality. The results are unrealistic to the case of microalgal biofuels; thus, from the class of generalized subspace clustering algorithms, this study selects ORCLUS, an extended version of the algorithm PROCLUS, and obtains acceptable results. This study has the drawback of not systematically evaluating all algorithms, which requires further research to remedy. Once the appropriate algorithm has been identified, future research can further compare the performances of text- and citation-based clustering approaches.

In this study, experts identify the clustering parameters and relationships between TFs and SFs. This process demands considerable time and effort, and future research can improve the process by referring to Shibata et al. (2011b), which calculates the similarity between paper and patent clusters. The resulting similarity can serve as reference for identifying relationships between TFs and SFs.

The second aspect of the conclusion relates to technological opportunities that this study extracts. On the production chain of microalgae biofuels, our results demonstrate that many scientists address research questions at the upstream side in two fields: algal ecosystems, and microalgal photosynthesis and light utilization. However, no technological literature published in these fields is apparent. The paper counts of the two fields are significantly large, so technological developers can advance by referring to the abundant research outcomes of the two scientific fields.

Science and technology interact and coevolve from the midstream to downstream fields of the production chain, which includes genetic engineering applications, microalgal cultivation, harvesting and dewatering processes, biomass extraction, biofuel conversion processes, and end-products of biofuel production. Among the fields where science and technology coevolve, the field of synthesis, harvesting, extraction, and conversion of lipids attract considerable scientific attention, and the number of papers has notably grown from 2009. The corresponding technological field has few patents and does not have an apparent growth trend; thus, technological opportunities potentially exist in this field.

In the second aspect, this study is limited by the scope of the document databases used for data collection. Papers indexed by the SCIE database and technological patents granted by the USPTO, to offer suggestions about technological opportunities for microalgal biofuel are collected. Future research can broaden the scope of document databases employed. In scientific databases, future research could also include Scopus, which is a well-structured citation database. In technological databases, future research could include the USPTO patent-application database to obtain relatively current technology information, and also include patents published by other patent offices, to enrich the sources of technology information.

This study shares some limitations with Shibata et al. (2010). The first is time lag. It takes several months from submitting scientific findings to the publications on academic journals and some years from filing technological innovations to being granted by patent offices. Consulting expert insight is a feasible means to capture the possible contents of papers or patents in examinations. Second, the analytical results should serve as an intellectual basis for constructing R&D strategies, rather than being strategies themselves. Third, not all academic knowledge is represented by patents, which hinders exploration on technological opportunities. Some technological developers protect their technological inventions by means of trade secret, rather than applying for patents; others may not apply for patents if they evaluate the technologies as lacking market potential or requiring more than 20 years' lead time.

Acknowledgment

The authors thank two anonymous referees for their helpful comments on this article and thank the Biotechnology Business

Section, Solvent and Chemical Business Division, CPC Corporation, Republic of China (Taiwan) for providing technology consultancy in biofuels. This research was supported by the Ministry of Science and Technology of the Republic of China (Taiwan). Grant number is NSC 102-2410-H-415-011.

