Technological Forecasting & Social Change xxx (xxxx) xxx–xxx

Contents lists available at ScienceDirect

Technological Forecasting & Social Change

journal homepage: www.elsevier.com/locate/techfore

Identifying translational indicators and technology opportunities for nanomedical research using tech mining: The case of gold nanostructures

Jing Ma[a,*], Natalie F. Abrams[b], Alan L. Porter[c,d], Donghua Zhu[e], Dorothy Farrell[b]

[a] College of Management, Shenzhen University, Shenzhen, China
[b] Office of Cancer Nanotechnology Research, NCI, NIH, Bethesda, MD, USA
[c] School of Public Policy, Georgia Institute of Technology, Atlanta, GA, USA
[d] Search Technology, Inc., Atlanta, GA, USA
[e] School of Management and Economics, Beijing Institute of Technology, Beijing, China

ARTICLE INFO

Keywords:
Tech mining
Bibliometric analysis
Translational science
Gold nanostructures
Nano-medical research

ABSTRACT

Clinical translation of scientific discoveries from bench to bedside is typically a challenging process with sporadic progress along its trajectory. Analyzing R&D can provide key intelligence on advancing biomedical innovation in target domains of interest. In this study, we explore the feasibility of using a streamlined tech mining approach for identification of translational indicators and potential opportunities, using observable markers extracted from selected research literature. We apply this strategy to analyze a set of 23,982 PubMed records that involved gold nanostructures (GNSs) research. Nine indicators are generated to assess what different GNSs research activities had achieved and to predict where GNSs research will likely go. We believe such analysis can provide useful translation intelligence for researchers, funding agencies, and pharmaceutical and biotech companies.

1. Introduction

The development of R&D activities follows respective trajectories regarding different domain features. These trajectories could be initiated or shifted by either findings from basic research or demands from customer markets. Along each technology developmental pathway, visible or invisible milestones, as readiness indicators, provide clues to identify further development opportunities. To delineate these opportunities, it is important to consider domain-specific progressive properties.

In biomedical fields, scientific and technical innovation has facilitated the development of countless new therapies and medical devices. Yet, developmental trajectories of specific technologies are often uneven and challenging. Clinical translation of early discoveries "from bench to bedside" is often described as a slow and incremental process (with an average lag of 17 years) (Morris et al., 2011). This time lag can be as long as 30 years for specific topics, such as nano-enabled drug delivery (Wilhelm et al., 2016). Capturing research activities and other observable events in technology development can provide a better understanding of the ongoing biomedical R&D process and innovation prospects.

Nanotechnologies hold much promise in biomedical research. Companies, agencies, and researchers are aware of the transformational potential of nanotechnology in biomedicine and continue to pursue its development. As a result, over the past decade, government support for nanotechnology-based research has increased dramatically all over the world. For example, the US National Institutes of Health (NIH) invests $450M per year into nanotechnology research and training, including $150M per year invested by the National Cancer Institute (NCI) alone (Dickherber et al., 2015).

Nanotechnology-based translational research often requires highly coordinated and cost-effective programs, as well as timely assessments of the nanotechnology research landscape. To take academic discoveries further, researchers need to identify, optimize, and validate the most promising nanotechnologies. However, it is difficult to measure translational readiness due to a lack of a consensus conceptual model of how research is translated into patient benefits.

To assess translational readiness, several studies have attempted to define nanomedical technology development models. Etheridge and coauthors proposed a linear nanotechnology development pipeline based on the T-phases model adopted by NIH (Etheridge et al., 2013). An alternative framework developed by Trochim et al. (2011) relies on the process marker model of clinical translation that involves many distinct and observable markers. These models can be useful for assessing

[*] Corresponding author.
E-mail address: majing88@szu.edu.cn (J. Ma).

https://doi.org/10.1016/j.techfore.2018.08.002
Received 27 December 2017; Received in revised form 30 May 2018; Accepted 2 August 2018

translational readiness, but involve time consuming steps, such as manual review and indexing of research literature (Venditto and Szoka, 2013; Weber, 2013).

Based on bibliometric, text mining and other tools applied to science and technology resources, especially literature compilations, "tech mining" has been proposed as an approach to analyze technological innovation progress and generate useful intelligence (Porter and Cunningham, 2005). In this study, we introduce tech mining to develop a framework for article classification and link prediction to track translational readiness of specific nanotechnologies. The framework relies on the process marker model of clinical translation introduced by Trochim et al. (2011). We hypothesize that observable translational markers and milestones can be extracted from published literature and used to categorize publications automatically and, ultimately, to assign specific nanotechnologies according to different stages of translation. We also assume that the research network constructed with various technological elements has a self-organized development pattern, which can predict future opportunities to some extent. We believe that the methodology described here can provide valuable insights into translational readiness of promising technologies, such as gold nanostructures (GNSs) analyzed here.

2. Background

2.1. Technology opportunities analysis (TOA)

Introduced by Dr. Alan Porter and his team in the early 1990s (Porter et al., 1995; Porter and Detampel, 1995), technology opportunities analysis (TOA) has been developed to take further advantage of abundant online sources, such as publication and patent abstract record compilations, to achieve a better understanding of science and technology development. One emphasis of TOA is to focus on quantitative analysis by building hybrid tech mining models combining bibliometric, statistical, data mining, and text mining approaches and tools for specific goals that may benefit researchers, policy makers and enterprises (Porter and Cunningham, 2005).

Based on this concept, technology opportunity has been defined from both macro and micro perspectives. These include consideration of developmental pathways, hot spots and technology emergence, technological vacancy, collaboration potentials, and so forth. This places TOA as a branch of technological forecasting (Noh et al., 2016; Song et al., 2015). Much effort has been made to explore TOA from both methodological and application points of view.

We draw upon a methodological stream that strives to extract intelligence from sets of abstract records on a topic of interest. Such sets are retrieved from global databases (such as MEDLINE) for further analyses. First, hybrid approaches have been explored to extract useful words/phrases from textual data, such as title and abstract, or full text. For example, Zhang et al. (2014a) proposed a streamlined cleaning process for extracting and combining keywords that extracted from free text to improve topical analysis. Ma and Porter (2015) compared keywords from different fields and combined them to generate more descriptive topics for tracing technological pathways. Another notion is to combine the meaning of terms with their part of speech (POS) information in sentences. Yoon et al. (2013), Yoon and Kim (2011) and Guo et al. (2016) both used SAO (subject-action-object) semantic structure to represent different technological elements and to track technological changes. Then, to further explore connections between different scientific/technological directions or elements, topical analysis, data clustering and similarity calculation are often used (Newman et al., 2014; Ogawa and Kajikawa, 2017; Tseng et al., 2009).

An application orientation focuses on introducing other methods or tools into TOA to solve specific problems, such as technology road mapping (Choi et al., 2013; Zhang et al., 2014b, 2013) and morphological analysis (Boon and Park, 2005). Most of these studies are case focused. Emerging fields, for example, nanotechnology (De Miranda Santo et al., 2006; Ma and Porter, 2015), dye-sensitized solar cells (Ma et al., 2014; Zhang et al., 2014b), and big data (Zhang et al., 2016), have attracted researchers' attention since they are more active and may have higher potential.

While most TOA studies have focused on tracking topical level changes within a field objectively, this paper aims to assess technology readiness by defining different indicators, and to further predict micro-level technological links by treating development as a self-organized system. Since there can be several developmental paths within a specific field, high-level readiness in basic research studies may signal high application potential, while emerging trends may represent potential opportunities to expand our knowledge in these areas. This study aims to assess the feasibility of bridging from research to potential opportunities by treating translational processes of scientific/technological development. [In our case, GNSs case, development entails intertwined advances in biomedical science, and materials science & technology; we use "technological" as shorthand for the R&D activity under study.]

