Technological Forecasting & Social Change 105 (2016) 179–191

Contents lists available at ScienceDirect

Technological Forecasting & Social Change

Topic analysis and forecasting for science, technology and innovation: Methodology with a case study focusing on big data research

Yi Zhang [a,b,*], Guangquan Zhang [a], Hongshu Chen [a,b], Alan L. Porter [c], Donghua Zhu [b], Jie Lu [a]

[a] Decision Systems & e-Service Intelligence Research Lab, Centre for Quantum Computation & Intelligent Systems, Faculty of Engineering and Information Technology, University of Technology Sydney, Australia
[b] School of Management and Economics, Beijing Institute of Technology, Beijing, PR China
[c] Technology Policy and Assessment Centre, Georgia Institute of Technology, Atlanta, USA

ARTICLE INFO

Article history:
Received 30 April 2015
Received in revised form 13 January 2016
Accepted 21 January 2016
Available online 1 February 2016

Keywords:
Topic analysis
Technological forecasting
Text mining
Text clustering
Technical intelligence

ABSTRACT

The number and extent of current Science, Technology & Innovation topics are changing all the time, and their induced accumulative innovation, or even disruptive revolution, will heavily influence the whole of society in the near future. By addressing and predicting these changes, this paper proposes an analytic method to (1) cluster associated terms and phrases to constitute meaningful technological topics and their interactions, and (2) identify changing topical emphases. Our results are carried forward to present mechanisms that forecast prospective developments using Technology Roadmapping, combining qualitative and quantitative methodologies. An empirical case study of Awards data from the United States National Science Foundation, Division of Computer and Communication Foundation, is performed to demonstrate the proposed method. The resulting knowledge may hold interest for R&D management and science policy in practice.

1. Introduction

The coming of the Big Data Age introduces big opportunities and big challenges for modern society. The focus on "data-driven", emphasizing information technology's (IT) role in leading decision making and innovation, has now evolved into both analytic and applied models (Bughin et al., 2010; McAfee et al., 2012). Meanwhile, research addressing Science, Technology, & Innovation (ST&I) activities is widening into multiple perspectives (Bengisu, 2003; Zhang et al., 2014c). Industry and national Research & Development (R&D) efforts are beginning to track these trends to compete globally. However, the number and extent of potential topics are changing all the time, and their induced accumulative innovation, or even disruptive revolution, has the ability to quickly and heavily influence much of society.

ST&I data sources, involving academic publications, patents, academic proposals, etc., provide possibilities for describing previous scientific dynamics and efforts, discovering innovation capabilities, and forecasting probable evolution trends in the near future (Porter and Detampel, 1995; Zhang et al., 2013). As a valuable instrument for ST&I analysis, text mining affords automatic techniques to explore insights into data structure and content, which helps augment and amplify the capabilities of domain experts when dealing with real-world problems (Kostoff et al., 2001). Information visualization techniques are also highly engaged in Technology Roadmapping (TRM) for R&D planning and strategic management. Current ST&I text analysis oriented toward TRM focuses on emerging technical topics via the Forecasting Innovation Pathways approach (Guo et al., 2012; Robinson et al., 2013), and the Keyword-based Patent/Knowledge Map (Yoon and Park, 2005; Lee et al., 2008; Lee et al., 2009b). Those contribute promising efforts to deal with industry-related technology assessment and forecasting tasks via both semi-automatic, bibliometric-oriented software tools and expert knowledge.

Previous studies on ST&I topic analysis and forecasting could be considered in two aspects: 1) IT techniques have been widely introduced for text clustering, but these intelligent algorithms usually concentrate on data dimensions, data scale, and cluster understanding (Beil et al., 2002), and lack the consideration to connect the stimulated experiments with real-world problems. As an example, an efficient text clustering algorithm in a simulated training set of business news would not be readily adaptable for scientific publications, since semantic structure and linguistic norms differ between the two data forms. 2) Current R&D and strategic management favor the contribution of expert knowledge, tending to shut the door on intelligent IT techniques.

We summarize concerns with recent research as follows: 1) Text clustering algorithms generally are able to obtain sound results on

[*] Corresponding author at: Decision Systems & e-Service Intelligence Research Lab, Centre for Quantum Computation & Intelligent Systems, Faculty of Engineering and Information Technology, University of Technology Sydney, Australia.

E-mail addresses: yizhang.bit@gmail.com (Y. Zhang), guangquan.zhang@uts.edu.au (G. Zhang), Hongshu.Chen@student.uts.edu.au (H. Chen), alan.porter@isye.gatech.edu (A.L. Porter), zhudh111@bit.edu.cn (D. Zhu), Jie.Lu@uts.edu.au (J. Lu).

http://dx.doi.org/10.1016/j.techfore.2016.01.015
0040-1625/© 2016 Elsevier Inc. All rights reserved.

simulated datasets, but show biases and limited scope; and cannot be readily adapted to real-world data; 2) New approaches combined with old, unsolved issues have increased the confusion of feature selection, e.g., in which situations does Term Frequency Inverse Document Frequency (TFIDF) analysis really benefit the text clustering process? Which one is better for text clustering — single words or phrases? 3) Similarity measurement is usually used to group similar items, however, is it possible to explore the relationships among topics, which would help identify significant topics or predict possible developmental directions?

Considering these concerns, this paper attempts to build up a semi-automatic method for ST&I topic analysis and forecasting. For the above concerns, we introduce a K-Means-based clustering approach for semi-supervised learning on semi-labeled ST&I records, and especially for the third concern, a topic analytic model is engaged in clustering, where we 1) apply a similarity measure approach to trace the interactions between topics and identify highly involved topics and 2) predict future trends via the changes of the TFIDF value of related topics in a time series. Based on the United States (US) National Science Foundation (NSF) Awards data, we construct a feature selection model to compare phrases and single words, TFIDF and normal term frequency value, and assembled sets of features. We then focus on the computer science domain and Big Data-related topics, and use TRM approaches to visualize both historical data-oriented analytic results and forecasting studies, where we creatively combine the objective quantitative evidence and expert knowledge in one TRM model.

The main contributions of this paper include: 1) we focus on the NSF data and construct a K-Means-based clustering methodology with high accuracy in a local K-value interval, where an optimized K value would be determined automatically; 2) we introduce a similarity measure function for topic relationship identification, which helps explore the interaction among TRM components quantitatively and predict possible future trends, and then, creatively visualize both objective analytic results and expert knowledge-based qualitative discussion of the TRM.

The rest of this paper is organized according to the following structure. “Related works” section reviews previous studies including text clustering, topic analysis, TRM, and a comparison between our research and related works. In the Methodology section, we present a detailed research methodology on the ST&I topic analysis and forecasting studies. The section “Empirical study” follows, using the US NSF Awards from 2009 to 2013 in the Division of Computer and Communication Foundation as a case. This section identifies topics by clustering approaches, illustrates the development trend visually, and engages expert knowledge in topic understanding and forecasting. Finally, we conclude our current research, noting limitations, and put forward possible directions for future work.

2. Related works

This section mainly reviews previous literature on text clustering, topic analysis, and TRM, and then, compares the significance of our work with related work.

