Scientometrics (2015) 102:811–827
DOI 10.1007/s11192-014-1392-6

Analyzing patent topical information to identify technology pathways and potential opportunities

Jing Ma · Alan L. Porter

Received: 11 March 2014 / Published online: 30 July 2014

Abstract As a basic knowledge resource, patents play an important role in identifying technology development trends and opportunities, especially for emerging technologies. However patent mining is restricted and even incomplete, because of the obscure descriptions provided in patent text. In this paper, we conduct an empirical study to try out alternative methods with Derwent Innovation Index data. Our case study focuses on nano-enabled drug delivery (NEDD) which is a very active emerging biomedical technology, encompassing several distinct technology spaces. We explore different ways to enhance topical intelligence from patent compilations. We further analyze extracted topical terms to identify potential innovation pathways and technology opportunities in NEDD.

Keywords Patent analysis · Text mining · Tech mining · Technology pathways · Technology opportunities analysis · Nano-enabled drug delivery

Introduction

As a basic knowledge resource, patents play an important role in identifying technology development trends and opportunities (Liu and Shyu 1997; Tsuji 2012), especially for emerging technologies. However, unlike research publications, patents do not have

J. Ma (✉)
School of Management and Economics, Beijing Institute of Technology, Beijing, China
e-mail: majing881003@163.com

A. L. Porter
School of Public Policy, Georgia Institute of Technology, Atlanta, GA, USA
e-mail: alan.porter@isye.gatech.edu

A. L. Porter
Search Technology, Inc., Atlanta, GA, USA

keywords to generalize their content. Usually it's difficult and time consuming to process or read the content of full texts of patents. And from the perspective of technology protection, most descriptions in titles and abstracts of patents are more obscure than those in articles gathered from global databases. So how to analyze patent content and understand topical information effectively and to identify potential innovation pathways and technology opportunities by using patent topical information is a key research question in patent mining.

Competitive technical intelligence (CTI) strives to answer the "reporter's questions"-who, where, when, what, how, and why? Effective information on the first three can be generated relatively well by treating field structured data gathered from database searches (e.g., in Web of Science or Derwent patents). The last two (how and why) usually require domain knowledge to interpret bibliometric tabulations. Our emphasis here is on answering the "what?" question-i.e., the topical content being treated in the patent set.

Analyzing and mining patent information can help derive information for CTI, technology development trend analyses and forecasting (Daim et al. 2006). Previous research has introduced text mining methods into patent content analysis. Tseng et al. (2007) described a series of text mining techniques applicable to patent analyses. From more detailed perspectives, these tools can be classified into diverse applications, including document processing, indexing, topic clustering, and topic mapping.

Many studies have been conducted on identifying technology trends and opportunities. Porter et al. (1994) and Porter and Detampel (1995) introduced Technology opportunities analysis (TOA). Watts et al. (1997) and Zhu and Porter (2002) advanced such analyses by using scientific and engineering literature more effectively. Kim et al. (2008) generated a patent map by using a semantic network of keywords from patent documents. The map offered a clear overview of patent information and enabled people to understand advances of emerging technologies. Hsu et al. (2006) proposed methodologies for patent map analysis, patent technology clustering, and patent document clustering to assist companies to gain information for R&D strategic planning. Choi et al. (2011), Yoon and Kim (2012) and Zhang et al. (to appear) detected signals of new technology opportunities using subject–action–object (SAO)-based semantic patent analysis. Trappey et al. (2011) analyzed Chinese RFID patents combining patent content clustering, and technology life cycle forecasting to identify a niche space of RFID development in China. Some traditional innovative methodologies suffer the limitations of being non-quantitative, while patent mining is an ideal supplementary tool. Daim et al. (2006) forecasted three emerging technology areas by integrating bibliometrics and patent analyses with other forecasting approaches like scenario planning and growth curves. Xu and Leng (2012) validated a technique of combining patent text mining and morphological analysis in an empirical study of liquid crystal display wide viewing angle patents. The method also provided advantages in terms of cost and time reductions in constructing morphological box and flexibility analyses. Yoon (2008) and Yoon et al. (2014) also combined morphological analysis and patent text mining to explore technological opportunities and roadmapping. These studies also extend past trends by fitting growth curves.

However, based on patent full text or abstract content itself, text mining is not overly effective. For a patent, there is often a considerable time lag between initial application and getting granted or denied (Trappey et al. 2010), let alone being indexed by patent databases, like Derwent Innovation Index (DII). And when we do text mining on patent content, some novel phrases or words are rarely extracted because of their low frequencies. However, such terms may offer significant clues about technology hotspots now and future technology development prospects. So it's desirable, but difficult, to gain first-hand

information from patent records to identify opportunities in terms of component technologies.

In this paper, we conduct an empirical study to try out alternative methods with DII data. Derwent patent titles and abstracts are preferred because technical specialists rewrite them to convey meaning more clearly than the original versions submitted to patent authorities.

Our case study focuses on nano-enabled drug delivery (NEDD) which is a very active emerging technology that engages several distinct technology spaces. We have been exploring different ways to enhance topical intelligence from patent compilations and trying to identify potential innovation pathways and technology opportunities in NEDD. Our previous study has focused on analyses of the development of NEDD using research and other publication data (Zhou et al. 2014; to appear). This paper keys on patent data to gain perspective on development (generally somewhat "downstream" from the research data) and prospective applications. And since patent topical information is more difficult to gather and analyze, the contribution of this paper is mainly on processes to generate more accurate and useful topical intelligence from patents.

There are five parts in this paper. This first part offers a general introduction of our research. In the second part, we introduce how we develop our data set and the process of patent topical analysis. The third section presents the results of our NEDD topical analysis. And the fourth part is analyzing pathways and opportunities of NEDD based on the topical mining results. In conclusion, we reflect on the advantages and disadvantages of our methods and point out promising research opportunities.

Data and methodology

In this paper, we use DII data-a high-quality data source for bibliometric analyses (Derwent World Patents Index 2014). Eight sets of search terms were applied to generate the data set (Zhou et al. to appear). The original data set we developed has 8,426 raw records from 1999 to 2012. In DII, each record reflects a patent family, a group of related inventions filed in one or multiple patent authorities.

Our general approach is to pursue bibliometric and text analyses of emerging technologies. We strive to develop methods to extract useful intelligence to inform technology roadmaps (Zhou et al. 2014; to appear) and to help forecast innovation pathways (Robinson et al. 2013).