2.2. Literature-based translational research in biomedical fields

Proposed by NIH in 2003, translational (*i.e.* "bench to bedside") research aims to connect basic research in laboratories with clinical applications. With abundant online resources, translational research has also become a popular topic in scientometric and social science fields. Text mining tools have facilitated these scientometric studies by enabling named entity recognition (NER), pathway extraction and reasoning, gene function prediction, and so on (Gonzalez et al., 2016). Efforts have also been made to measure and trace translational progress in biomedical research using automated literature analysis. Research at a journal level has used predefined word lists to classify biomedical research journals from "basic" to "clinical" types (Lewison and Paraje, 2004). But this is not enough to obtain detailed knowledge about research content. To go further, literature content or types should be considered. Cambrosio et al. mapped translational cancer research by clustering high frequency terms (Cambrosio et al., 2006), while Venditto and Szoka added citation and clinical trials data to support portraying the translational research process of cancer nanomedicine (Venditto and Szoka, 2013). Indexed information, such as Medical Subject Headings (MeSH) terms, also provides effective topical content. Weber used MeSH terms to build a triangle framework to track research publication changes over time from three dimensions – "animal," "cell," and "human" (Weber, 2013). And Leydesdorff et al. (2012) used three groups – "C", "D", "E" – of MeSH terms to describe changes of research focus over time.

Most of these previous studies assessed translational processes using predefined markers or annotated data, like MeSH terms, and they focused more on the pathway of research development. But on some micro level, translational indicators should be domain-specific. And besides assessing translational readiness, it is important to look forward. This is the reason for this study seeking to classify publications using a supervised classification method, and to combine translational stage analysis with technology opportunities identification.

3. Data and methods

The main goal in conducting this analysis was to test the feasibility of our tech mining approach to identify "process markers" of clinical translation and potential research opportunities in published research literature. We applied this methodology to biomedical studies of GNSs. We used a hybrid lexical query to extract relevant records and metadata from MEDLINE, *via* the PubMed interface.

GNSs were selected as the focus of this study because they represent one of the most popular and diverse nanotechnologies that have been applied in biomedical research (cancer research in particular). They present a rapidly developing research domain with many potential applications in diagnosis and treatment of human diseases (Alkilany

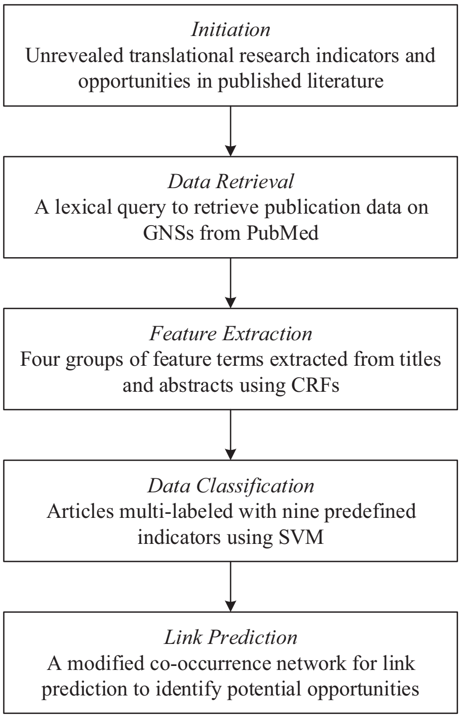
Initiation
Unrevealed translational research indicators and opportunities in published literature
Data Retrieval
A lexical query to retrieve publication data on GNSs from PubMed
Feature Extraction
Four groups of feature terms extracted from titles and abstracts using CRFs
Data Classification
Articles multi-labeled with nine predefined indicators using SVM
Link Prediction
A modified co-occurrence network for link prediction to identify potential opportunities

Fig. 1. Tech mining approach applied to identify "process markers" of clinical translation and opportunities in published research literature on GNSs.

et al., 2013; Janib et al., 2010). GNSs are defined as structures that are made of gold and range between 1 nm (molecular scale) and 100 nm in at least one dimension. Gold nanostructures have many favorable properties – such as surface chemistry, localized surface plasmon resonance (LSPR), and morphology – that enable a wide variety of applications. Presently, over 20,000 publications on this topic are indexed in MEDLINE, and about 200 new articles are published every week. Most of these studies describe potential applications in biomedical research and care.

Our analysis contains three major elements (Fig. 1): (i) extraction of grouped feature terms from titles and abstracts using a Conditional Random Fields (CRFs) model; (ii) classification of research articles by translational stage and application markers; and (iii) network construction using feature terms and link prediction for potential research opportunities. Details of these three steps will be described in the following sections.

3.1. Data

In this study, original GNSs research articles published from 2001 to 2015 (last retrieved in Oct. 2015) were retrieved from PubMed using a lexical query targeting keywords in titles and abstracts and MeSH terms. After several iterations followed by reviews, the final query was constructed and evaluated (see Suppl. Table 1). Several rules were applied to refine the query results and generate the final dataset of 23,982 articles.

1. Only original research records were retrieved, while reviews, comments, letters, and other types were excluded. Reviews not tagged as such in PubMed were identified (using a search term "review" in abstracts and titles);
2. Only articles written in English were selected;
3. Records with abstracts containing fewer than 3 sentences or 50 words were excluded.

3.2. Feature extraction and term grouping

Feature extraction is a dimensionality reduction tool that facilitates further analysis by transforming unstructured abstracts and titles into derived values. It is often followed by a series of cleaning steps – *e.g.*, combining frequently co-occurring words into clumps (Zhang et al., 2014a). In this study, in addition to extraction of key words/phrases from the corpus for data classification, we also analyzed links between different types of terms (*e.g.*, between specific nanostructures with given drugs or diseases).

In addition to named entities such as gene and cell types, we extracted multiple "feature terms" from titles and abstracts. After manually reviewing hundreds of GNSs articles in our dataset, we divided feature terms into four groups: "C – Chemicals," "G – Gene," "E – Experiment," and "O – Organism" (Table 1). These four groups roughly corresponded to three main MeSH system categories, "C – Diseases," "D – Chemicals and Drugs," and "E – Analytical, Diagnostic and Therapeutic Techniques and Equipment" (Leydesdorff et al., 2012). However, gene and protein related terms were removed from group C and taken to form a new group, G. Additionally, disease-related feature terms were added to group O, because they are semantically related to terms associated with organs and cells. In addition to named entities, common terms were also extracted.

To achieve feature extraction, we introduced a Conditional Random Fields (CRFs) model, which was conducted in CRF + + (http://taku910.github.io/crfpp/). As a discriminative model, CRFs provide a viable solution to label bias issues and modeling of statistical dependency of sequence data (Lafferty et al., 2001). CRFs provide good data fitting and are suited to including overlapping features (Sutton and McCallum, 2012; Yan and Zhu, 2015). As such, CRFs have been reported successfully applied to Named Entity Recognition (NER) and relation extraction tasks, especially in biomedical fields (Bundschus et al., 2008; McCallum and Li, 2003; Settles, 2005).

In CRFs modeling, sentences are treated as a random input variable $X = (X_1, X_2, \ldots, X_n)$ to be labeled and the corresponding label sequences $Y = (Y_1, Y_2, \ldots, Y_n)$ obey the Markov property (Lafferty et al., 2001). CRFs define the conditional probability of a label sequence y, given an input sequence x, to be

$$p(y \mid x, \lambda) = \frac{1}{Z_{(x)}} \exp\left(\sum_{i=1}^{n} \sum_{j} \lambda_j f_j (y_{i-1}, y_i, x, i)\right),$$

where $Z_{(x)}$ is a normalization factor over all label sequences; $f_j(y_{i-1}, y_i, x, i)$ is an arbitrary feature function over its arguments; and λ_j is a learned weight for each feature function. For example,

$$f_j (y_{i-1}, y_i, x, i) = \begin{cases} b(x, i), \text{ if } y_{i-1} = \mathrm{O},\ y_i = \mathrm{B} - \mathrm{G} \\ 0,\text{else} \end{cases},$$

Table 1
Grouping strategy of feature terms.