2.1. Text clustering

The purpose of clustering analysis is to explore potential groups for a set of patterns, points, or objects (Jain, 2010). Analogously, text clustering concentrates on textual data with statistical properties and semantic connections between phrases or terms. Its algorithms seek to calculate the similarity between documents and reduce rank by grouping a large number of items into a small number of meaningful factors (Chen et al., 2013; Zhang et al., 2014a). Text clustering emphasizes statistical properties and semantic connections of words or phrases, and it is popular, while not necessary, to introduce TFIDF analysis for feature extraction (Aizawa, 2003; Wu et al., 2008). On one hand, various statistics-based approaches are available for text clustering, e.g., Principal Components Analysis (PCA) (Zhu and Porter, 2002), K-Means (Huang, 2008; Jain, 2010), and hierarchical cluster (Cutting et al., 1992; Beil et al., 2002). These approaches measure document similarity via a term–document matrix, in which co-occurrence analysis is most involved. On the other hand, the Topic Models approach, evolving from Latent Dirichlet Allocation (LDA) into a family of methods, has more recently been playing an active role in clustering. It engages a hierarchical Bayesian analysis for discovering latent semantic groups in a collection of documents (Blei and Lafferty, 2006; Blei, 2012).

2.2. Topic analysis

Several studies have applied text clustering analysis to information search and retrieval (Voorhees, 1986; Chang and Hsu, 1997; Begelman et al., 2006). Currently, in the ST&I studies these generated semantic clusters are usually identified as “topics,” and learning these topics extends to newer sub-domain topic analyses. Topic analysis comprises topic identification (Boyack et al., 2011; Small et al., 2014), topic detection and tracking (Cataldi et al., 2010; Dai et al., 2010; Lu et al., 2014), and topic visualization (Huang et al., 2014; Zhang et al., 2014b). In particular, Kontostathis et al. (2004) concluded this related research as Emerging Trend Detection (ETD), which was described as a system with components containing linguistic and statistical features, learning algorithms, training and test set generation, visualization, and evaluation. An important ancestor of ETD is Topic Detection and Tracking (TDT) — the first to afford systematic methods to discover topics in a textual stream of broadcast news stories (Allan et al., 1998). Significant systems for technology management include Technology Opportunity Analysis (TOA) and Tech Mining (Porter and Detampel, 1995; Porter and Cunningham, 2004), both of which perform value-added data analysis by extracting useful information from ST&I documents for a specified domain and identifying related component technologies, market stakeholders, and relations.

2.3. Technology Roadmapping

TRM is defined as a future-oriented strategic planning approach to connect technologies, products, and markets over time (Phaal et al., 2004; Winebrake, 2004). Researchers have contributed to construct basic criteria and schemes for qualitatively based TRM models (Garcia and Bray, 1997; Phaal et al., 2004; Walsh, 2004; Phaal et al., 2006; Robinson and Propp, 2008; Tran and Daim, 2008). At the same time, traditional bibliometric approaches (e.g., co-occurrence, co-citation, and bibliographic coupling) and information visualization techniques have been involved in various kinds of automated software routines to help build more intelligent TRM composing models (Zhu and Porter, 2002; Chen, 2006; Waltman et al., 2010). A general observation is that IT techniques take active roles in data pre-processing and expert knowledge makes good sense for result evaluation and refinement (Zhang et al., 2015). Hybrid TRM models that blend qualitative and quantitative methodologies have become a trend in current TRM studies (Yoon and Park, 2005; Lee et al., 2008; Choi and Park, 2009; Lee et al., 2009a; Lee et al., 2009b; Porter et al., 2010; Zhang et al., 2014c). In addition, considering the shortages of terms and phrases, subject–action–object structures have been introduced to probe for relationships among TRM components (Choi et al., 2011; Choi et al., 2013; Zhang et al., 2014b), and these novel attempts hold out the possibility to more deeply understand underlying development chains to help compose TRMs.

2.4. Comparison with related work

Based on a 2.15-million-MEDLINE-publication dataset, Boyack et al. (2011) presented an outstanding comparison study on several text-based similarity approaches, e.g., TFIDF, Latent Semantic Analysis, Topic Models, BM25, and PubMed's own Related Articles (PMRA) approach. The study covered almost all mainstream text clustering algorithms and included a detailed discussion summarizing the advantages and disadvantages of these approaches. However, Boyack et al.'s (2011) study only applied single word based analyses, and one possible reason that the PMRA approach achieved the highest accuracy would be that it was MEDLINE data-oriented. In addition, the TFIDF analysis was used as a separate similarity measure approach when a combination with other approaches could have provided more benefit.

Yau et al. (2014) used the LDA approach and its extensions to group labeled scientific publications from the Web of Science (WoS) data source, and they also compared the model with a basic K-Means approach in a clustering experiment and showed excellent precision and recall values with their approach. Comparably, Newman et al. (2014) proposed a similar comparison between LDA and PCA on Dye Sensitized Solar Cell (DSSC)-related publications from WoS data, and then analyzed possible reasons and discussed benefits of both approaches. They both considered combining the Term Clumping process (Zhang et al., 2014a) with LDA and PCA. Since LDA and its extension were only able to deal with single words, the attempt on this combination was to simply use underlining to link words in phrases; Additionally, the seven pre-set categories in Yau et al.'s (2014) experiment – e.g., solar cells, RNAi, tissue engineering, graphene – were only lightly coupled, while the DSSC data in Newman et al.'s (2014) research was very narrow and highly coupled, which required deeply engaged expert knowledge to make judgments. Moreover, Yau et al.'s (2014) study only emphasized topic identification and did not provide more thoughts on the understanding of topics.

Gretarsson et al. (2012), in another interesting work related to our research, proposed a Topic Model approach "TopicNets" for visual analysis of large amounts of textual data. They also selected NSF grants to the University of California as one data sample. This is one of the few existing text analysis studies on NSF data, and the Topic Model approach they used held our interest. However, their emphasis was to construct adaptive software for textual visualization, while our research pays more attention to clustering. Also of considerable interest is the NSF "portfolio explorer" using topic modeling to overview research emphases (Nichols, 2014).

Small et al. (2014) introduced citation and co-citation analysis for topic identification and validated these emerging science topics via multiple data sources, e.g., Nobel prizes and other awards on selected topics. They contributed excellent work on identifying emerging topics and exploring the insights they hold. Comparing with document similarity measures, citation and co-citation analysis was another way to think about the clustering, and Small et al.'s (2014) empirical study addressed an entity of science and technology domains and related more to national R&D strategy and science policy.

There is a large number of contributions on hybrid TRM approaches, and we briefly compare our method with two representative models: Forecasting Innovation Pathways (Porter et al., 2010; Guo et al., 2012; Huang et al., 2012; Porter et al., 2013; Robinson et al., 2013) and Keyword-based Patent/Knowledge Mapping (Yoon and Park, 2005; Lee et al., 2008; Choi and Park, 2009; Lee et al., 2009a; Lee et al., 2009b; Jeong and Yoon, 2015). Both models provided systematic methodologies to apply TRM for topic identification and visualization in industrial applications. These match the scope of our model closely. However, our method also seeks to improve the clustering algorithm in accuracy. Rather than automatic application of the IT techniques [e.g., keyword-based self-correlation analysis (Zhu and Porter, 2002; Lee et al., 2008)], we emphasize the term–record pair for record-based clustering — the term vector included in one record would have more complete semantic meanings than isolated keyword clumps. This strategy shares similarities with the "word–topic–record" structure in the LDA approach.