To conduct this study, we introduce two software tools, VantagePoint [www.theVantagePoint.com] and ClusterSuite [program developed by J.J. O'Brien, with Stephen J. Carley, at Georgia Tech-to be available at www.VPInstitute.org] (O'Brien et al. 2013). VantagePoint is desktop (Windows environment) software for bibliometrics, natural language processing (NLP), data cleaning, analyses, and visualization. ClusterSuite is a set of routines developed through our Co-Lab collaboration to further refine text content extracted from the records in VantagePoint. This suite of algorithms consolidates noun phrases to provide more refined topical information.

As we began our analyses, we checked the accuracy of the NEDD data set by doing a rough principal components analysis (PCA) using about 500 top phrases consolidated from titles and abstracts in VantagePoint. During this process, we gained a general idea of topical information in these NEDD patents. The results showed that some patents addressed plants/transgenic plants/soybeans. We presented this rough result to a NEDD specialist, and he suggested these records should be removed. Some overlapped search

items between NEDD and transgenic plants may lead to this result. To remove these irrelevant records, we introduced some words or phrases to filter out those records-including soybean, insect resistance, transgenic plant, herbicide resistance/herbicide, plant, and seed. After applying the filters in titles & abstracts, we have 7,906 NEDD-related records for further analyses presented herein.

In this paper, our novel approach is to explore three different sources for patent topical content:

- *noun phrases from patent titles and abstracts,*
- *patent classes (especially, Manual Codes), and*
- *external imported keyword list [Medical Subject Headings (MeSH) terms from MEDLINE].*

Titles & abstracts are drawn from the patent abstract records themselves, as rewritten by Derwent personnel. Manual codes are a supplementary field from DII for patent classification, offering more technical emphasis than do International Patent Classifications (IPCs). And we introduce MeSH terms from MEDLINE to detect whether these external topical items can serve as keywords to provide new perspectives.

We propose a systematic process of patent topical analysis, as outlined in Fig. 1. We compare what can be gleaned from these three topical content fields separately, and in combination, to improve the accuracy of our analyses. Then, we use a suite of algorithms to further refine text content extracted from the records. And third, we use semantic techniques to improve our understanding of how different fields work within NEDD. The process helps us generate topics and sub-systems from patents so as to frame development trajectories for this technology field (and then to delve further for subfields). Then by analyzing the correlations and tracking R&D activities in these major sub-topics, we figure out a systematic way to identify opportunities and possibilities for NEDD application.

NEDD patent topical analysis

Keyword lists

The first task to conduct this topical analysis is to obtain term and phrase lists from the three sources noted. For titles and abstracts, we use VantagePoint's NLP (somewhat tailored for science and technology data analyses) to extract noun words/phrases. For Manual Codes, since they are hierarchical, we only use their most detailed levels. We then perform NLP on those detailed Manual Codes in VantagePoint to extract key words/phrases. To gather MeSH terms, we analyze 6,585 NEDD records retrieved from MEDLINE for 2012, and take the top 519 primary MeSH terms (those that appeared in at least 7 records). We import those 519 terms using VantagePoint's "MyKeywords" feature to tag these MeSH terms as they occur in titles and abstracts in our NEDD patent data set. We then create a field giving the frequencies of occurrences of these terms in our patent records. Some 441 out of the 519 MeSH terms are recognized in the NEDD patent data set.

The original NLP lists of words/phrases from titles and abstracts, and from manual codes, are not directly of great use. First, there are many meaningless items (e.g., numbers and general phrases). Second, many words or phrases only appear in one record. And then, inspection indicates that many similar items should be combined. To overcome these deficiencies in term quality, we use ClusterSuite to clean these lists. Numbers, punctuations, common and basic words, and extreme terms that appeared in too many or too few

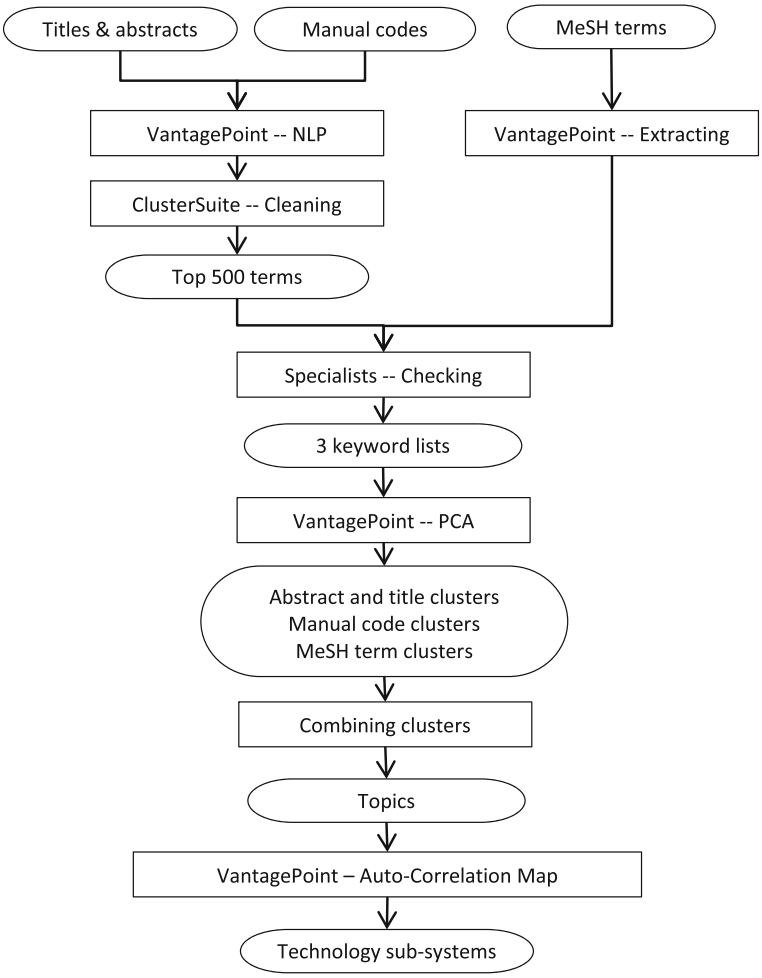

Fig. 1 Topical analysis process

records are eliminated. Similar words are clumped based on their similarity and co-occurrence in the records. Applying this set of cleaning algorithms dramatically reduces these lists, at the same time significantly improving the quality of the words/phrases.