Group	Description	Examples
C	Chemicals and drugs related terms, including nanostructures.	HAuCl4; self-assembled monolayers; gold nanoparticles; PEG
E	Experimental, methodological, analytical, and equipment related terms.	Cytotoxicity; drug delivery; transmission electron microscopy; *in vivo*
G	Gene and protein related terms.	Human immunoglobulin; DNA; bovine serum albumin; antibody
O	Organism, human, animal and cell related terms, including diseases.	Mouse; human serum; HeLa cancer cells; breast cancer

where

$$b(x, i) = \begin{cases} 1, \text{ if the suffix of } x_i \text{ is ase} \\ 0,\text{else} \end{cases}.$$

Feature functions are defined with different meanings, positions and other features of words. In this study, the following model features are used: word, stem, prefix and suffix, POS tag, stop word, capitalization, phrase, and their combinations. A pre-annotated keyword list with the most common nouns in the GNSs corpus (dataset) was also used as a model feature. For example, the word “gold” was pre-tagged in group C, while “microscopy” was put in group E.

3.3. *Data classification using predefined indicators*

One key task of this study involved tracing translational tracks based on research publications. Four questions were asked to select translational research indicators. The first question was used to restrict the research area to biomedical sciences, since GNSs can also be developed for environmental detection and other applications. Other questions were related to disease, application, and translational stage of the technology described in publications:

1. Is this study related to biomedical research?
2. Does it focus on some specific disease or treatment, especially cancer?
3. What is its target application – detection, treatment, or imaging?
4. Which translational stage does this study address – physicochemical characterization (Stage-1), *in vitro* (Stage-2), *in vivo* (Stage-3), or clinical and human subjects (Stage-4)?

As with most biomedical text resources, the GNSs records analyzed in this study warrant a multidimensional framework. This means that each article could be labeled with more than one tag. For example, one article could be tagged as related to biomedicine, cancer, a specific clinical application, and a specific translational stage. To identify individual translational pathways within our dataset of 23,982 PubMed records, we first classified these records across several dimensions as shown in Fig. 2.

After comparing results from several common classification models, such as K-nearest neighbors (KNN), decision tree, and support vector machine (SVM), we chose SVM (Cortes and Vapnik, 1995) based on its performance characteristics and ease of use. SVM is a popular supervised modeling technique used for data classification. It is a probabilistic, binary, linear classifier using the maximum-margin hyperplane (Fig. 3), and it can efficiently perform a non-linear classification using the kernel trick, implicitly mapping inputs into high-dimensional feature spaces.

In this study, feature terms for classification models were selected using Chi square; and an SVM model using a Gaussian kernel was performed in R using the kernlab package (https://CRAN.R-project.org/package=kernlab). A randomly selected sample set of records was manually annotated by two NIH coauthors for training and testing.

3.4. *Identifying potential research opportunities using link prediction*

We assumed that the development of some specific domain can be abstracted as an evolutionary network that is constructed with different research and technological elements. For biomedical research, the network develops when potential effects or links between drugs or nanostructures with certain diseases are identified and verified. That is, the network should not be connected randomly, and to some extent, we should be able to predict it.

Link prediction is often used to identify missing and future links from a certain network. Including common neighbors (Adamic and Adar, 2003; Lu and Zhou, 2010), Katz (1953), and random walk (Liu and Lü, 2010; Lu and Zhou, 2010), several link prediction indexes have been developed for calculating proximities between different nodes in networks (Table 2). For these indexes, node pairs with higher proximity

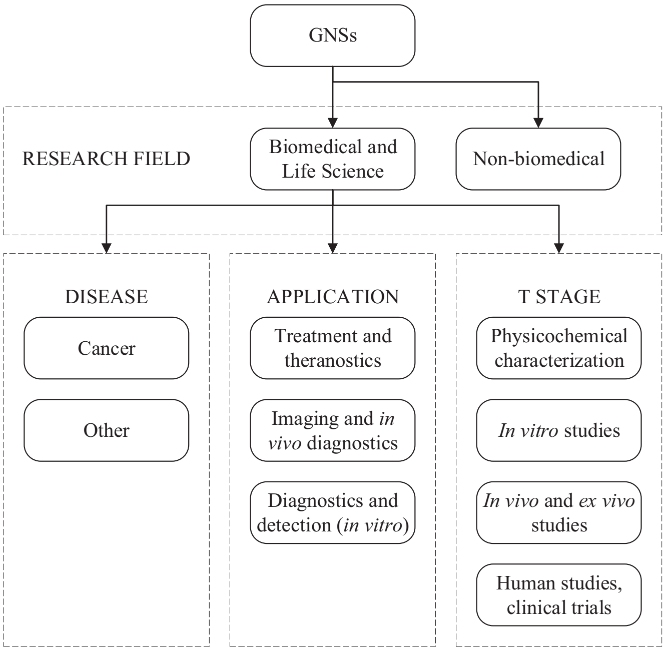
GNSs
RESEARCH FIELD
Biomedical and Life Science
Non-biomedical
DISEASE
Cancer
Other
APPLICATION
Treatment and theranostics
Imaging and *in vivo* diagnostics
Diagnostics and detection (*in vitro*)
T STAGE
Physicochemical characterization
In vitro studies
In vivo and *ex vivo* studies
Human studies, clinical trials

Fig. 2. Classification framework. [Articles could be classified based on the following characteristics: research field, disease, application and T-stage; each article may have more than one classification tag.]

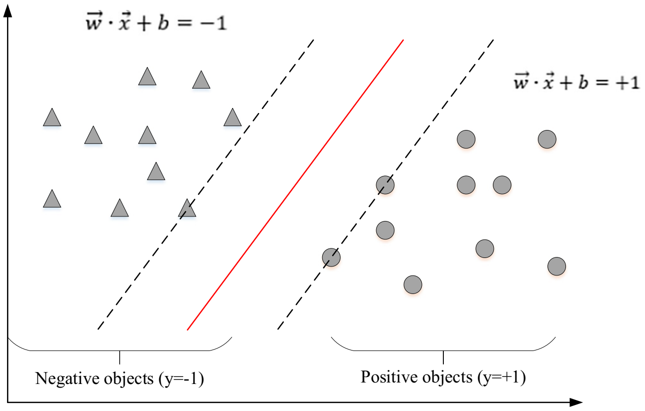
$\vec{w} \cdot \vec{x} + b = -1$
$\vec{w} \cdot \vec{x} + b = +1$
Negative objects (y=-1)
Positive objects (y=+1)

Fig. 3. Maximum-margin hyperplane for an SVM. [Samples on the margin are called the support vectors.]

Table 2
Four common examples of link prediction indexes. [Γ(*x*) consists of all neighbors of node *x*.]

Index	Proximity score
Common neighbors (CN)	${S_{xy}}^{CN} = \lvert\Gamma(x) \cap \Gamma(y)\rvert$
Jaccard	$S_{xy}^{Jaccard} = \frac{\lvert\Gamma(x) \cap \Gamma(y)\rvert}{\lvert\Gamma(x) \cup \Gamma(y)\rvert}$
Adamic-Adar (AA)	$S_{xy}^{AA} = \sum_{z \in \Gamma(x) \cap \Gamma(y)} \frac{1}{\log k_z}$ k_z is the degree of z.
Katz	$S_{xy}^{Katz} = \sum_{l=1}^{\infty} \beta^l \bullet \lvert paths_{xy}^{\langle l \rangle}\rvert = \beta A + \beta^2 A^2 + \beta^3 A^3 + \ldots$ ${paths_{xy}}^{\langle l \rangle}$ is the set of all paths with length l connecting x and y. β controls the path weights and must be lower than the reciprocal of the largest eigenvalue of adjacency matrix A.

are more likely to be connected. Common relations that have been predicted using link prediction include: collaboration, knowledge evolution, and citation networks (Choudhury and Uddin, 2016; Getoor and Diehl, 2005; Guns and Rousseau, 2014; Oakleaf, 2009).