References

Aggarwal, C.C., Yu, P.S., 2000. Finding generalized projected clusters in high dimensional spaces. Proceedings of the 2000 ACM SIGMOD international conference on Management of dataACM, Dallas, Texas, USA, pp. 70–81.
Beyer, K., Goldstein, J., Ramakrishnan, R., Shaft, U., 1999. When is "nearest neighbor" meaningful? Database Theory—ICDT'99, Springer, Proceedings of the 7th International Conference on Databse Theory (ICDT), Jerusalem, Israel, pp. 217–235.
Cohen, W.M., Levinthal, D.A., 1989. Innovation and learning: the two faces of R&D. Econ. J. 99, 569–596.
Cohen, W.M., Levin, R.C., Mowery, D.C., 1987. Firm size and R&D intensity: a reexamination. J. Ind. Econ. 35, 543–565.
Darzins, A., Pienkos, P., Edye, L., 2010. Current status and potential for algal biofuels production. IEA Bioenergy Task 39—commercializing liquid biofuels. The International Energy Agency, p. V.
Davis, R.,Aden, A.,Pienkos, P.T., 2011. Techno-economic analysis of autotrophic microalgae for fuel production. Appl. Energy 88, 3524–3531.
Feldman, R.,Sanger, J., 2007. The Text Mining Handbook: Advanced Approaches in Analyzing Unstructured Data. Cambridge University Press, Cambridge.
Frenz, M., Prevezer, M., 2012. What can CIS data tell us about technological regimes and persistence of innovation? Ind. Innov. 19, 285–306.
Glänzel, W., Meyer, M., 2003. Patents cited in the scientific literature: an exploratory study of 'reverse' citation relations. Scientometrics 58, 415–428.
Günnemann, S.,Färber, I.,Müller, E.,Assent, I.,Seidl, T., 2011. External evaluation measures for subspace clustering. Proceedings of the 20th ACM international conference on Information and knowledge management. ACM, Glasgow, Scotland, UK, pp. 1363–1372.
Hellmann, T., 2007. The role of patents for bridging the science to market gap. J. Econ. Behav. Organ. 63, 624–647.
IPCC, 1992. Climate Change: The IPCC 1990 and 1992 Assessments. Intergovernmental Panel on Climate Change.
Klevorick, A.K., Levin, R.C., Nelson, R.R., Winter, S.G., 1995. On the sources and significance of interindustry differences in technological opportunities. Res. Policy 24, 185–205.
Konur, O., 2011. The scientometric evaluation of the research on the algae and bio-energy. Appl. Energy 88, 3532–3540.
Kostoff, R.N., 2006. Systematic acceleration of radical discovery and innovation in science and technology. Technol. Forecast. Soc. Chang. 73, 923–936.
Kostoff, R.N., 2008. Literature-Related Discovery (LRD): introduction and background. Technol. Forecast. Soc. Chang. 75, 165–185.
Kriegel, H.-P., Kröger, P., Zimek, A., 2009. Clustering high-dimensional data: a survey on subspace clustering, pattern-based clustering, and correlation clustering. ACM Trans. Knowl. Disc. Data 3, 1–58.
Lee, S., Yoon, B., Lee, C., Park, J., 2009. Business planning based on technological capabilities: patent analysis for technology-driven roadmapping. Technol. Forecast. Soc. Chang. 76, 769–786.
J.A. Lefstin, The formal structure of patent law and the limits of enablement, in, University of California, Hastings College of Law, 2008
Martino, J.P., 2003. A review of selected recent advances in technological forecasting. Technol. Forecast. Soc. Chang. 70, 719–733.
Meyer, M., 2000. Patent citations in a novel field of technology—what can they tell about interactions between emerging communities of science and technology? Scientometrics 48, 151–178.
Meyer, M., 2002. Tracing knowledge flows in innovation systems—an informetric perspective on future research science-based innovation. Econ. Syst. Res. 14, 323–344.
Moise, G., Zimek, A., Kröger, P., Kriegel, H.-P., Sander, J., 2009. Subspace and projected clustering: experimental evaluation and analysis. Knowl. Inf. Syst. 21, 299–326.
Müller, E., Günnemann, S., Assent, I., Seidl, T., 2009. Evaluating clustering in subspace projections of high dimensional data. Proc. VLDB Endowment 2, 1270–1281.
Narin, F.,Hamilton, K.S.,Olivastro, D., 1997. The increasing linkage between US technology and public science. Res. Policy 26, 317–330.
Nelson, R.R., 1982. The role of knowledge in R&D efficiency. Q. J. Econ. 97, 453–470.
Nieto, M., Quevedo, P., 2005. Absorptive capacity, technological opportunity, knowledge spillovers, and innovative effort. Technovation 25, 1141–1157.
Olsson, O., 2005. Technological opportunity and growth. J. Econ. Growth 10, 35–57.
Parsons, L., Haque, E., Liu, H., 2004. Subspace clustering for high dimensional data: a review, SIGKDD. Explor. Newsl. 6, 90–105.