After cleaning these lists, we pick out about 500 top (most frequent) items from each of them for further analyses. For words/phrases extracted from the titles and abstracts, the top 512 items appear in at least 81 records. For manual code terms, the top 502 items each appear in at least 35 records. And all 441 MeSH terms found in the patent abstract records are retained (see Table 1). Note the huge data reduction for the title & abstract terms and phrases.

As our next step, we send our three lists to a NEDD specialist. He checks the three lists for items that are too generally used in NEDD research or do not relate strongly to NEDD interests; those terms are removed. Especially in the MeSH term list, many words are too general to get at sub-topical emphases within NEDD (e.g., cell, water, oxide); so these words are removed. Finally, we get three lists for further topical analysis (the last column

Table 1 Three lists for topical analyses

#	List	Original No.	After ClusterSuite	Top "500"	For PCA
1	Titles and abstracts	253,939	7,208	512 >=81	345
2	Manual codes	5,213	1,538	502 >=35	344
3	MeSH	519	–	441	276

of Table 1)-345 items from the titles & abstracts; 344, from manual codes; and 276, from the MeSH terms. Terms in list 1 appear in 7,384 out of the 7,906 records; list 2 covers 7,640; and list 3 covers 7,844. So for the NEDD patent data, these three key words/phrases lists appear quite inclusive in tapping the content of the full set.

The three lists are relatively distinct, but with some overlap. Table 2 lists the numbers of duplicates between pairs of lists. For example, 63 items out of 345 from titles and abstracts are found in the list of manual codes. Conversely, only 36 items from the 344 manual codes are found in the list extracted from the titles and abstracts. Why are the two numbers not equal? Since the descriptions of these lists are not the same, we use fuzzy matching to assess these duplicates. For example, in MeSH terms, there is only one "nucleic acid," but in the list of titles & abstracts, we have "nucleic acid," "nucleic acid sequence," and "nucleic acid encoding"; in the list of Manual codes, we have "nucleic acid & hybridization probes," "peptide nucleic acid," and other descriptions. So this MeSH "nucleic acid" item can be found in titles & abstracts, and in Manual codes, but similar "nucleic acid" items may not be found in the MeSH terms. From Table 2 we can see, even though there are some duplicates in these three lists, the content is largely distinct.

The aforesaid cleaning process should be adaptable to diverse keywords and phrases lists. It is an efficient way to obtain substantial basic topical information from patent data. Especially when information from patents themselves is not qualified or sufficient, some imported or related external information should be considered as an aid.

NEDD topic clusters and sub-systems

To obtain topical information descriptive of NEDD patents, we use Factor Maps in VantagePoint, based on principal components analysis (PCA) (Newman et al. 2012), to cluster each of the three keyword lists. We get 18 clusters from List 1 (titles and abstract phrases), 18 clusters of List 2 (manual codes), and 17 clusters for List 3 (MeSH terms).

These three cluster results showcase different characteristics of NEDD. The three cluster sets are not equally descriptive of all types of topics. List 1 (titles & abstracts) describes drugs and targets more richly; in other words, the patents prominently including these words/phrases are applicable to some treatments for many kinds of diseases. But we cannot find many clear clues about processes or how they work. In patent abstracts, the description attends more to the results instead of how to get them. In the Derwent data, titles & abstracts of patents have been rewritten to enhance the quality of description. Even though they are better and descriptive, we find that most patent abstracts describe more about uses and advantages. That is they have detailed descriptions of what diseases a new formulation or protein composition can be used to treat, but less description of how these formulations or compounds are prepared and what the mechanisms are. The results for List 2 present some clusters about methods that these patents use, not very detailed but better

Table 2 Overlaps between the three topical lists

	X: Titles and abstracts	X: Manual codes	Y: MeSH terms
Y: Titles and abstracts	–	36/63	65/51
Y: manual codes	63/36	–	55/22
Y: MeSH terms	51/65	22/55	–
# of X found in Y/# of Y found in X			

than List 1. Clusters in List 3 offer relatively clearer descriptions about nanoparticles and drugs. But in the results of List 2 and 3, most targets and diseases are not specified, just words like "anticancer," "antiallergic," or so on.

For a better understanding of NEDD patent topics, we combine the results from these three lists. Some clusters are merged based on their similar and overlapping items. We also delete some clusters with low record coverage, unclear boundaries, or weak linkage with other clusters. For example, we find that all three lists have some clusters or items about neurodegenerative diseases, like Alzheimer's and Parkinson's, so we combine these clusters into one topic (#7). Another example is #1, this topic is mainly from the titles and abstracts group. It covers a large portion of the records and is closely connected with many other clusters, so it is kept. We also delete some clusters that do not appear significant-e.g., from the MeSH group, one with just two items, herpesvirus 1 and herpesvirus 4, covering two records. In all we manually compose 13 basic topic clusters-see Table 3. They cover 7,466 records−94.43 % of the patent set. The last column of Table 3 indicates from which lists each topic mainly derives.

To clarify the relationships among these 13 topics, Fig. 2 maps them based on the degree to which they associate with the same patent records. Based on the details of these clusters and their correlations, we divide them into four sub-systems:

- P: process/mechanism
- D: drug/composition
- C: carrier/vector
- T: target/disease.

In Fig. 2, we can see the four sub-systems have relatively clear boundaries. However, topics connect within and across sub-systems. And these connections reflect their correlations.

For "P: process/mechanism" sub-system, most topics are about gene or protein expression. Patents within this subsystem mentioned much about treating diseases through new kinds of nucleic acid sequences or other agents to alert gene expression. And in this sub-system, methodologies like RNA interference, hybridization probes, and so on were applied. But it's not clear how to deliver these substances to targets or cells.

In the "D: drug/composition" sub-system, we identify two topics. One is anti-cancer drugs, containing doxorubicin, paclitaxel, and some other agents to treat cancers. We name the other topic pharmaceutical compounds. In this topic, we don't identify specific agents, but, rather, some general types of compounds as formulas or preparation materials, such as fused ring, polyamine, and so on.