3. Methodology

This study develops a methodology which contains a data pre-processing approach, a K-Means-based clustering analysis approach, and a forecasting approach. It uses NSF Awards data as a case study. The methodology seeks to define an ST&I textual data-driven, but adaptive, method for topic analysis and forecasting. The general research framework is given in Fig. 1.

Step 1. *Data pre-processing*

Normally, ST&I textual data have common fields (e.g., Title, Abstract) and special ones (e.g., International Patent Classifications in patent data, Program Element/Program Reference codes in NSF Awards data). Our purpose, in this step, is twofold: to remove meaningless data and retrieve relevant fields from raw data records.

In our previous study, we developed a Term Clumping process for technical intelligence that aims to retrieve core terms (words and phrases) from ST&I resources by performing term cleaning, consolidation, and clustering approaches (Zhang et al., 2014a). This paper introduces a modified Term Clumping process for feature extraction (core term retrieval), and generates a term–record-matrix at the end of this step.

We have focused on terms and phrases for quite a long time, and we value more meaningful semantic structures over single words. However, the increasing popularity of the LDA approaches pulls researchers back to rethink the desirable balance between phrases and single words. This prompts us to revisit consideration of which one is better for ST&I textual studies. In the context, it becomes promising to compare unigrams to n-grams via our method. At the same time, although it has been decades since TFIDF analysis was first introduced to weight terms for information retrieval, it is still critical to fully consider such techniques in our studies. Thus, we compare the efficiency of TFIDF with normal TF for clustering analyses. There are quite a few variations of TF and IDF weighting (e.g., log normalization of TF, inverse frequency smoothing, and probabilistic inverse frequency), but, in our design, the TFIDF value goes with the term-record pair, and the extreme cases (e.g., total record number in the set is 0, total number of records with specified term is 0, or total instances of specified term in a specified record is 0) would not happen. Thereupon, we apply the classical formula of TFIDF analysis (Salton and Buckley, 1988) to our method:

$$\mathrm{TFIDF} = \mathrm{TF} \times \mathrm{IDF} = \frac{\text{frequency of rerm } t_i}{\text{total instances of terms in record } D_j} \times \log \frac{\text{total records number in the set}}{\text{total number of records with term } t_i}.$$

Step 2. *Topic analysis*

This step sets up a training set of labeled data for machine learning and proposes a data-driven K-Means-based clustering approach. Several aiding models are added as described below:

1) Cluster validation model

Referring to the common performance measures in information retrieval, Recall, Precision, and F Measure are three target values for cluster validation. However, in our framework, the Recall value for the whole dataset is meaningless, since all records have been clustered and there is no missing record. In this context, we only use Total Precision as the target value in our model. The definition is given as follows:

$$\text{Total Precision} = \frac{\text{number of records clustered to the correct category}}{\text{total number of records in training set}}.$$

It is also necessary to mention that, in the cluster validation model, we label all records with the real category of the Centroid into which they are grouped. Therefore, the selection of the Centroid is one sensitive issue that will influence the cluster validation process, and generally we set the records with the largest Euclid length as the initial Centroids.

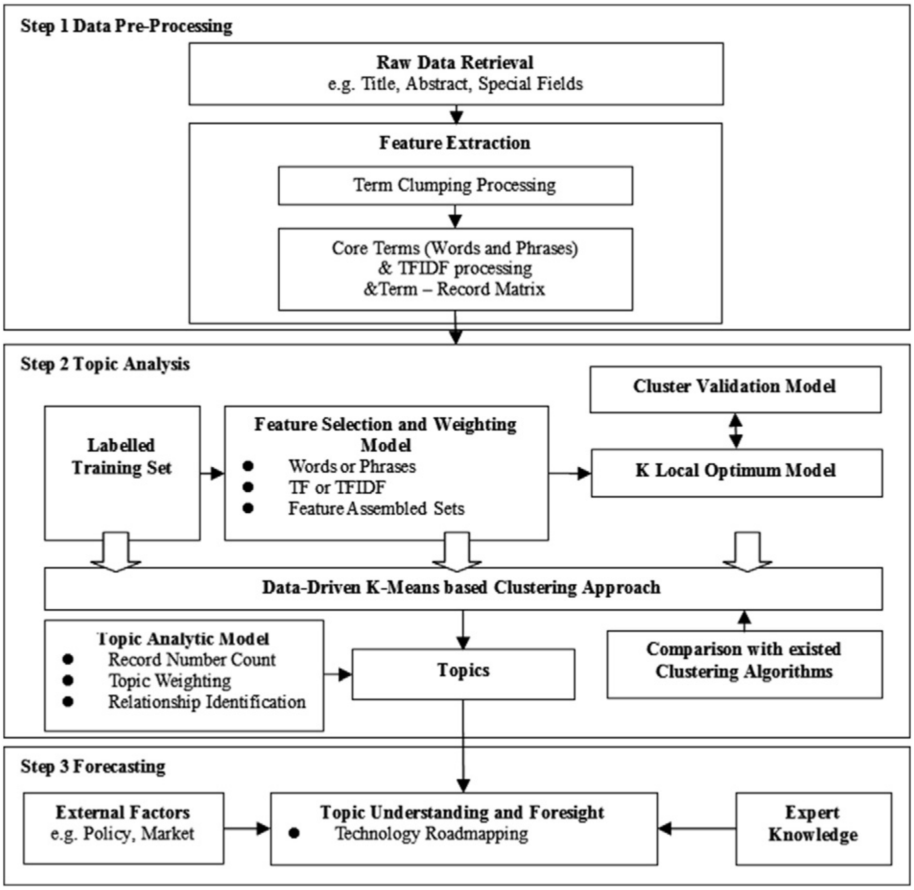


Fig. 1. Research framework for ST&I topic analysis and forecasting.

2) Feature selection and weighting model

This step uses NSF Awards as the sample to present our method. Title and Abstract (described as Narration in NSF Awards) are the most common fields used for text analysis. For NSF Awards data we also introduce the Program Element (PE) Code and Program Reference (PR) Code for our study. In NSF Awards, one record will be classified to at most 2 PE codes and at least 1 PR code, both of which are comprised of semantic terms. However, whereas these codes sometimes make good sense to help explore relations between records, sometimes they mislead. E.g., in the case of PR codes, most relate to techniques or methodologies while one or two codes would be used to describe the empirical domain. We develop an automatic way to assemble the best Title terms, Narration terms, PE codes, and PR codes. Six assembled sets are compared in this model:

- #1 Narration + Title terms
- #2 Narration + Title terms + PE code
- #3 Narration + Title terms + PE/PR code
- #4 Narration + weighted Title terms
- #5 Narration + weighted (Title terms + PE code)
- #6 Narration + weighted (Title terms + PE/PR code).

We set the six assembled sets for comparing three aspects: 1) are title-terms better than abstract ones? 2) Do the PE and PR codes help improve accuracy (both or only one of them)? 3) Does the weighting approach make sense? Thus, the comparison between #1 and #4 focuses on aspect 1; the comparison between/within #2 & #3 and #5 & #6 is for aspect 2, and #1–3 and #4–6 are used to indicate the efficiency of the weighting approach as we designed in the 3rd aspect.