In this study, four groups of feature terms in titles were used to represent research and technological elements in the network. Co-occurrence generated links between different nodes. To be more descriptive, only links between different feature groups were considered, such as “anticancer drug (C) – cytotoxicity (E)” and “gold nanorods (C) – colon cancer (O).” We first tested predictability of the co-occurrence network using data from three intervals for training, and new links in 2015 for testing. AUC was used (Liu and Lü, 2010) to evaluate the performance of each index. That can be interpreted as the probability that a randomly selected new link in 2015 is given a higher score of similarity than a randomly selected nonexistent link.

$$AUC = \frac{n_1 + 0.5n_2}{n}.$$

Among n independent comparisons, there are n_1 times that the link from 2015 has a higher score and n_2 times with the same score. If a network is randomly connected, AUC should be around 0.5. How much AUC exceeds 0.5 indicates to what extent an index can be used to predict missing links in a network.

Secondly, the index with the best performance was used to predict potential links. Alternative links with high scores were selected and interpreted as latent opportunities and topics.

4. Results

4.1. Feature terms identified in GNSs records

Randomly selected records (500) were manually annotated for training and testing. The annotated records were double-checked by two NIH co-authors. We tried eight different CRFs models by using varying model features from words only to all features. A 10-fold validation was used, which means for each model, we did ten experiments using 450 records as the training set and the other 50 records as test, each. Model performance was evaluated with precision, recall and F value. A prediction was considered true positive only when the group label and term boundary were both predicted. The best performance (Table 3) was observed when all model features (word, stem, prefix and suffix, POS tag, stop word, capitalization, phrase, keyword, and their combinations) were included with an average F value of 76%. Group C-Chemicals and E-Experiment had better F values, and these two groups also had the most feature terms both in the training set and the full GNSs corpus.

The final model was then trained with these 500 sample records and used to extract feature terms from the other 23,482 records. A primary cleaning step was carried out to combine single and plural forms and abbreviations. As listed in Table 3, in total, 91,073 feature terms were identified in the GNSs corpus, including 40,021 group C – Chemical terms, 40,797 E – Experiment terms, 4905 G – Gene terms, and 5350 O – Organism terms. Although groups G and O each contain relatively few terms (~5% and ~6%, respectively, of all feature terms), they cover 41% and 35% records in GNSs corpus, respectively. This result also reveals a heavy bias toward non-biomedical records in our GNSs dataset. Since biomedical research usually includes some use of biological materials – cells, animals, and genes, these articles may have feature terms that belong to group O and/or G. Therefore, the relatively low coverage of G and O in our corpus indicates that many GNSs studies included in PubMed do not focus on the development of biomedical applications.

Table 3
Feature extraction result using CRFs.

Group	Precision	Recall	F value	No. of extracted terms
All terms	81%	72%	76%	91,073
C	82%	74%	78%	40,021
E	79%	72%	75%	40,797
G	84%	63%	72%	4905
O	82%	66%	73%	5350

Table 4
Examples of annotated sentences using CRFs.

#	Sentences from titles/abstracts
1	The effect of **carboxylic acid** (**C**) functionality present in **polymer backbone** (**C**) is reported on **electrochemical sensing** (**E**) of **dopamine** (**C**).
2	This technique could be useful in **cancer** (**O**) **treatment** (**E**) if a cancer-specific **antibody** (**G**) is used to localize **gold nanoparticles** (**C**) to **malignant cells** (**O**).
3	**68Ga-labeled gold glyconanoparticles** (**C**) for exploring **blood-brain barrier permeability** (**E**): **preparation** (**E**), **biodistribution** (**E**) studies, and improved **brain uptake** (**E**) via **neuropeptide conjugation** (**E**).
4	As a proof of principle, a **biodistribution** (**E**) study in **rats** (**O**) is performed for the different (68)**Ga-GNPs** (**C**).

CRFs allowed us to annotate each sentence in the GNSs dataset – *i.e.*, add C, E, G and O tags. Table 4 shows four examples of annotated sentences. Most feature terms were extracted correctly with clear boundaries and meaning, although we found most false cases were due to the incompleteness of annotation. For example, in the second example, “cancer-specific antibody” would be more accurate, but the model only identified “antibody.”

We then compared the results obtained using CRFs and term clumping. Even after a complete cleaning process, the term clumping approach resulted in many meaningless, but high frequency, terms such as “first time” and “experimental results,” as well as several terms that only appeared once in the corpus. Feature terms extracted using CRFs were more high frequency and could be classified with less noise.

4.2. Translational research development from different dimensions

Based on Fig. 2, we chose nine tags for record classification (Table 5). We gave clear boundaries for all tags as far as possible. However, articles can be multi-tagged, for example, “gold nanoshell serving as both CT contrast agents and photoabsorbers for photothermal therapy” (Ke et al., 2014); so this article should be labeled both “Imaging” and “Treatment.” To develop training and test data, in this stage, we expanded our training set to 1000 records by sampling another 500 articles from the corpus. Each record from the training set was manually assigned one or more tags based on their titles, abstracts and MeSH terms, if MeSH indexed. Some articles were not annotated with any of these nine tags. Two of our authors from NIH manually indexed the sample data individually, and then combined their results together by double-checking inconsistent cases. Some dependency roles were also applied, for example, records annotated with “Treatment” should also be in the “Biomed” group.

To select appropriate feature terms for SVM models, we tried different group sizes of feature terms with high Chi square scores from 200 to 700. The step size was 50. Models with the best performance are listed in Table 5. For example, there are 610 tagged biomedical records in the 1000 sample records. The best performance of SVM model for Biomed group was observed when using the top 550 feature terms. Also, two decision tree models (one used the same feature terms with SVM – DT; one used only the top 200 feature terms – DT*) and a KNN model were conducted to compare results. All performance results were tested by 10-fold cross validation. The best results were observed in SVM with all F values above 80% and three of them were over 90%. Precision performed relatively better than recall.

It's rational to apply this model to the other 22,982 records. The classification result was not perfectly accurate, but it provides valuable

Table 5
Classification results using different models.

Group	No. of records	Model	No. of terms	Precision	Recall	F1	Accuracy
Biomed	610	SVM	550	96%	90%	93%	92%
		DT	550	94%	80%	87%	85%
		DT*	200	93%	80%	86%	85%
		KNN	200	92%	70%	80%	78%
Cancer	143	SVM	550	94%	93%	93%	98%
		DT	550	91%	79%	85%	96%
		DT*	200	92%	79%	85%	96%
		KNN	200	94%	23%	37%	89%
Detection	327	SVM	350	91%	81%	86%	91%
		DT	350	85%	78%	82%	89%
		DT*	200	85%	78%	82%	89%
		KNN	200	90%	53%	67%	83%
Treatment	162	SVM	650	92%	91%	91%	97%
		DT	650	91%	69%	79%	94%
		DT*	200	91%	69%	79%	94%
		KNN	200	94%	30%	45%	88%
Imaging	105	SVM	350	90%	86%	88%	98%
		DT	350	82%	52%	64%	94%
		DT*	200	82%	52%	64%	94%
		KNN	200	89%	15%	26%	91%
Stage-1	204	SVM	650	88%	73%	80%	92%
		DT	650	70%	42%	52%	85%
		DT*	200	70%	42%	52%	85%
		KNN	200	67%	15%	24%	81%
Stage-2	264	SVM	300	84%	86%	85%	92%
		DT	300	77%	78%	77%	88%
		DT*	200	77%	78%	77%	88%
		KNN	200	90%	44%	59%	84%
Stage-3	77	SVM	250	89%	88%	89%	98%
		DT	250	76%	70%	73%	96%
		DT*	200	76%	70%	73%	96%
		KNN	200	100%	10%	19%	93%
Stage-4	59	SVM	200	94%	80%	86%	99%
		DT	200	73%	49%	59%	96%
		DT*	–	–	–	–	–
		KNN	200	–	0%	–	94%