An important component of NEDD is use of nanoparticles as drug carriers. In the "C: carrier/vector" sub-system, we have three topical concentrations. One is about liposomes and capsules; one, polymers, like polyethylene glycol; the other one, some kinds of metal used to prepare nanotubes and other nanoparticles. Also based on the connection between

Table 3 Basic topics in NEDD patents

#	No. of records	Topics	Words or phrases	Lists
1	3,437	P: expression, nucleic acid encoding	Expression, polypeptide, polynucleotide, host cell, isolated polypeptide, isolated polynucleotide, inhibitors, recombinant vector, microarrays, polypeptide activity, nucleic acid amplification reaction, gene chips, exogenous polynucleotide, expression dose	List 1
2	3,422	P: nucleic acids and hybridization probes	Detection, nucleic acid & hybridization probes, nucleic acid hybridization test methods and nucleic acid probes, polarography and enzyme processes, DNA amplification method	List 1 List 2
3	3,319	P: RNA interference	RNA, RNA interference, siRNA, sense strand, antisense strand	List 1 List 2 List 3
4	3,241	P: immune response	Antibody, antigen, T cell, interferon, cytokine, interleukin, macrophage, CD4, CD8	List 3
5	1,446	T: autoimmune diseases	Multiple sclerosis, Crohn's disease, systemic lupus erythematosus, ulcerative colitis, glomerulonephritis, myasthenia gravis, scleroderma, atopic dermatitis, Addison's disease, Sjogren's syndrome, adult respiratory distress syndrome, uveitis, allergic rhinitis, Grave's disease, spondylitis, Hashimoto's thyroiditis, autoimmune hemolytic anemia, dermatomyositis, polymyositis, urticaria, autoimmune hepatitis, primary biliary cirrhosis	List 1 List 2
6	1,433	C: metal nanoparticles	Nanoparticle, temperature, colloid, silicon, iron, gold, nanotube, nanostructure, titanium, silver, nanowire	List 3
7	1,368	T: neurodegenerative diseases	Alzheimer's disease, Parkinson's disease, Huntington's disease, Amyotrophic lateral sclerosis, neurodegenerative diseases, septic shock, spinal cord injury	List 1 List 2
8	1,028	D: pharmaceutical compounds	Fused ring, mononuclear heterocyclic, polyamines, carboxylic acid & phenol present, carboxylic amides, polycarboxylic acid, aliphatic and cycloaliphatic mono and sulfur acid, amidine and biguanide and guanidine, iso cyanide, hydrazine	List 1 List 2
9	997	T: cancers	Breast cancer, prostate cancer, lung cancer, ovarian cancer, pancreatic cancer, colon cancer, liver cancer, cervical cancer, bladder cancer, skin cancer, colorectal cancer, stomach cancer, brain cancer, neck cancer, testicular cancer, esophageal cancer, gastric cancer	List 1
10	676	P: fusion proteins	Fusion protein, polypeptide production, fusion genes and transgenes, encoding fusion protein	List 2
11	584	C: polymers	Polyethylene, polyethylene glycol, imine, polyethyleneimine, polyether, polyalcohol	List 2 List 3

Table 3 continued

#	No. of records	Topics	Words or phrases	Lists
12	495	C: micro capsules and liposomes	Liposome, emulsion, capsule, micelle	List 3
13	341	D: anticancer drugs	Doxorubicin, paclitaxel, cisplatin, methotrexate, fluorouracil, camptothecin, dexamethasone	List 1 List 3