We treat these four kinds of terms separately and introduce a weighting model into #4, #5, and #6 in order to calculate similarities. Normally, in the first three assembled sets, we calculate the similarity for Narration terms, Title terms, PE codes, and PR codes respectively, and use the mean as the final similarity value of the assembled set. In the last 3 assembled sets, with the help of the weighting model, the inverse ratio of the term amount is engaged. Let #4 serve as an example and we come out with the weights below:

$$\mathrm{V(D_n) = VN(D_n) + VT(D_n)}$$

$\mathrm{VN}(D_n)$ is the Term–Record Vector with only Narration terms, while $\mathrm{VT(D_n)}$ with only Title terms

$$\text{Let } \mathrm{T_N = VN(D_n) \cap VN(D_m)} \text{ and } \mathrm{T_T = VT(D_n) \cap VT(D_m)}$$

$$w_N = \frac{T_T}{T_{N+T_T}}, w_T = \frac{T_N}{T_N + T_T}.$$

Weighted similarity value: $\mathrm{s_w(D_n,D_m) = w_N \times S(VN(D_n),VN(D_m)) + w_T \times S(VT(D_n),VT(D_m))}$.

This model also attempts to compare additional topics in text analysis: the clustering accuracy of words and phrases, and normal TF and TFIDF values.

According to common sense, comparing to use of single words, phrases should be more specific and would help create a more accurate cluster, since relations among phrases seem more meaningful than among individual words. However, phrases appear much less frequently, leading to less overlap between records, and thus, might be detrimental to a similarity measure.

Our data-driven clustering approach is comprised of the above models and the clusters, identified as topics, to be generated at the end of this step. We also compare our clustering approach with two popular mainstream text clustering algorithms: LDA and Hierarchal Aggregative Clustering (HAC).

3) K-local optimum model

A traditional K-Means algorithm needs to set the K-value manually, and this value affects the clustering results heavily (Jain, 2010). Aiming to reduce this influence and to find the best K-value in a specific interval, our approach situates the cluster validation model in the loop for a specified interval, and decides the best K-value in the interval based on its F Measure.

The main concepts of K-Means are described as follows:

A. Initialization: Select the top K records with the highest Euclid length as the Centroid of K clusters;

Let tf_{in} as the frequency of term t_i in Record D_n

Record–Term Vector : $V(D_n) = \left\{tf_{1n}, tf_{2n}, \ldots, tf_{(i-1)n}, tf_{in}\right\}$

Euclid length of Record D_n : $ELEN(D_n) = \dfrac{1}{\sqrt{\sum tf_{in}^2}}$.

B. Record assignment: Classify each record to the Centroid with the highest similarity value;

Let $V(D_n)$ and $V(D_m)$ as the Record–Term Vector of Record D_n and Centroid D_m

Similarity value : $S(D_n, D_m) = \cos(V(D_i), V(D_m))$.

C. Centroid refine: Calculate the similarity between a record and its cluster; set the record with the highest similarity value as the Centroid of this cluster;

Let cluster $C = \{D_1, D_2, \ldots, D_{l-1}, D_l\}$

Similarity value : $S(D_n, C) = \dfrac{\sum_{k=1}^{l} S(D_n, D_k)}{l}$.

D. New & old Centroid comparison: If all new Centroids are the same as the old ones, the loop ends. Or else, return to Step B.

4) Topic analytic model

In this model, our endeavors are to explore statistical information on topics and provide these objective results for expert knowledge-based forecasting studies in the next step. We weight the topics via the TFIDF value (we apply the TFIDF formula of Step 1), where both the record number and the total term frequency of each topic will be counted. The TFIDF value would be used as the Y axis of topics in the TRM processing. Although it is critical to treat high-TFIDF weighted topics as important ones, the phenomenon that the TFIDF values of related topics increases or decreases in a time series indicates that the topics are becoming more important or not.

We also apply the similarity measure approach in the K-Means optimum model to calculate the semantic relationships between topics in the adjacent years. We define the highly similar topics as related topics, since it is not always easy to locate the same topics. We trace the dynamics of the statistical information of related topics, and we consider such changes as a kind of evolution in specific ST&I fields. In addition, we introduce a prevalence value (Zhang et al., 2014b) to identify the most representative terms for labeling topics, rather than labeling as the highest frequency term directly.

Step 3. *Forecasting*

In the past, based on terms, we proposed a semi-automatic TRM composing approach (Zhang et al., 2013). We also engage expert knowledge and understanding with external factors, e.g., policy, technique and development status. In particular, as mentioned by Kostoff et al. (2001), quantitative results are considered as objective evidence to assist decision making by domain experts. Expert knowledge plays a more important role in forecasting studies. The general steps of this section are outlined below:

1) Sort the generated topics by year and consolidate possible duplicate topics;
2) Send the topic list to domain experts for assessment, which includes topic names, representative terms, record numbers, and TFIDF values. Then, the experts are asked to mark the topics as "1" for "interesting topic at that time," "0" for "not interesting at that time," and "0.5" for "not sure" — there is no specific definition for "interesting," but our purpose is to let domain experts think about whether the topic relates to "emerging technology," "hot research question," or "it just makes sense to them;"
3) Calculate the marks for each topic and obtain rankings;
4) With the help of experts, consolidate similar topics, classify topics into appropriate phases of TRM, and locate them on the map;
5) Discuss the draft TRM (only the part of historical data-oriented analytic results) with domain experts via multi-round interviews, workshops, or seminars; obtain insights for future-oriented technology evolution and external relationships; and address specific concerns on forecasting studies.

The final output of our method is a TRM that blends both historical data-oriented analytic results and forecasting insights.

4. Empirical study

This section details the processes in the empirical study, which demonstrates the feasibility and efficiency of our methodology. This study uses NSF Awards data and focuses on computer science-related techniques, which dives into the origins of scientific innovation and draws reference for technical intelligence studies on other mature technologies or even emerging technologies.

4.1. Data

In the book *Lee Kuan Yew*: *The Grand Master's Insights on China, the United States, and the World*, the founding father of modern Singapore mentioned that "America's creativity, resilience and innovative spirit will allow it to confront its core problems, overcome them, and regain competitiveness" (Allison et al., 2013). Researchers and institutions are trying to evaluate the status of the competition for global innovation and to date no conclusion has been made. Undoubtedly, the United States currently is, and still will be, the world leader for a while to come due to its powerful capability to produce innovation.

As arguably the most important government agency in the US for funding fundamental research and education, in most fields of science and engineering, the US National Science Foundation accounts for about one-fourth of federal support to academic institutions for basic research. It receives approximately 40,000 proposals each year for research, education and training projects, approximately 11,000 of which are granted as awards (United States National Science Foundation, 2014). Understanding of NSF Awards data, which contains the most intelligent and innovative basic research and is more advanced than other regions by several years, could be considered an express path to revealing how the innovation evolution pathways of the US work. Such a research approach brings the core of the world's innovation and research to the forefront and the resulting knowledge could strongly support R&D management plans and science policy both in the US and other countries.