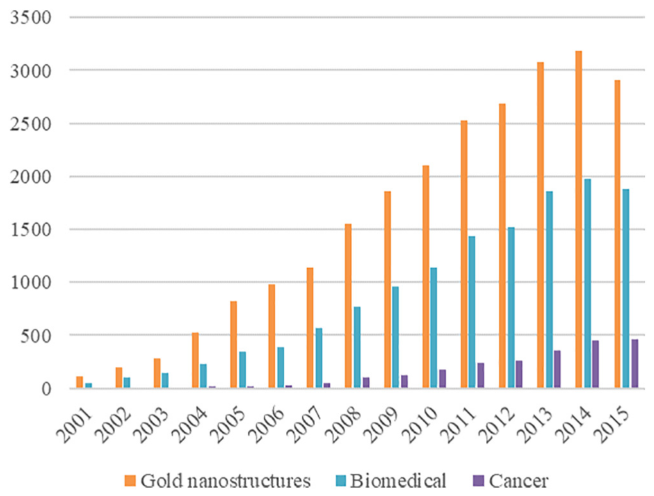

Fig. 4. Publication trends for biomedical and cancer subsets. [Note: Biomedical research accounts for over 55% of GNSs publications.]

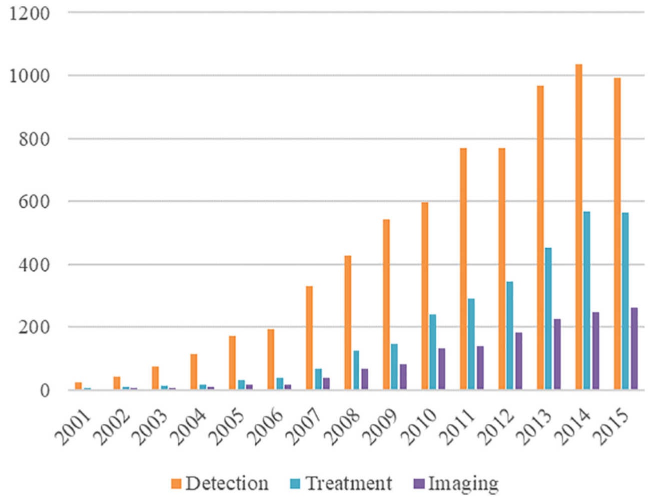

Fig. 5. Publication trends for applications.

clues on how GNSs research is distributed across various applications and topics (Table 6). There were 13,403 biomedical articles and 2314 cancer related articles. *In vitro* detection had the most records, followed by treatment and, then, imaging. Among translational stages, Stage-1 and Stage-2 had the most articles.

Figs. 4 to 6 display publication trends of different subsets based on the classification results. The number of GNSs publications in PubMed has grown dramatically from 115 in 2001 to 3185 in 2014 (data were last updated in Oct. 2015; 2015 data are therefore incomplete). Biomedical research accounts for over 55% of GNSs publications and this ratio has been steadily increasing. Other research directions of GNSs include environmental, agricultural and catalytic applications. Cancer related research emerged in 2004 or so and began to occupy a significant proportion, emerging as a vital domain in GNSs biomedical research in 2008–2009 (Fig. 4).

Due to their remarkable electrical and optical properties, gold nanostructures have been widely developed for biosensor, protein detection, and other detection uses. The number of publications related to Detection rose quickly in the early 2000's and remained a leading GNSs trend for over 15 years. Intriguingly, treatment has been developing rapidly, especially since 2010, in chemotherapy and photothermal therapy, and its gap with detection has been narrowing. Compared to detection and treatment, the upward trend of imaging is relatively moderate.

As for translational stage, Stage-1 (physicochemical characterization) accounted for a bigger proportion of GNSs research before 2007. After that, Stage-2, *in vitro* studies, has been on a sharp upward trend,

Table 6
Indicators for data classification and classification results.

Dimension	Indicator (tag)	Description	No. of articles (%)
Research area	Biomedical	Biomedical research; clinical application; biology, basic science; bioremediation research	13,403 (56%)
Disease	Cancer	Cancer, neoplasm	2314 (10%)
Application	Detection	Diagnostics and detection; sensors (*in vitro*); disease processes; drug discovery	7068 (29%)
	Treatment	Therapy and theranostics; tissue regeneration; *in vivo* biomedical applications; implantations	2921 (12%)
	Imaging	Imaging and *in vivo* diagnostics; image guided therapy	1432 (6%)
T stage	Stage-1	Physicochemical characterization	3077 (13%)
	Stage-2	*In vitro* – activity; compatibility; availability; immunology	5364 (22%)
	Stage-3	*In vivo* – activity; toxicity; biocompatibility; efficacy and safety; *ex vivo* activity	1105 (5%)
	Stage-4	Human subjects – human testing; clinical samples	713 (3%)

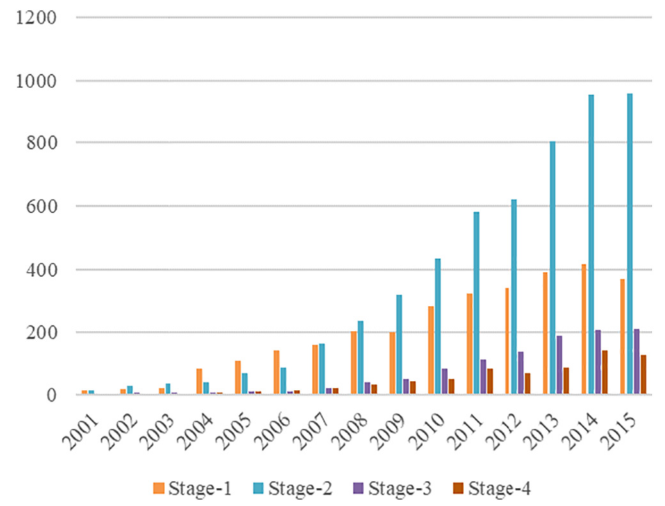

Fig. 6. Publication trends for translational stages.

and has become dominant since then, while Stage-1 has been slowing recently. *In vivo* studies (Stage-3) have also been increasing at a mild pace. We found only a small amount of Stage-4 publications in our dataset, and most of them belong to detection related studies using human samples (human serum, urine samples), with no clinical trials for new therapies or drugs. Notably, not all applications covered all these stages. For example, Stage-3 (*in vivo*), was barely detected among detection studies, as expected, based on the nature of these applications, since most detection applications like biosensor, protein detection were not carried out *in vivo*.

As seen in Table 7, detection applications, such as immunoassays, ECL biosensors and cancer diagnosis, had higher translational readiness than treatment and imaging. For example, specific *in vitro* detection targets, including carcinoembryonic antigen, carbohydrate antigen, glucose oxidase and IgG protein, were reported in human subject studies (Stage-4). Treatment related research had achieved animal experiments (Stage-3) with little further progress. Combining term clustering results of high frequency term lists for each cell of Table 7, we found that after 2010, detection application barely had new topics or terms, indicating that research into these applications was relatively mature. Recent "hot" topics included brain delivery and photothermal therapy, indicating great potential in these directions. Like treatment related applications, imaging research also included reports of *in vivo* animal experiments. Notably, in addition to cancer imaging, neurodegenerative diseases had become emerging targets for GNSs imaging applications. In general, the detection focused area displayed more potential for practical application and commercialization, while treatment and imaging studies appeared to still remain "emerging topics" in preclinical studies, which may point toward additional research opportunities.