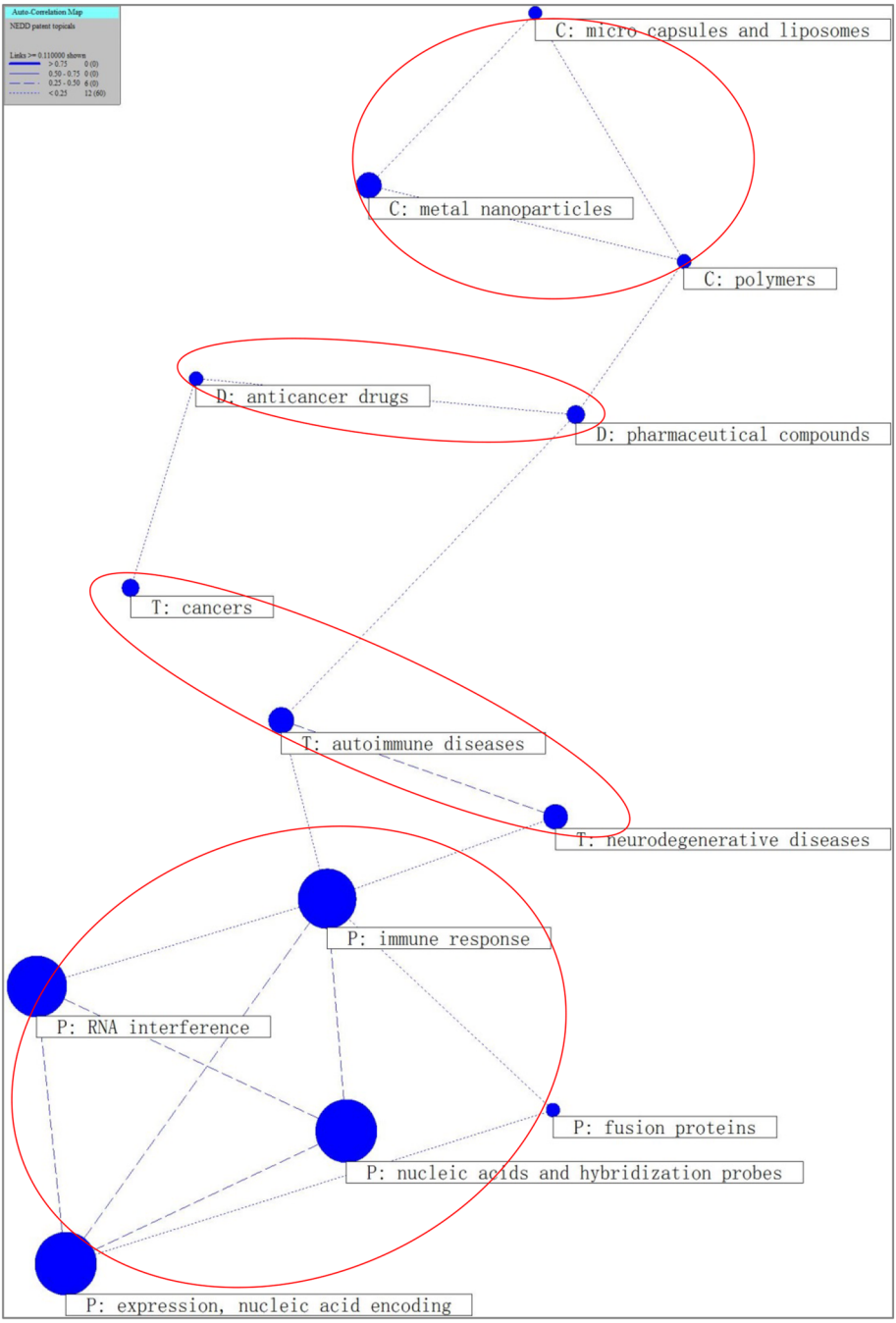

Fig. 2 Auto-correlation map of the 13 NEDD topics

"polymers" and "pharmaceutical compounds," we deduce that there may be some overlaps for some kinds of polymers applied in NEDD both as drugs and carriers.

"T: target/disease"-at last, the main purpose of NEDD is treating diseases. Here we only keep three main kinds of diseases-autoimmune disease, neurodegenerative disease, and cancer. These three are all generated from PCA results instead of our existing knowledge. There are also descriptions about other diseases in our data set, such as virus infection, which are scattered and not systematic. So at this stage, we ignore them.

In a workshop held at Georgia Tech (February 2014), we discussed these topics and sub-systems. Five NEDD researchers deemed the 13 topics and 4 sub-systems to be suitable. This offers valuable validation of this text mining.

Also, from the perspective of patent analysis, if we only use text from patent records themselves, in Table 3, we will totally miss Topics 4, 6, and 12 and maybe parts of Topics 3, 11, and 13. If so, in Fig. 2, we may fail to generate sub-system "C: carrier/vector", which may make the correlation map incomplete.

Analyzing pathways and opportunities in NEDD

NEDD development pathways

To observe the development of these topics and sub-systems, we draw Fig. 3. Since data in 1999 and 2012 may not be complete, we focus on 2000–2011. In Fig. 3, the trends show that in 2001 and 2002, many topics-especially in sub-system "P: process/mechanism"-had high patent productivities. Even though there was a decline in 2003, these topics have been at a relatively high level. Patents from these topics overlap. These similar trends confirm that people have been exploring the secrets of our genome for decades. The publishing of working drafts of the Human Genome Project (HGP) consortium and Celera in February 2001 is regarded as one of the most significant milestones in the history of human genetic research. It also corresponds with a revolution in vaccine development, searching for new drugs, pharmacogenomics and pharmacogenetics, and gene testing and gene therapy (Ikekawa and Ikekewa 2001). We believe this event also contributed to the spurt of patenting in related NEDD topics in 2001 and 2002. Also some policy changes may be related to this spurt, for example, the 'American Inventors Protection Act' of USPTO (the United States Patent and Trademark Office) in 2000.

In Fig. 3, RNA interference has experienced another two peaks in 2006 and 2011, showing an increasing trend. From 2006 to 2008, there was a rise in patenting of pharmaceutical compounds, but in recent years, this declines. Anticancer drugs, the other topic in the "D: drug/composition" sub-system, keeps increasing, but not significantly.

An exciting fact is that all three topics in "C: carrier/vector" have been on the uptrend, especially metal nanoparticles, which shows rapid growth in recent years. At the same time, three topics in "T: target/disease" stay essentially stable, without remarkable changes.

A better way to understand Fig. 3 is to clarify these topics. In Fig. 4, we locate these topics along a time sequence based on the sub-system to which they belong. We separate time into three periods. Each oval stands for one topic. The positions of these ovals represent their most active development time period. From this perspective, we can nominate a general innovation pathway of NEDD patents. The general idea of NEDD is to deliver drugs in the right place at the right time, so as to improve their effectiveness.