The NSF Awards database is open access, and all data can be downloaded from the NSF website. All awards are classified according to specific award type and division, and our study concentrates on Standard Grants, the most meaningful and the largest part of the NSF Awards. Moreover, most NSF Awards data is labeled by its Program Type, while a lesser part is unlabelled. Program Type sometimes entails very extensive classification (e.g., Collaborative Research or Early Concept Grants for Exploratory Research), or is very specific (e.g., Cyber Physical System, Information Integration and Informatics). Statistically, less than half of the NSF Awards data is labeled in detail or with any kind of "usable" classification, while others have common or meaningless labels or no label at all. As a result, we have treated the NSF Awards data as semi-labeled.

Step 1. *Raw data retrieval & feature extraction*

Although the NSF funds more than 10,000 proposals per year and online, open-access data dates back to 1959, considering our background, social networks, and the purpose of this paper, we only selected awards relating to computer science under the Division of Computer and Communication Foundations with an organization code that fell between 5,010,000 and 5,090,000. This narrowed the data set to 12,915 records. Since one of the main motivations for topic analysis is to address the innovation possibilities from NSF Awards data, we removed awards granting support for travel, summer school, and further education to arrive at a final total of 9274 records. We then applied the Term Clumping steps (Zhang et al., 2014a) for core term retrieval. The process for each step is given in Table 1. However, we did not choose the clustering steps, including Term Cluster Analysis and Combine Terms Network, from the Term Clumping steps because that reduces the number of similar terms and increases the difficulty of seeking similar pairs.

Before further processing, we dealt with a training set first. Given that the NSF Awards data are semi-labeled, we screened all 1124 records in 2009, chose 10 categories (shown in Table 2) associated with 587 records, and established a training set, which included 369 Title terms and 2161 Narration terms. We also imported 56 PE codes and 64 PR codes associated with these 587 records. After that, we calculated the TFIDF value for each Term–Record vector and generated a Term–Record matrix. We note that the 10 chosen categories (Table 2) are highly-coupled, which is generally considered a big challenge for existing text clustering approaches (Yau et al., 2014).

Table 1
Steps of Term Clumping processing.

	Step	# of N. terms*	# of T. terms*
1	9274 Records, with 9274 Titles and 8975 Narrations	–	–
2	Natural Language Processing via VantagePoint (VantagePoint, 2015)	254,992	17,859
3	Basic cleaning with thesaurus	214,172	16,208
4	Fuzzy matching	184,767	15,309
5	Pruning (remove terms appearing only in one record)	42,819	2470
6	Extra fuzzy matching	40,179	2395
7	Computer science based common term cleaning	38,487	2311
8	Deep Cleaning: expert-aided screening^	30,037	–

* N. = Narration, T. = Title.

^ Deep Cleaning: One computer-related PhD candidate and one data analyst help to screen and refine the term list.

Table 2
List of the ten selected categories.

No.	Category	Record num.	Notes
1	AF	46	Algorithmic foundations
2	CIF	51	Communications and information foundations
3	CPS	46	Cyber-physical systems
4	CSR	42	Computer systems research
5	III	47	Information integration and informatics
6	MRI	52	Major research instrumentation program
7	NeTS	66	Networking technology and systems
8	RI	75	Robust intelligence
9	SHF	94	Software and hardware foundations
10	TC	69	Trust worthy cyberspace

Step 2. *Topic analysis*

Based on the training set in the K local optimum model and considering the balance of the best number of clusters to treat at a time (fewer topics make results easier to understand, but more topics lead to a greater degree of accuracy), we set the interval of K value as [15, 20]. We compared the accuracy of the six assembled sets in Fig. 2. We also listed the maximum and mean of Total Precision of six Assemblies against Word & TFIDF, Phrase & TF, and Phrase & TFIDF in Table 3.

Before looking into Table 3 and Fig. 2, comparison between the results with/without the Deep Cleaning step is interesting. Our previous approach did not apply the Deep Cleaning step to the training set, where general thinking is that the TFIDF might yield surprising results with good precision, but this kind of "surprising" is not stable and the target value is lower than those with normal TF. However, after the Deep Cleaning step in the Term Clumping process, it is obvious that the deep-cleaned terms benefit significantly in the TFIDF analysis. This is because the TFIDF analysis introduces document frequency into the feature space along with term frequency, and helps increase the weighting of special terms. Thus, the more special terms there are, the better the results in the TFIDF. As discussed in Zhang et al. (2014a), we reaffirm that a good term cleaning step can be considered as basic pre-processing for TFIDF analysis.

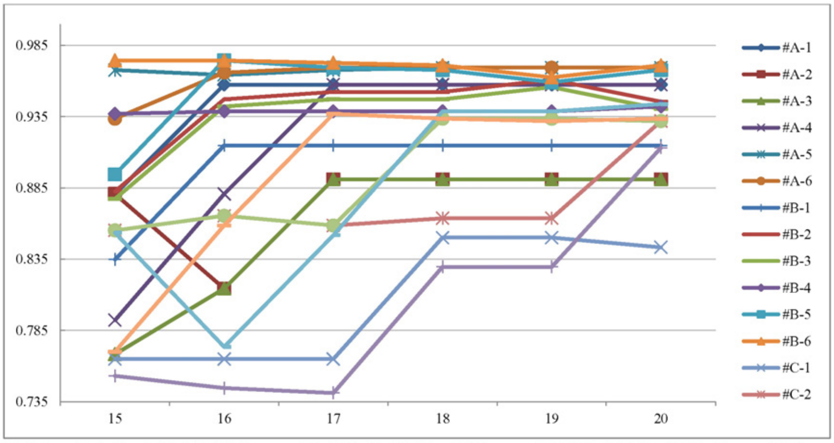
Generally, the combination of phrases and TFIDF values were the most accurate assembled sets, and the phrase-based ones worked better than those with words as shown in both Table 3 and Fig. 2. In comparing the efficiency of feature combinations, #5 and #6 were the best assembled sets. We try to explore the reasons behind these differences and outline some of our deductions below:

1) Comparison between phrase and single word

Comparison between #B and #C indicates the advantage of phrases. The lower term frequency that a phrase might have does not influence the clustering approach in generating accurate results, and TFIDF analysis also weakens the gap in term frequency. Phrases hold much stronger semantic relationships for similarity measures, and the possible negative effects from single words – for example the word "mining" occurring in both "data mining" and "mining industry" – are substantially reduced. In this case, based on the NSF Awards data and our text clustering approach, phrases matched our scope better.

2) Comparison between the TFIDF and normal TF value

The assembled sets #2, #3, #5, and #6 generated better results on #A, and both #A-5 and #A-6 arrive at the highest value of Total Precision. This helps to claim that the TFIDF analysis makes good sense for feature extraction, which weights common terms as lower value and is positive on cutting down noise terms. However, an interesting exception still exists in #1 and #4 — the only two assemblages without PE/PR



*#A before the tie line is for TF with Phrases, while #B is for TFIDF with Phrases and #C is for TFIDF with Words, e.g., #A-1 means Feature Combination #1 with TF & Phrase.

Fig. 2. Total Precision of the 6 assembled sets with TF and TFIDF.

codes. We will discuss the significance of the PE/PR codes in the next comparison, but our understanding is that our training set includes 10 high-coupling categories, and the narration and title might contain sizable noise terms, and the TFIDF analysis will mishandle these special terms. An example is that "neural network" is a basic algorithm for computer techniques and is applied in various domains (as shown in Table 2, it might belong to AF, CIF, CSR, III, MRI, NeTS, RI and SHF), but this term will be ranked highly by the TFIDF analysis.