4.3. Potential opportunities in GNSs cancer treatment research

Our analysis demonstrates that cancer has become one of the most important areas of GNSs research. Specifically, cancer treatment has emerged as a fast-developing application in recent years. To examine this trend further, we built a network using the method described in Section 3.4 with 1598 cancer treatment records.

Table 7
Publication distribution in two dimensions [each record can be multi-labeled or labeled only in one dimension; the total number of each row or column does not correspond to Table 2].

	Detection	Treatment	Imaging
Stage-1	604	607	306
Stage-2	2188	2409	999
Stage-3	215	878	620
Stage-4	665	61	16

Table 8
AUC scores using 4 link prediction algorithms.

Training set	No. of records	CN	Jaccard	AA	Katz
2001–2014	1265	72.77%	69.54%	72.96%	86.87%
2010–2014	1052	72.88%	70.26%	73.21%	86.21%
2014	327	66.13%	64.57%	66.14%	75.89%

To that end, we used new links identified in 2015 as test sets, while other intervals (2014, 2010–2014 and 2001–2014) were used as training sets (Table 8). We tried 4 different algorithms to assess the predictability of the network. For these different algorithms, indexes based on common neighbors (CN, Jaccard and AA) offered relatively poor performance, while Katz, considering all paths between two nodes, showed the best AUC scores. This result implies that several latent patterns exist between different elements of this network. These patterns not only depend on common neighbors between them, but also rely on some specific paths indicating medical research principles and development in the GNSs cancer treatment area.

The results also reveal that a medium time interval around 5 years offers a more stable link prediction performance. For 2015, its research activities were more connected with recent years. However, when we only use 2014 data, there was too much uncertainty. A bigger training set did not lead to better performance. Links from early years may have been eliminated later and disappear from more recent studies. When more records from early years are introduced into the training set, the proximity indexes based on common neighbors go down, while only Katz results increase.

Since the records from 2001 to 2009 did not improve model performance significantly, we used records of more recent R&D activities (2010–2015) as the final data to predict potential links. The Katz index was applied to predict unknown links between different groups of terms using GNSs cancer treatment records from 2010 to 2015. Proximity scores between each unconnected node pair were sorted in descending order. Theoretically, a node pair with high proximity score is more likely to get connected but analyzing these potential links one by one is not feasible. So, to explore potential emerging terms, we list the high-frequency terms found in 500 node pairs with the highest proximity scores (Table 9).

Table 9 shows high-frequency terms that remain highly connected with *in vitro* studies, including human cell, MCF-7 breast cancer cell, and osteosarcoma cell. This shows that most cancer treatment-focused studies will still be in Stage 2 recently. Although most of these studies involved nanorods and nanoparticles, we also identified several emerging nanostructures, such as nanoprisms and nanobeacons, which first appeared in the cancer treatment context in 2013 and 2015, respectively. The numbers of nanoprism and nanobeacon related records were still low, but they exhibited a consistent upward trend. Other potential emerging topics include photothermal treatment, hyaluronic acid, and lectin. There are also other promising indications not listed in Table 9 that include gene therapy related terms, such as antisense DNA and anticancer drugs (like 5-fluorouracil). These results may be useful in informing R&D activities and improving research efficiency regarding switching research focuses or exploring novel materials or targets barely detected.

5. Discussion

In this study, we explored the possibility of using tech mining to better understand and describe translational readiness of GNS-based nanotechnologies and to identify related emerging topics in biomedical

Table 9
High-frequency terms in link prediction results (2010–2015).

Feature term	Frequency
Human cell (O)	89
Radiographically dense mammary tissue (O)	65
MCF-7 breast cancer cell (O)	41
Hyaluronic acid-fabricated nanogold (C)	33
Infrared triggered photodynamic therapy (E)	32
Polyethyleneimine (C)	30
Anticancer efficacy (E)	23
Partial inhibition (E)	22
Gold nanoprisms (C)	21
Vasculature damage (O)	18
Anticancer efficiency (E)	17
Dendrimers (E)	14
Colon cancer cell growth (O)	13
Vitro administration (E)	13
Covalently-coupled bombesin peptide (G)	12
Drug-treated cancer cell nucleus (O)	12
Lectin (G)	12
Osteosarcoma cell (O)	11
Antibody nanoparticle (C)	11
Photothermal treatment (E)	11
Gold nanobeacons (C)	11
Gold nanoshell-decorated silicone (C)	11

research. To assess research readiness of specific technologies, we classified publications using a predefined translational indicator framework. To achieve the first goal, we compared different classification models using dynamic feature terms. We explored the feasibility of using Conditional Random Fields (CRFs) and network analysis to delineate future trends in GNS-related research. One of the most important reasons to use CRFs for feature extraction is that tagging feature terms can be very helpful for network construction using textual connections. The predictability of the network for GNSs cancer treatment was verified and potential opportunities were screened in the form of missing links. This quantitative analysis may provide researchers with future possibilities for more efficient R&D activities. Tech mining approaches such as the one proposed in this study can complement traditional literature reviews of target research landscapes.

Classification analysis results revealed that most GNSs biomedical studies are focused on early stages of technology development. This highlights the need for more advanced data curation systems and open computational platforms to enable meta-analyses and avoid duplicate efforts. The analysis also revealed limitations of strictly literature-based efforts lacking streamlined access to standardized GNS-related data. Further analysis showed that most application-ready studies focused on detection, including immunoassays, DNA detection, and biosensors. More recently, treatment applications have been gaining traction, predominantly in areas related to cancer research.

To the best of our knowledge, this study offers the first computational assessment of the GNS-based research landscape to identify the most promising nanotechnologies. To predict emerging topics, we used a link prediction approach, which develops a self-organized pattern of a research network. The link prediction analysis confirmed the predictability of GNSs research network. Some technology elements, including nanoprism and nanobeacon, photothermal treatment, and gene therapy, were identified as potential hot topics.

Such translational processes exist not only in the biomedical domain, but also in many other technology fields. A transplantation of such analyses to other topics could be fruitful – *e.g.*, in exploring technology transfer progression from R&D publication and patent information resources.

This study is not without inherent limitations. At this stage, published literature is the most abundant source of information about nanomedical research. However, to study more advanced technologies effectively, we may need to add multi-sourced data from patents, clinical trials and medical records. Furthermore, even the most thorough analysis of publications will not provide a comprehensive picture of translational innovation pathways and bottlenecks for any research field. Our goal was to develop a methodology for rapid assessment of translational innovation and to evaluate it in a rapidly developing and growing area of biomedical research. As tech mining tools, data repositories, and consensus conceptual models continue to mature, new opportunities will present themselves to improve and integrate the approach described here with other resources.

As for link prediction, its major limitation is that it can only predict potential linkages based on present nodes. One potential solution for this problem is to apply link prediction to a bigger training network. Another limitation is that feature term annotation is a very labor-intensive process and more manual term selection should reveal additional emerging areas of research.

More studies need to be conducted to confirm our results about the feasibility of using tech mining to explore research translation to innovation.

6. Conclusions

In this study, we have developed a methodology to assess translational readiness and to identify potential opportunities in biomedical GNSs research using tech mining approaches. We have used a combination of multi-sourced tools to generate more diversified results. These tools were used to develop a pipeline to explore innovation pathways of biomedical research from a translational point of view. Nine indicators were generated to locate different GNSs research activities on a progression toward clinical use. GNSs research has demonstrated upward trends in the areas of cancer research, therapeutic applications, and *in vitro* studies. Treatment and imaging applications attracted more research efforts recently with many emerging topics, such as photothermal treatment and gene therapy, that may lead to new opportunities. We believe such analyses can be useful for researchers, funding agencies, and even pharmaceutical and biotech companies involved in relevant research activities and research planning.