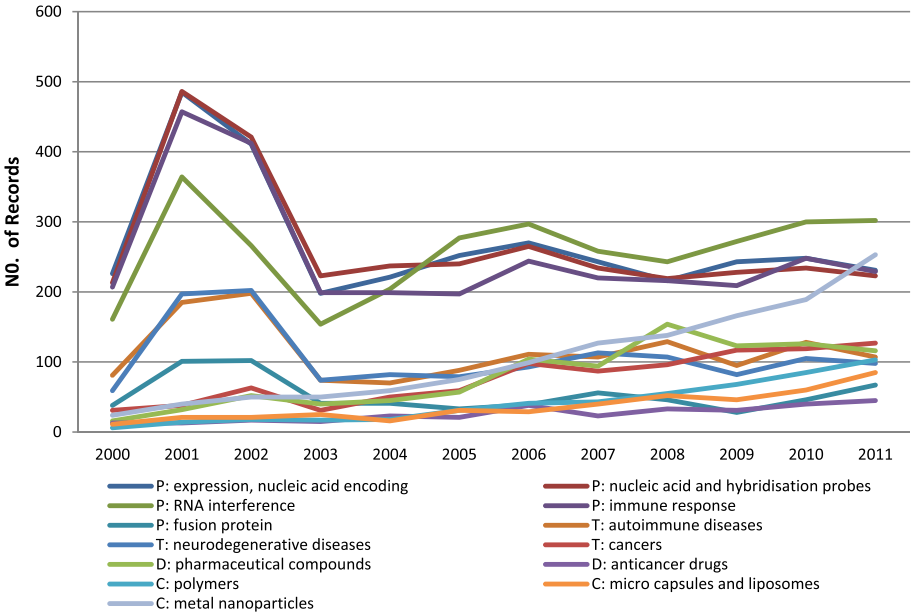

Fig. 3 Development trends of the 13 NEDD topics

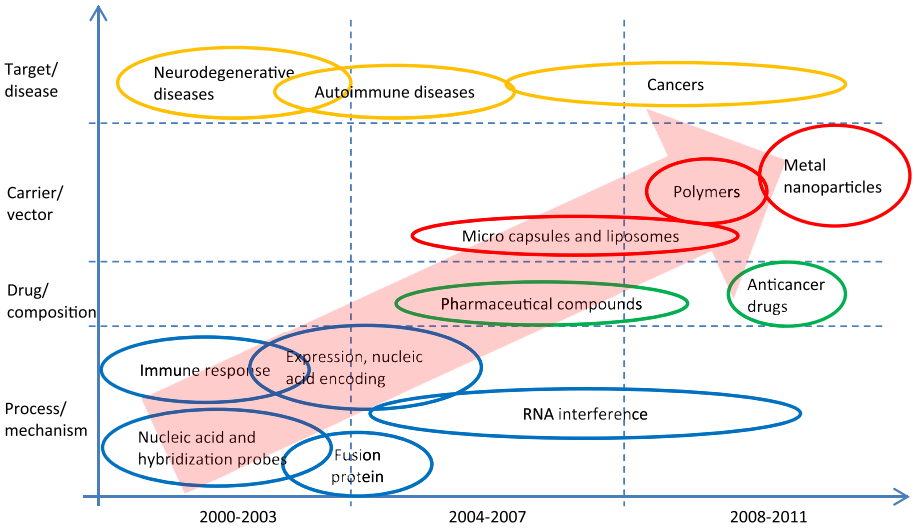

Fig. 4 Developmental pathways, locating the 13 NEDD topics

Consider the four sub-systems we include for a general NEDD framework. Drugs or gene therapy agents are passengers. Their combination with carriers is the basic unit of treatment. After the unit arrives at the destination, and passengers (active agents) are released, they will influence target cells or other mechanisms of the body (human or animal)-e.g., alerting gene expression or suppressing protein production, and finally have a

treatment effect. From Fig. 4, we can see people first work to solve the problems about the processes and mechanisms in the body, and the materials to complete these processes, and then start to solve the problems about how to deliver them to targets, with better carriers. Topics about process on the left are more exploratory, but on the right, topics about carriers and drugs are more specific and focused on NEDD application. *The hotspot in NEDD invention seems to have shifted from process/mechanism to carrier/vector.*

In terms of topical analysis, we find that most recent topics, like metal nanoparticles, polymers, and anticancer drugs, mainly come from the MeSH terms (see Table 3). Since these topics are relatively new, there is only a small portion of the NEDD patents on these topics, and the frequencies of key words or phrases of these patents are small. It's difficult to identify these new items in lists of titles and abstracts, and manual codes, and they rarely appear in the results of the corresponding PCA analyses. Research publications may reveal more novel topics. Since we use the top 519 MeSH terms in MEDLINE for NEDD 2012 publications, this resource is an efficient supplement to the original topical information from patents. It offers some new, lower-frequency items in patents' titles and abstracts and Manual codes, and helps us to find new hotspot applications in patents that have been relatively well-studied in publications recently.

Through analyzing the development trends of the different topics, we also figure out some promising topics that catch our attention and suggest that there may be special opportunities. Here we choose the one that has increased most significantly recently (Fig. 3) to conduct some detailed analysis-metal nanoparticles.

Metal nanoparticles in NEDD

For NEDD, many kinds of materials can be used to produce nanoparticles, such as proteins, peptides, polymers, lipids, metals and metal oxides, and carbon. Among these materials, metal nanoparticles are attractive options for drug and contrast agent delivery because they provide a stable platform on which multiple functionalities can be grafted; some can be imaged directly; and certain formulations allow magnetic-directed guidance (Janib et al. 2010). In our data, the number of patents involving metal nanoparticle topics increased from 24 in 2,000–253 in 2011 (Fig. 3).

Four metals appear prominently in our topic items-iron, gold, titanium, and silver. Gold nanoparticles first appeared in 2004, and have become more and more popular in recent years. There are 43 patents related to gold in our data set. To detect how these metals are applied and their potential usage, we analyze titles & abstracts of these 43 patents. We clarify some key points in Fig. 5.

There are three patents out of 43 not directly used in the medical area, but related to gold nanoparticle or polymer preparation. Two patents are about traditional Chinese medicine, in which "gold" is used in herbs' or materials' names. There is one patent in which gold is treated as a kind of impurity. The rest of these patents are related to NEDD and gene therapy. Fig. 5 suggests the main applications and materials in the gold nanoparticle patents. Gold nanoparticles are applied most frequently in nanoparticle carriers and reagent strips. Fourteen patents are about gold nanoparticles used in carriers, and seventeen are used in reagent strips. Colloidal nanosilver or nanogold composition is mentioned in one patent to be used as an antimicrobial agent. Another five patents are about how to prepare gold nanoparticles, like using thermal decomposition or phage, and producing wire-form gold particles. As an important material, colloidal gold is widely used for reagent strips. They are applicable to many kinds of targets, like nucleic acids, drugs

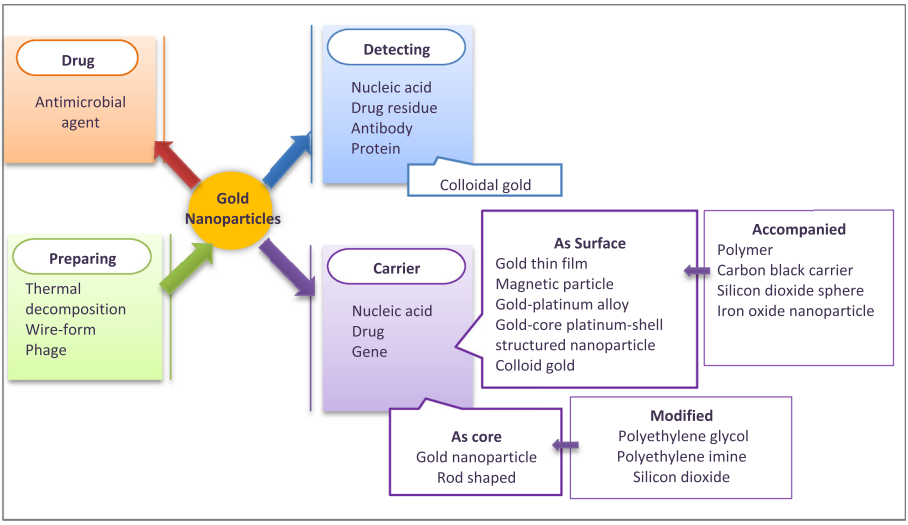

Fig. 5 Applications and preparing of gold nanoparticles

antibodies, and other substances. Almost all of these reagent strips describe their advantages as sensitive and rapid detection.