3) Comparison among the six assembled sets of feature combinations

PE and PR codes can be treated as keywords of publications. They are special and meaningful, but far fewer terms originate from them than from the narration and title texts. Thus, it is beneficial to engage the PE/PR codes, but the difference between a PE code and PR code is not as obvious as other comparisons. We also attempted to read the PE and PR codes manually to distinguish differences among them, and discovered that the PE code acts as the main keyword for a proposal, while the PR codes contain a large amount of noise which might obfuscate relationships among proposals. For example, it is common to add one or two terms describing the empirical study, such as earthquake engineering or gene and drug delivery, or to use some general terms to emphasize the purpose of the research, such as science, math, eng. & tech education, and science of science policy. This may explain the reason that #2 and #A-5 are slightly ahead of #3 and #A-6, respectively. Additionally, if we also consider TFIDF scores, the minor reversion between #B-5 and #B-6 could be due to reducing the TFIDF-weights of the common and empirical study-related PR codes.

The sequence of #B-5 (weighted PE code), #B-2 (non-weighted PE code), and #B-4 (without PE code) definitively proves the advantages derived from PE codes. However, when exploring the reason why #A-2 was worse than #A-4, we ran Feature Combination #A-5.1 which uses a direct ratio to weight the PE code. The highest Total Precision of #A-5.1 was 0.8739, which was worse than those of #A-2 and #A-4. Therefore, a reasonable explanation is that the narration terms are more negatively misleading in the clustering analysis than the PE/PR code, and a direct ratio enlarges this negative impact while an inverse ratio weakens it.

Table 3
Maximum and average value of Total Precision of the six assembled sets with Word & TFIDF, Phrase & TF, and Phrase & TFIDF.

		#1	#2	#3	#4	#5	#6
#A Phrase & TF	Max	0.9574	0.8910	0.8910	0.9574	0.9693	0.9693
	Avg.	0.8456	0.8440	0.7666	0.8289	0.9106	0.8998
#B Phrase & TFIDF	Max	0.9148	0.9608	0.9557	0.9421	*0.9744*	*0.9744*
	Avg.	0.9015	0.9401	0.9350	0.9390	0.9554	0.9710
#C Word & TFIDF	Max	0.8501	0.9319	0.9336	0.9131	0.9438	0.9370
	Avg.	0.8064	0.8731	0.8964	0.8018	0.8833	0.8941

The italicize data are the highest accuracy combination in this comparative study.

In addition, considering #B-1, a remarkable improvement exists with #B-4. The possible driving force is that the title terms are much more specific than the narration terms, while the former one has a fewer amount, which enlarges the advantage in inverse-ratio weighting assemblages.

Based on the results and analysis of the above experiments, we chose "Phrases," "TFIDF value," "#6 Feature Combination of Narration terms, and weighted Title terms and PE/PR code," and "K = 15" as the most suitable K-Means Clustering approach for NSF proposal data.

4.2. Comparison with two text clustering algorithms

Aiming to demonstrate the accuracy and adaptability of our approach with NSF Awards data, we used the cluster validation model above to compare our results with those derived from the LDA approach and the HAC approach. One promising feedback on our comparison study was that the PE and PR codes were generated by human experts for classifying topics of NSF Awards, and these human-intervened codes would be a possible factor to improve the clustering results. At this stage, in order to avoid a higher extent of human intervention and ensure the fairness of this comparison, we compared LDA and HAC approaches with the results of the #0 assembled set, which did not distinguish Narration terms and Title terms and just combined them together for clustering without TFIDF weights. We recorded the results in Table 4 (#29, #50, and #58 were specifically used for comparing with related configurations in LDA or HAC approaches).

1) Comparison with the LDA approach

The purpose of this comparison is to focus on the efficiency of the clustering ability of the LDA approach and the possible usability of the Term Clumping process for the LDA approach. Since the only permissive input of the LDA approach is a set of single words, we set the input as: the raw content of the combined title and narrations; and the Term Clumping-cleaned core phrases. Both of them were pre-processed by their own Natural Language Processing (NLP) function in the LDA approach. We used the basic LDA approach proposed by Yang et al. (2013), and also set the fixed topic number as the interval [15, 20]. The results of the LDA approach are listed in Table 5.

We concluded that the LDA approach, which is effective for text clustering, has increased efficacy with single words, large data sets and low-coupling domains. The LDA approach is single-word based

Table 4
Total Precision of the results with #0 assembled set.

Topic number	#15	#16	#17	#18	#19	#20	#29	#50	#58
#A-0 Phrase & TF	0.8010	0.8120	0.8090	0.8103	0.8341	0.8341	0.8474	0.8622	*0.8883*

The italicize data are the highest accuracy combination in this comparative study.

Table 5
Total Precision of the results of the LDA approach.

	Term Clumping-cleaned Phrases						Raw content of Title and Narration						
Topic number	#15	#16	#17	#18	#19	#20	#15	#16	#17	#18	#19	#20	#50
Total Precision	0.3782	0.4736	0.4617	0.4838	0.4225	*0.5451*	0.4566	0.5145	0.5145	0.4855	0.3867	0.5247	0.4682

The italicize data are the highest accuracy combination in this comparative study.

and therefore term frequency is its most important factor, however, removing and consolidating terms during the Term Clumping process drastically reduce the term frequency and render the cleaning process pointless. Our training set only contained 587 records with approximately 3000 single words, which just met the bottom of LDA's data requirements. As mentioned above, the training set involved 10 high-coupling sub categories in the computer science-related domain, which also posed extreme challenges for the word-based LDA approach. E.g., "neural network," "social network," and "computer network" belong to different sub domains, but the shared word "network" made them relate at a high possibility. We also tried to run the LDA approach in our dataset with 50 topics to confirm the "large fixed topic number" preference, since Yau et al. (2014) fixed the topic number as 50, but the results were not prospective and even worse than those with 20 topics.

2) Comparison with the HAC approach

In previous studies, the HAC approach has received rave reviews for its accuracy. In our study, we ran a basic HAC approach for comparison. The algorithm used is shown below.

A. Initialization: to set each record as a cluster;
B. Iteration: to calculate similarity among clusters and group the two clusters with the highest similarity value;
C. Terminal condition: to set the cluster number as the terminal condition, where a threshold will be used to illustrate the percentage of the cluster number in the record number.

The results of the HAC approach are shown in Table 6, and it is obvious that the HAC approach had better efficacy with large topic numbers and the Term Clumping process, but was not as accurate as our K-Means-based clustering approach. Another interesting reference value was that we spent hours running this basic and raw HAC approach for the small 587-record data set while our approach only took several minutes to generate results.

This comparison study highlights that our K-Means-based clustering approach adapted to the NSF Awards well, but we did not decry the LDA and HAC approaches. In this comparison, we first emphasize that the Term Clumping process played nicely to retrieve meaningful terms and phrases and helped increase the accuracy of clustering. It is also obvious that the Term Clumping process usefully involves human intervention, which would lead to better clustering results. Second, our method, including the Term Clumping process and the K-Means clustering model prefers terms and phrases. This preference would be unfair for the LDA approach in this comparison; thus, the unfavorable results above are understandable. Third, the purpose for the clustering model is to identify ST&I topics for further studies; therefore, we also prefer to control the number of topics, which facilitate qualitative follow-on approaches. This comparison helps demonstrate the advantage of our K optimum model.