Supplementary data to this article can be found online at https://doi.org/10.1016/j.techfore.2018.08.002.

Acknowledgments

This paper is a revised and expanded version of a paper entitled “Delineating Translational Innovation Pathways for Nanomedical Research Using Tech Mining” presented at the PICMET '17 Conference – Technology Management for Interconnected World, Portland, Oregon, USA, 9–13 July 2017. This research was partially supported by the Innovation Co-lab from Beijing Institute of Technology, China, Georgia Tech, United States and the University of Manchester, United Kingdom. We also acknowledge support from the US National Science Foundation, United States (NSF) (#1064146), the Natural Science Foundation of Shenzhen University, China (SZU) (#2017013). The findings and observations contained in this paper are those of the authors and do not necessarily reflect the views of US NSF, SZU. *VantagePoint* software [www.theVantagePoint.com] and R were used in this study for text analyses of MEDLINE abstract records.

References

Adamic, L.A., Adar, E., 2003. Friends and neighbors on the Web. Soc. Networks 25, 211–230. https://doi.org/10.1016/S0378-8733(03)00009-1.

Alkilany, A.M., Lohse, S.E., Murphy, C.J., 2013. The gold standard: gold nanoparticle libraries to understand the nano-bio interface. Acc. Chem. Res. 46, 650–661. https://doi.org/10.1021/ar300015b.

Boon, B., Park, Y., 2005. A systematic approach for identifying technology opportunities: keyword-based morphology analysis. Technol. Forecast. Soc. Chang. 72, 145–160. https://doi.org/10.1016/j.techfore.2004.08.011.

Bundschus, M., Dejori, M., Stetter, M., Tresp, V., Kriegel, H.-P., 2008. Extraction of semantic biomedical relations from text using conditional random fields. BMC Bioinf. 9, 207. https://doi.org/10.1186/1471-2105-9-207.

Cambrosio, A., Keating, P., Mercier, S., Lewison, G., Mogoutov, A., 2006. Mapping the emergence and development of translational cancer research. Eur. J. Cancer 42, 3140–3148. https://doi.org/10.1016/j.ejca.2006.07.020.

Choi, S., Kim, H., Yoon, J., Kim, K., Lee, J.Y., 2013. An SAO-based text-mining approach for technology roadmapping using patent information. R D Manag. 43, 52–74. https://doi.org/10.1111/j.1467-9310.2012.00702.x.

Choudhury, N., Uddin, S., 2016. Time-aware link prediction to explore network effects on temporal knowledge evolution. Scientometrics 108, 745–776. https://doi.org/10.1007/s11192-016-2003-5.

Cortes, C., Vapnik, V., 1995. Support-vector networks. Mach. Learn. 20, 273–297. https://doi.org/10.1023/A:1022627411411.

De Miranda Santo, M., Coelho, G.M., dos Santos, D.M., Filho, L.F., 2006. Text mining as a valuable tool in foresight exercises: a study on nanotechnology. Technol. Forecast. Soc. Chang. 73, 1013–1027. https://doi.org/10.1016/j.techfore.2006.05.020.

Dickherber, A., Morris, S.A., Grodzinski, P., 2015. NCI investment in nanotechnology: achievements and challenges for the future. Wiley Interdiscip. Rev. Nanomed. Nanobiotechnol. 7, 251–265. https://doi.org/10.1002/wnan.1318.

Etheridge, M.L., Campbell, S.A., Erdman, A.G., Haynes, C.L., Wolf, S.M., McCullough, J., 2013. The big picture on nanomedicine: the state of investigational and approved nanomedicine products. Nanomedicine. https://doi.org/10.1016/j.nano.2012.05.013.

Getoor, L., Diehl, C.P., 2005. Link mining: a survey. In: ACM SIGKDD Explor. Newsl. 7. pp. 3–12. https://doi.org/10.1145/1117454.1117456.

Gonzalez, G.H., Tahsin, T., Goodale, B.C., Greene, A.C., Greene, C.S., 2016. Recent advances and emerging applications in text and data mining for biomedical discovery. Brief. Bioinform. 17, 33–42. https://doi.org/10.1093/bib/bbv087.

Guns, R., Rousseau, R., 2014. Recommending research collaborations using link prediction and random forest classifiers. Scientometrics 101, 1461–1473. https://doi.org/10.1007/s11192-013-1228-9.

Guo, J., Wang, X., Li, Q., Zhu, D., 2016. Subject-action-object-based morphology analysis for determining the direction of technological change. Technol. Forecast. Soc. Chang. 105, 27–40. https://doi.org/10.1016/j.techfore.2016.01.028.

Janib, S.M., Moses, A.S., MacKay, J.A., 2010. Imaging and drug delivery using theranostic nanoparticles. Adv. Drug Deliv. Rev. https://doi.org/10.1016/j.addr.2010.08.004.

Katz, L., 1953. A new status index derived from sociometric analysis. Psychometrika 18, 39–43. https://doi.org/10.1007/BF02289026.

Ke, H., Yue, X., Wang, J., Xing, S., Zhang, Q., Dai, Z., Tian, J., Wang, S., Jin, Y., 2014. Gold nanoshelled liquid perfluorocarbon nanocapsules for combined dual modal ultrasound/ct imaging and photothermal therapy of cancer. Small 10, 1220–1227. https://doi.org/10.1002/smll.201302252.

Lafferty, J., McCallum, A., Pereira, F.C.N., 2001. Conditional random fields: probabilistic models for segmenting and labeling sequence data. In: ICML '01 Proc. Eighteenth Int. Conf. Mach. Learn. 8. pp. 282–289. https://doi.org/10.1038/nprot.2006.61.

Lewison, G., Paraje, G., 2004. The classification of biomedical journals by research level. Scientometrics 60, 145–157. https://doi.org/10.1023/B:SCIE.0000027677.79173.b8.

Leydesdorff, L., Rotolo, D., Rafols, I., 2012. Bibliometric perspectives on medical innovation using the medical subject headings of PubMed. J. Am. Soc. Inf. Sci. Technol. 63, 2239–2253. https://doi.org/10.1002/asi.22715.

Liu, W., Lü, L., 2010. Link prediction based on local random walk. Europhys. Lett. 89, 58007. https://doi.org/10.1209/0295-5075/89/58007.

Lu, L., Zhou, T., 2010. Link prediction in complex networks: a survey. Phys. A 390, 1150–1170. https://doi.org/10.1016/j.physa.2010.11.027.

Ma, J., Porter, A.L., 2015. Analyzing patent topical information to identify technology pathways and potential opportunities. Scientometrics 102, 811–827. https://doi.org/10.1007/s11192-014-1392-6.

Ma, T., Porter, A.L., Guo, Y., Ready, J., Xu, C., Gao, L., 2014. A technology opportunities analysis model: applied to dye-sensitised solar cells for China. Tech. Anal. Strat. Manag. 26, 87–104. https://doi.org/10.1080/09537325.2013.850155.

McCallum, A., Li, W., 2003. Early results for named entity recognition with conditional random fields, feature induction and web-enhanced lexicons. In: Proc. seventh Conf. Nat. Lang. Learn. HLT-NAACL 2003–4, pp. 188–191. https://doi.org/10.3115/1119176.1119206.

Morris, Z.S., Wooding, S., Grant, J., 2011. The answer is 17 years, what is the question: understanding time lags in translational research. J. R. Soc. Med. 104, 510–520. https://doi.org/10.1258/jrsm.2011.110180.