As carriers, gold nanoparticles can be used as a core part or surface part in a vector. These carriers can be used to deliver drugs, nucleic acids, or genes. When gold nanoparticles are applied as a coating material, they are usually in thin film, alloy, or colloid form. Also, they may cooperate with other core nanoparticles, like polymer nanoparticles, carbon black carriers, silicon dioxide spheres, and so on. Because gold nanoparticles are very stable (Alkilany et al. 2013), they are an attractive choice for the surface of carriers. They can help to reduce the degradation of nucleic acids. By applying gold magnetic particles and magnetic fields, the vectors can be transferred to targets rapidly and accurately without damage to the cell. As core in vectors, gold nanoparticles usually should be modified by other particles, like polyethylene glycol, polyethylene imine, and silicon dioxide. It's also important that gold nanoparticles are often rod shaped when they are used as main parts in vectors.

This analysis has shown that gold nanoparticles are a promising material in NEDD. In our data set, there was only one patent about gold nanoparticles in 2004, rising to 14 in 2011. Were it not for use of the terms derived from recent NEDD research publication activity (List 3, MeSH terms imported), patent analysis would not likely attend to this component "hot topic." Gold nanoparticles have many remarkable characteristics, like stability, and can be widely used in reagent strips, drug carriers, and imaging. Fig. 5 can be treated as a clue for innovations of gold nanoparticles in the NEDD field. By combining different factors from different perspectives, new opportunities may be found.

Conclusions

Patents provide an important resource in identifying technology development trends and opportunities. But based on the limited and obscure topical descriptions in their text, it is

hard to analyze patent data accurately. In this paper, we try to analyze technology development pathways and figure out high potential opportunities in the NEDD field through patent topical information analysis from the perspective of non-domain experts.

There are mainly two parts in this paper. First, we propose a systematic process to obtain patent topical information. This process combines a) three different topical information sources for patent abstract data sets, b) via professional software for data extracting and cleaning, and c) eliciting input of specialists to help interpret the empirical analyses. We believe this composite approach offers good potential for scientometricians to pursue patent analyses. A workshop including five NEDD researchers and several science policy researchers found the topical analyses meritorious. They advocated further analyses at a finer grain-i.e., focusing on one target (disease or organ) and/or one delivery mechanism (etc.) at a time. They suggested that the patterns discerned at more specific levels could provide sharper insights. Furthermore, such component level patterns could constitute leading indicators to anticipate developmental trajectories for other, new specific foci (technologies and/or targets).

We get 13 NEDD topics by using PCA and divide them into four sub-systems. This result helps us to gain a general framework of NEDD invention activity. These 13 topics cover 94 % of our patent search set, and the four sub-systems have relatively clear boundaries. These results are efficient and are the bases for our following analyses. This process of topical analysis proves to be effective, especially through import of external topical information sources (pertinent MeSH terms) to supplement the insufficient original topical information of patent abstract records. It provides an approach to identify novel but low-frequency words/phrases from patent content.

In "Analyzing pathways and opportunities in NEDD" section, we analyze the developmental pathway in NEDD and focus on a specific topic, metal nanoparticles, especially gold nanoparticles. The hotspot in NEDD invention seems to have transferred from process/mechanism to carrier/vector development. Metal nanoparticles are quite promising. We analyze the latest gold nanoparticles that are applied to reagent strips and drug carriers. These applications are split into different components, and their combinations may indicate new opportunities. Characteristics of gold nanoparticles are also generated to help nominate future innovations.

Based on a relatively accurate and efficient patent topical analysis, we can generate a framework of NEDD development pathways and identify some potential opportunities. Such an analytical methodology can significantly improve the availability of technical intelligence for multiple purposes.

Some cautions are in order. The data we use are DII patent abstract records; these have been rewritten by Derwent indexers. The behavior of the terms extracted by NLP, as well as those from the frequently occurring Derwent Manual Codes, could well be peculiar to Derwent data. It would be interesting to compare the differences of topical analyses from DII data with other data for the same records. This might be done by performing the same search in a first level patent database (e.g., Micropatents, Questel). Or, one could go further to explore differences in content emphases by retrieving full patent texts for a suitable sample set to compare with analyses of the corresponding abstract records.

Also it will be more useful to domain specialists if we can combine the patent topical analysis results with the results of other data sources, such as publications (Zhou et al. 2014) and web pages, to obtain a better understanding of technology development patterns and trends. We have conducted two workshops with domain experts-the one noted above at Georgia Tech and a later one in conjunction with a drug delivery conference. Both confirmed the value in combining empirical analyses with expert perspectives to interpret

those analyses. In this case, the domain experts pointed to more value from sub-level analyses. These encouraged the sub-system analyses noted and additional pointed probes. We are presently exploring NEDD application to brain chemistry and Alzheimer's disease. We aspire to gain understanding in parallels between those target applications. We are also comparing NEDD cancer applications to look for commonalities that may speak to R&D opportunities (e.g., might research keying on improved drug delivery for cancer X hold additional promise for application to other cancers or other diseases with common target organs-such as the brain).

Acknowledgments We acknowledge support from the US National Science Foundation, Science of Science & Innovation Policy (SciSIP) Program (Award No. 1064146-"Revealing Innovation Pathways: Hybrid Science Maps for Technology Assessment and Foresight"). The findings and observations contained in this paper are those of the authors and do not necessarily reflect the views of the US National Science Foundation. The authors would like to thank thank Xiao Zhou, Douglas K.R. Robinson, Ying Guo, and Min Suk Shim for contributions to the NEDD analyses. We also thank our colleagues of the Innovation Co-lab from Beijing Institute of Technology, Georgia Tech, and the University of Manchester for their feedback.

References

Alkilany, A. M., Lohse, S. E., & Murphy, C. J. (2013). The gold standard: Gold nanoparticles libraries to understand the nano-bio interface. *Accounts of Chemical Research, 46*(3), 650–661. doi:10.1021/ar300015b.

Choi, S., Yoon, J., Kim, K., Lee, J. Y., & Kim, C. H. (2011). SAO network analysis of patents for technology trends identification: a case study of polymer electrolyte membrane technology in proton exchange membrane fuel cells. *Scientometrics, 88*(3), 863–883. doi:10.1007/s11192-011-0420-z.