Table 6
Total Precision of the results of the HAC approach.

	Raw Phrases			Term Clumping-cleaned Phrases		
Threshold	0.1	0.05	0.035	0.1	0.05	0.035
Topic number	58	29	20	58	29	20
Total Precision	0.7462	0.7428	0.7172	*0.8556*	0.8391	0.8187

The italicize data are the highest accuracy combination in this comparative study.

4.3. Sensitivity analyses for topic analysis

We compared six assembled sets and three feature extraction approaches (e.g., words & phrases, and TF & TFIDF) to select best combinations, and also compared our method with the LDA and HAC approaches. We noticed that several factors would influence our analytic results, thus, aiming to help extend our method to a broader scope, we summarize such contingent factors here:

A. ST&I data sources: diverse ST&I data sources have different biases for title terms and abstract terms, and also have different special features. Thus, depending on actual research targets and data, it is necessary to consider which is preferred — title terms, abstract terms, or some other content?
B. ST&I fields: analytic results vary with diverse ST&I applications. As an example, for NSF Awards, a vertical case in a specific technical field might prefer to use only the PE codes (PR codes are noisier). However, multidisciplinary studies would be fine for both PE and PR code analyses. In addition, clustering on a high-coupling field would need to enlarge the weights on distinctive features, while low-coupling cases would favor other content treatments.
C. Term cleaning: well-cleaned terms support TFIDF weighted results, while normal TF based analyses fare better using raw terms;
D. Topic number: almost all approaches prefer larger topic numbers. However, lager topic numbers increase difficulties in understanding, thus, it is vital meaningful to figure out an appropriate topic number for further studies;
E. Record volume: our method has no preference for record volume, but a larger dataset would help avoid possible negative impacts from data distribution or other abnormalities;
F. Words and phrases: the present comparisons favor the advantages of phrases, especially for analyses of large-scale data set.

Step 3. *Forecasting*

We applied our method to NSF Awards datasets from 2009 to 2013, numbering approximately 1000 each year and 4847 in total. After the K-Means clustering approach, we obtained 75 topics, then consolidated six duplicate topics, and retrieved 69 topics for further processing. The topic list (a selected sample is shown in Table 7) included topic label, description, record number, TFIDF value, and the similarity information — we recorded the top 2 most similar topics in the last year.

Table 7
Interesting topics and their statistical information (2009 and 2013 as samples).

Year	Topic label	Topic description	#R	TFIDF	Similarity	Rank	Lvl.
2009	Adaptive grasping & machine learning	Automatic speech recognition, computer vision, hierarchical visual categorization, robotic intelligence	102	0.1292	N/A	0.9306	1.5
	Behavior modeling & static analysis	Software maintenance, human centered computing, citizen science, dynamic environments, programming verification	142	0.1585	N/A	0.8861	1.5
	online social networks	Large scale, social network, machine learning, measurement	258	0.2115	N/A	0.8417	2.0
	Trustworthy cyber space & computer system research	Comparative analysis, internet, public security, software tools	120	0.1458	N/A	0.8013	2.5
	High performance computer & major instrument research	Computer engine, certifiably dependable software, consortium, virtual organization	83	0.1766	N/A	0.7722	3.0
	Cyber physical device & data mining	Cyber Physical system, research instrument, database system, trust worthy cyber, large scale	188	0.1857	N/A	0.7528	2.5
	Bayesian model computation & peta scale data	High dimensional data sets, information integration, computer supported cooperative work, computer graphics	95	0.1538	N/A	0.7278	1.0
	Human centered computing & virtual world	Social interaction, human computer interaction, computer vision, internet, virtual environment	57	0.1246	N/A	0.7252	2.0
	Hidden web databases & future internet	Internet, programmable measurement architecture, web data, network coding	41	0.0864	N/A	0.6389	1.0
	Ad hoc wireless networks & cross layer protocols	Ad hoc wireless networks, cross layer optimization, data centers, higher layer protocols, wireless sensor network	58	0.0940	N/A	0.6139	2.0
2013	Big data & machine learning	Large scale hydrodynamic Brownian simulations, parallel structured adaptive mesh refinement calculation, asynchronous learning, scalable system software, complex time series data	262	0.2345	Large scale data collection — 0.30 Robotic intelligence — 0.12	1	2.0
	Robotic intelligence & large scale	Robotics engineering, open source data center, data mining, sustainable future, software needs	182	0.2185	Robotic intelligence — 0.93 RFID system — 0.21	1	3.0
	Software system & automatic graphical analysis	Machine learning, optimization, engineering practice, prediction, software foundation	117	0.1817	Integrated system usage — 0.41 Heterogeneous architecture — 0.26	0.7972	2.5
	Trustworthy cyberspace & secure protocols	Cyber learning, computer security, internet, ethical complexities, cyber security	162	0.1765	Cyber security — 0.91 Robotic intelligence — 0.21	0.7532	2.5
	Cyber physical system & power system	Software system, verification, semiconductor, cyber infrastructure, major research instrument, software foundation	50	0.1302	Grid networks cyber — 0.86 Cyber infrastructure eco sys. — 0.11	0.7217	2.5
	Supporting knowledge discovery & social media	Scientific visualization, real time, mobile device, social science, human computer interaction	81	0.1567	Cooperative activity analysis — 0.42 Grid networks cyber — 0.11	0.7028	1.5
	Virtual organization & socio technical system	System design, virtual world, design guideline, meta-analysis, long tail science	24	0.0711	Artificial human agents — 0.08	0.6914	2.5
	NSF smart health & signal processing	Health influences, medical signals, pattern recognition, information theory, communication and information foundation	47	0.1066	Algorithm foundation — 0.21 Robotic intelligence — 0.14	0.6833	3.0

As mentioned in the Methodology section, we treated our auto-generated results as objective evidence for decision-making, and sought to combine quantitative and qualitative methods for topic analysis and forecasting studies. Hence, we engaged experts on computer-related subjects for topic confirmation and modification. Nine experts, comprising four senior researchers who have focused on computer-related studies for more than 10 years and five PhD candidates, from the Faculty of Engineering and Information Technology, University of Technology Sydney (UTS), Australia, were invited to serve as our panel. Based on their research experience and academic backgrounds, they helped us to consolidate similar topics and confirm whether the topics were interesting or not.