Newman, N.C., Porter, A.L., Newman, D., Trumbach, C.C., Bolan, S.D., 2014. Comparing methods to extract technical content for technological intelligence. J. Eng. Technol. Manag. 32, 97–109. https://doi.org/10.1016/j.jengtecman.2013.09.001.

Noh, H., Song, Y.K., Lee, S., 2016. Identifying emerging core technologies for the future: case study of patents published by leading telecommunication organizations. Telecommun. Policy 40, 956–970. https://doi.org/10.1016/j.telpol.2016.04.003.

Oakleaf, M., 2009. Link prediction in citation networks. Commun. Inf. Lit. 3, 80–90. https://doi.org/10.1002/asi.

Ogawa, T., Kajikawa, Y., 2017. Generating novel research ideas using computational intelligence: a case study involving fuel cells and ammonia synthesis. Technol. Forecast. Soc. Chang. 120, 41–47. https://doi.org/10.1016/j.techfore.2017.04.004.

Porter, A.L., Cunningham, S.W., 2005. Tech Mining: Exploiting New Technologies for Competitive Advantage. John Wiley & Sonshttps://doi.org/10.1002/0471698466.

Porter, A.L., Detampel, M.J., 1995. Technology opportunities analysis. Technol. Forecast. Soc. Chang. 49, 237–255. https://doi.org/10.1016/0040-1625(95)00022-3.

Porter, A.L., Jin, X.-Y., Gilmour, J.E., Cunningham, S., Xu, H., Stanard, C., Wang, L., 1995. Technology opportunities analysis: integrating technology monitoring, forecasting, and assessment with strategic planning. Technol. Forecast. Soc. Chang. 49, 237–255. https://doi.org/10.1016/0040-1625(95)00022-3.

Settles, B., 2005. ABNER: an open source tool for automatically tagging genes, proteins and other entity names in text. Bioinformatics 21, 3191–3192. https://doi.org/10.1093/bioinformatics/bti475.

Song, K., Kim, K.S., Lee, S., 2015. Discovering new technology opportunities based on patents: text-mining and F-term analysis. Technovation 60–61, 1–14. https://doi.org/10.1016/j.technovation.2017.03.001.

Sutton, C., McCallum, A., 2012. An Introduction to Conditional Random Fields. 4. Found. Trends® Mach. Learn., pp. 267–373. https://doi.org/10.1561/2200000013.

Trochim, W., Kane, C., Graham, M.J., Pincus, H.A., 2011. Evaluating translational research: a process marker model. Clin. Transl. Sci. 4, 153–162. https://doi.org/10.1111/j.1752-8062.2011.00291.x.

Tseng, Y.H., Lin, Y.I., Lee, Y.Y., Hung, W.C., Lee, C.H., 2009. A comparison of methods for detecting hot topics. Scientometrics 81, 73–90. https://doi.org/10.1007/s11192-009-1885-x.

Venditto, V.J., Szoka, F.C., 2013. Cancer nanomedicines: so many papers and so few drugs!. Adv. Drug Deliv. Rev. https://doi.org/10.1016/j.addr.2012.09.038.

Weber, G.M., 2013. Identifying translational science within the triangle of biomedicine. J. Transl. Med. 11, 126. https://doi.org/10.1186/1479-5876-11-126.

Wilhelm, S., Tavares, A.J., Dai, Q., Ohta, S., Audet, J., Dvorak, H.F., Chan, W.C.W., 2016. Analysis of nanoparticle delivery to tumours. Nat. Rev. Mater. https://doi.org/10.1038/natrevmats.2016.14.

Yan, E., Zhu, Y., 2015. Identifying entities from scientific publications: a comparison of vocabulary- and model-based methods. J. Inf. Secur. 9, 455–465. https://doi.org/10.1016/j.joi.2015.04.003.

Yoon, J., Kim, K., 2011. Identifying rapidly evolving technological trends for R&D planning using SAO-based semantic patent networks. Scientometrics 88, 213–228. https://doi.org/10.1007/s11192-011-0383-0.

Yoon, J., Park, H., Kim, K., 2013. Identifying technological competition trends for R&D planning using dynamic patent maps: SAO-based content analysis. Scientometrics 94, 313–331. https://doi.org/10.1007/s11192-012-0830-6.

Zhang, Y., Guo, Y., Wang, X., Zhu, D., Porter, A.L., 2013. A hybrid visualisation model for technology roadmapping: bibliometrics, qualitative methodology and empirical study. Tech. Anal. Strat. Manag. 25, 707–724. https://doi.org/10.1080/09537325.2013.803064.

Zhang, Y., Porter, A.L., Hu, Z., Guo, Y., Newman, N.C., 2014a. Term clumping for technical intelligence: a case study on dye-sensitized solar cells. Technol. Forecast. Soc. Chang. 85, 26–39. https://doi.org/10.1016/j.techfore.2013.12.019.

Zhang, Y., Zhou, X., Porter, A.L., Vicente Gomila, J.M., 2014b. How to combine term clumping and technology roadmapping for newly emerging science & technology competitive intelligence: "problem & solution" pattern based semantic TRIZ tool and case study. Scientometrics 101, 1375–1389. https://doi.org/10.1007/s11192-014-1262-2.

Zhang, Y., Zhang, G., Chen, H., Porter, A.L., Zhu, D., Lu, J., 2016. Topic analysis and forecasting for science, technology and innovation: methodology with a case study focusing on big data research. Technol. Forecast. Soc. Chang. 105, 179–191. https://doi.org/10.1016/j.techfore.2016.01.015.

Jing Ma received a Ph.D. of Management Science and Engineering from Beijing Institute of Technology in 2017. She was a visiting scholar in Georgia Tech from 2013 to 2014. Her research interests include innovation management, scientometrics and text mining, and has published over 10 peer-reviewed articles. Presently she is an assistant professor in College of Management, Shenzhen University.

Natalie F. Abrams is a Program Director in the Division of Cancer Prevention at the US National Cancer Institute, where she manages a portfolio of cooperative agreements and grants, develops research initiatives, and conducts portfolio analysis. Before joining NCI in 2011, she was a faculty at the J. Craig Venter Institute in Rockville, Maryland, USA, where she conducted research in the areas of comparative genomics and evolutionary classification of proteins encoded in complete eukaryotic genomes. She has over 15 years of experience in bioinformatics and genomics sciences and published 50 peer-reviewed articles. Dr. Abrams received a Ph.D. in Biology/Biochemistry from the Engelhard Institute of Molecular Biology in Moscow, Russia.

Alan L. Porter is Professor Emeritus of Industrial & Systems Engineering, and of Public Policy, at Georgia Tech, where he is Co-director of the Technology Policy and Assessment Center. He is also Director of R&D for Search Technology, Inc., Norcross, GA. He is author or co-author of some 240 articles and books, including Tech Mining (Wiley, 2005) and Forecasting and Management of Technology (Wiley, 2011).

Donghua Zhu is Professor of Management Science and Engineering, at Beijing Institute of Technology, and he was a Senior Visiting Scholar at School of Public Policy, Georgia Institute of Technology. His current research focuses on data mining, technology assessment and forecasting.

Dorothy Farrell currently works at the American Association of Colleges of Pharmacy. Prior to joining AACP in 2017, she worked at the National Cancer Institute as a project manager in the Alliance for Nanotechnology in Cancer, where she managed research awards and participated in the development of new research initiatives. Dr. Farrell received her B.S. in Physics from Brooklyn College, the City University of New York, USA and her doctorate in Physics from Carnegie Mellon University, where her thesis research focused on the synthesis and characterization of self-assembled arrays of magnetic nanoparticles. She then spent two years at University College London in the UK, working on the preparation of magnetic alloy nanoparticles and nanoparticle-antibody conjugates for use in cancer therapy before returning to the U.S. to work at the Naval Research Laboratory developing biocompatible ligands for nanoparticle functionalization.