Daim, T. U., Rueda, G., Martin, H., & Gerdsri, P. (2006). Forecasting emerging technologies: Use of bibliometrics and patent analysis. *Technological Forecasting and Social Change, 73*(8), 981–1012. doi:10.1016/j.techfore.2006.04.004.

Derwent World Patents Index, [online] http://en.wikipedia.org/wiki/Derwent_World_Patents_Index (Accessed 13 Feb 2014).

Hsu, F. C., Trappey, A. J. C., Trappey, C. V., Hou, J. L., & Liu, S. J. (2006). Technology and knowledge document cluster analysis for enterprise R&D strategic planning. *International Journal of Technology Management, 36*(4), 336–353. doi:10.1504/IJTM.2006.010271.

Ikekawa, A., & Ikekewa, S. (2001). Fruits of Human Genome Project and Private Venture, and Their Impact on Life Science. *Yakugaku Zasshi Journal of The Pharmaceutical Society of Japan, 121*(12), 845–872. doi:10.1248/yakushi.121.845.

Janib, S. M., Moses, A. S., & MacKay, J. A. (2010). Imaging and drug delivery using theranostic nanoparticles. *Advanced Drug Delivery Reviews, 62*(11), 1052–1063. doi:10.1016/j.addr.2010.08.004.

Kim, Y. G., Suh, J. H., & Park, S. C. (2008). Visualization of patent analysis for emerging technology. *Expert Systems with Applications, 34*(3), 1804–1812. doi:10.1016/j.eswa.2007.01.033.

Liu, S. J., & Shyu, J. (1997). Strategic planning for technology development with patent analysis. *International Journal of Technology Management, 13*(5–6), 661–680. doi:10.1504/IJTM.1997.001689.

Newman, N.C., Porter, A.L., Newman, D., Trumbach, C.C., and Bolan, S.D. (2012). 'Comparing methods to extract technical content for technological intelligence' In *PICMET '12: Proceedings-Technology Management for Emerging Technologies*, IEEE, Vancouver, Canada, 1279–1285.

O'Brien, J.J., Carley, S., & Porter, A.L. (2013) Keyword field cleaning through ClusterSuite: A term-clumping tool for VantagePoint software. *Poster presented at Global Tech Mining Conference*. 25 Sep 2013. Atlanta, USA.

Porter, A. L., & Detampel, M. J. (1995). Technology Opportunities Analysis. *Technological Forecasting and Social Change, 49*(3), 237–255. doi:10.1016/0040-1625(95)00022-3.

Porter, A. L., Jin, X.-Y., Gilmour, J. E., Cunningham, S., Xu, H., Stanard, C., et al. (1994). Technology opportunities analysis: Integrating technology monitoring, forecasting & assessment with strategic planning. *SRA Journal (Society of Research Administrators), 21*(2), 21–31.

Robinson, D. K. R., Huang, L., Guo, Y., & Porter, A. L. (2013). Forecasting innovation pathways (FIP) for new and emerging science & technologies. *Technological Forecasting and Social Change, 80*(2), 267–285. doi:10.1016/j.techfore.2011.06.004.

Trappey, C. V., Trappey, A. J. C., & Wu, C. Y. (2010). Clustering patents using non-exhaustive overlaps. *Journal of Systems Science and Systems Engineering, 19*(2), 162–181. doi:10.1007/s11518-010-5134-x.

Trappey, C. V., Wu, H. Y., & Taghaboni-Dutta, F. (2011). Using patent data for technology forecasting: China RFID patent analysis. *Advanced Engineering Informatics, 25*(1), 53–64. doi:10.1016/j.aei.2010.05.007.

Tseng, Y. H., Lin, C. J., & Lin, Y. I. (2007). Text mining techniques for patent analysis. *Information Processing and Management, 43*(5), 1216–1247. doi:10.1016/j.ipm.2006.11.011.

Tsuji, Y. S. (2012). Profiling technology development process using patent data analysis: a case study. *Technology Analysis & Strategic Management, 24*(3), 299–310. doi:10.1080/09537325.2012.655417.

Watts, R.J., Porter, A.L., Cunningham, S.W., and Zhu, D.H. (1997) TOAS intelligence mining; Analysis of natural language processing and computational linguistics in *Proceedings of First European Symposium on Principles of Data Mining and Knowledge Discovery*, Springer-Verlag, Trondheim, Norway, 323–334.

Xu, F., & Leng, F. H. (2012). Patent text mining and informetric-based patent technology morphological analysis: an empirical study. *Technology Analysis & Strategic Management, 24*(5), 467–479. doi:10.1080/09537325.2012.674669.

Yoon, J., & Kim, K. (2012). Detecting signals of new technological opportunities using semantic patent analysis and outlier detection. *Scientometrics, 90*(2), 445–461. doi:10.1007/s11192-011-0543-2.

Yoon, B., Park, I., & Coh, B. (2014). Exploring technological opportunities by linking technology and products: Application of morphology analysis and text mining. *Technological Forecasting and Social Change, 86*, 287–303. doi:10.1016/j.techfore.2013.10.013.

Yoon, B., Phaal, R., & Probert, D. (2008). Morphology analysis for technology roadmapping: applicationof text mining. *R&D Management, 38*(1), 51–68. doi:10.1111/j.1467-9310.2007.00493.x

Zhang, Y., Zhou, X., Porter, A.L., and Gomila, J.M.V., and Yan, A. (to appear), Triple Helix innovation in China's dye-sensitized solar cell industry: Hybrid methods with Semantic TRIZ and technology roadmapping, *Scientometrics.*

Zhou, X., Porter, A.L., Robinson, D.K.R., Zhang, Y., and Guo, Y. (to appear) Nano-enabled drug delivery: Recent trends, emerging issues, and future directions, in Islam, N. (Eds.), *Nanotechnology: Recent Trends, Emerging Issues and Future Directions*, Nova Science Publishers, Hauppauge, NY.

Zhou, X., Porter, A.L., Robinson, D.K.R., Shim, M.S., and Guo, Y. (2014), Nano-enabled drug delivery: A research profile, *Nanomedicine: Nanotechnology, Biology and Medicine*. *10*(5), 889-896. doi:10.1016/j.nano.2014.03.001

Zhu, D. H., & Porter, A. L. (2002). Automated extraction and visualization of information for technological intelligence and forecasting. *Technological Forecasting and Social Change, 69*(5), 495–506. doi:10.1016/S0040-1625(01)00157-3.