Petrescu, A.S., 2009. Science and technology for economic growth. New insights from when the Data Contradicts Desktop Models1. Rev. Pol. Res. 26, 839–880.
Price, D.J.D., 1965. Is technology historically independent of science? A study in statistical historiography. Technol. Cult. 6, 553–568.
Robinson, D.K.R., Huang, L., Guo, Y., Porter, A.L., 2013. Forecasting Innovation Pathways (FIP) for new and emerging science and technologies. Technol. Forecast. Soc. Chang. 80, 267–285.
Rosenberg, N., 1982. How exogenous is science? In: Rosenberg, N. (Ed.), Inside the Black Box: Technology and Economics. Cambridge University, Cambridge, pp. 141–160.
Salton, G., Wong, A., Yang, C.S., 1975. A vector space model for automatic indexing. Commun. ACM 18, 613–620.
Shane, S., 2001. Technological opportunities and new firm creation. Manag. Sci. 47, 205–220.
Shibata, N., Kajikawa, Y., Sakata, I., 2010. Extracting the commercialization gap between science and technology—case study of a solar cell. Technol. Forecast. Soc. Chang. 77, 1147–1155.
Shibata, N., Kajikawa, Y., Sakata, I., 2011a. Measuring relatedness between communities in a citation network. J. Am. Soc. Inf. Sci. Technol. 62, 1360–1369.
Shibata, N., Kajikawa, Y., Sakata, I., 2011b. Detecting potential technological fronts by comparing scientific papers and patents. Foresight 13, 51–60.
Singh, J., Gu, S., 2010. Commercialization potential of microalgae for biofuels production. Renew. Sustain. Energy Rev. 14, 2596–2610.
Smith, V.H., Sturm, B.S.M., de Noyelles, F.J., Billings, S.A., 2010. The ecology of algal biodiesel production. Trends Ecol. Evol. 25, 301–309.
Szepannek, G., 2013. Package ORCLUS. The R Project for Statistical Computing.
U.S. DOE, 2010. National algal biofuels technology roadmap. U.S. Department of Energy, Office of Energy Efficiency and Renewable Energy, Biomass Program.
USPTO, 2106.01 Subject matter eligibility analysis of process claims involving laws of nature, in, The United States Patent and Trademark Office, 2013.
Wang, M.-Y., Chang, D.-S., Kao, C.-H., 2010. Identifying technology trends for R&D planning using TRIZ and text mining. R. Manag. 40, 491–509.
Weiss, S.M.,Indurkhya, N.,Zhang, T.,Damerau, F.J., 2005. Text Mining: Predictive Methods for Analyzing Unstructured Information. Springer Science + Business Media, Inc., New York.
Wu, F.-S., Hsu, C.-C., Lee, P.-C., Su, H.-N., 2011. A systematic approach for integrated trend analysis—the case of etching. Technol. Forecast. Soc. Chang. 78, 386–407.
Yoon, J., Kim, K., 2012. Detecting signals of new technological opportunities using semantic patent analysis and outlier detection. Scientometrics 90, 445–461.
Yoon, B., Park, Y., 2005. A systematic approach for identifying technology opportunities: keyword-based morphology analysis. Technol. Forecast. Soc. Chang. 72, 145–160.
Ziman, J., 1988. An Introduction to Science Studies. Cambridge University Press, Cambridge.

Ming-Yeu Wang is an associate professor of the Department of BioBusiness Management (formerly known as Department of Bio-industry and Agribusiness Administration), National Chiayi University, Taiwan. She received her Master degree in management of technology and the Ph.D. degree in industrial engineering and management from the National Chiao Tung University, Taiwan. Her works have appeared in *Technological Forecasting & Social Change*, *R&D Management*, *International Journal of Technology Management and Journal of Engineering and Technology Management*. Her current research focuses on technological forecasting, patent analysis and R&D management.

Shih-Chieh Fang is a Professor of the Department of Business Administration and Institute of International Business, National Cheng Kung University, Taiwan. He received his Ph.D. degree from Department and Graduate Institute of Business Administration, National Taiwan University, Taiwan. He published his papers in *Industrial Marketing Management*, *Journal of Business and Industrial Marketing and Scientometrics*. He has handled a program of energy policy at Center for Energy Technology and Strategy, National Cheng Kung University, Taiwan.

Yu-Hsuan Chang is a patent engineer in Gainia Intellectual Asset Services, Inc. in Taiwan. She received her master's degree from the Department of Bio-industry and Agribusiness Administration, National Chiayi University, Taiwan.