The effort to blend expert knowledge with our analytic results included:

A. We sent the raw topic list to the nine experts personally via emails, and then, the experts marked these topics as interesting, not interesting, and not sure;
B. We used an inverse ratio to weight the 4-researcher group and 5-PhD-candidate group, and then, removed all topics marked below "not sure — rank 0.5" resulting in 50 remaining topics. We list some of the final topics – all 10 topics in 2009 and 8 topics in 2013 – in Table 7;
C. We re-ran the topic analytic model in Step 2, and TFIDF analysis was applied to weight the topics (shown in Table 7);
D. Based on Table 7 and related information, we followed the TRM composing model (Zhang et al., 2013) to draw the historical data-oriented analytic results in Fig. 3 — X axis is the time series from 2009 to 2013, Y axis is the TFIDF value of the topics, and the topics are linked with similar previous ones;
E. The statistical information derived from the topic analytic model was used to identify several significant topics: six increasing topics (marked by up-arrows), the record number of which was more than similar topics in the past years, and seven most highly-involved topics (marked with dark boxes), which had interactions with two or more previous topics;
F. We separated interviews with one senior researcher and four PhD candidates of our expert panel. They, based on their knowledge, highlighted six significant topics (marked with stars), which were considered as fundamentally important IT techniques or research topics for Big Data. In this context, we finished a draft version of the historical data-based analysis in Fig. 3.
G. A two-round workshop then was used for the modification of the historical analytic results and the forecasting studies, where all nine of our expert panel and two external experts were engaged. We first presented the detailed statistical information of all topics and a raw version of Fig. 3 to the experts, and then, they discussed our analytic results, especially focused on the selected topics (e.g., increasing, highly-involved, and significant topics). We led the discussion with designed directions that included "do you think the current increasing topics would maintain such trends in the near future?" "Which topics/techniques revealed on the graph hold strong potential over the next three to five years?" and "Which topics/techniques that did not exist on the graph would also be emerging ones?". Finally, we summarized several evolutionary trends and five Big Data issues — software tools, techniques & algorithms, analytic scope, analytic capability, and data management, and identified four Big Data-related emerging topics, which would be considered as the future directions in the following decade.

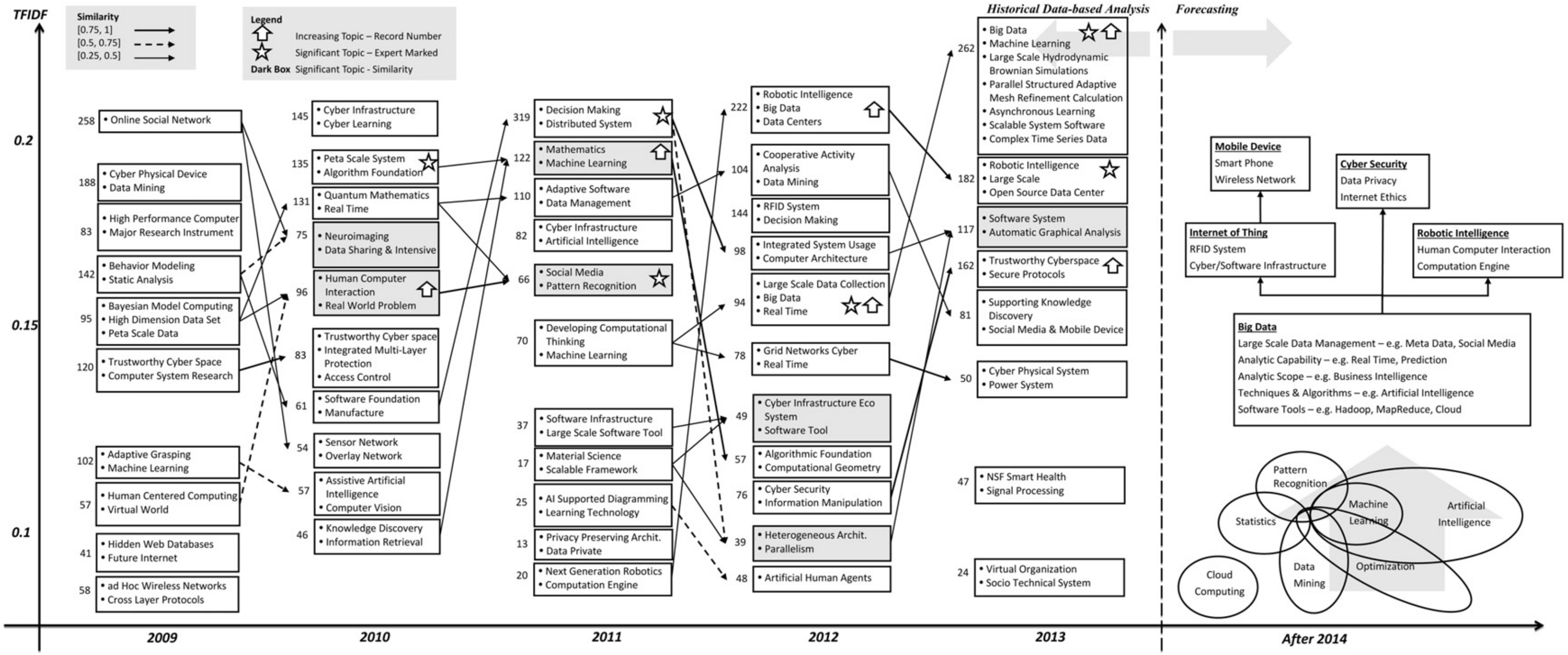

TFIDF
Similarity
[0.75, 1]
[0.5, 0.75]
[0.25, 0.5]
Legend
Increasing Topic – Record Number
Significant Topic – Expert Marked
Dark Box Significant Topic - Similarity
0.2
0.15
0.1
258 • Online Social Network
188 • Cyber Physical Device • Data Mining
83 • High Performance Computer • Major Research Instrument
142 • Behavior Modeling • Static Analysis
95 • Bayesian Model Computing • High Dimension Data Set • Peta Scale Data
120 • Trustworthy Cyber Space • Computer System Research
102 • Adaptive Grasping • Machine Learning
57 • Human Centered Computing • Virtual World
41 • Hidden Web Databases • Future Internet
58 • ad Hoc Wireless Networks • Cross Layer Protocols
145 • Cyber Infrastructure • Cyber Learning
135 • Peta Scale System • Algorithm Foundation
131 • Quantum Mathematics • Real Time
75 • Neuroimaging • Data Sharing & Intensive
96 • Human Computer Interaction • Real World Problem
83 • Trustworthy Cyber space • Integrated Multi-Layer Protection • Access Control
61 • Software Foundation • Manufacture
54 • Sensor Network • Overlay Network
57 • Assistive Artificial Intelligence • Computer Vision
46 • Knowledge Discovery • Information Retrieval
319 • Decision Making • Distributed System
122 • Mathematics • Machine Learning
110 • Adaptive Software • Data Management
82 • Cyber Infrastructure • Artificial Intelligence
66 • Social Media • Pattern Recognition
70 • Developing Computational Thinking • Machine Learning
37 • Software Infrastructure • Large Scale Software Tool
17 • Material Science • Scalable Framework
25 • AI Supported Diagramming • Learning Technology
13 • Privacy Preserving Archit. • Data Private
20 • Next Generation Robotics • Computation Engine
222 • Robotic Intelligence • Big Data • Data Centers
104 • Cooperative Activity Analysis • Data Mining
144 • RFID System • Decision Making
98 • Integrated System Usage • Computer Architecture
94 • Large Scale Data Collection • Big Data • Real Time
78 • Grid Networks Cyber • Real Time
49 • Cyber Infrastructure Eco System • Software Tool
57 • Algorithmic Foundation • Computational Geometry
76 • Cyber Security • Information Manipulation
39 • Heterogeneous Archit. • Parallelism
48 • Artificial Human Agents
Historical Data-based Analysis
262 • Big Data • Machine Learning • Large Scale Hydrodynamic Brownian Simulations • Parallel Structured Adaptive Mesh Refinement Calculation • Asynchronous Learning • Scalable System Software • Complex Time Series Data
182 • Robotic Intelligence • Large Scale • Open Source Data Center
117 • Software System • Automatic Graphical Analysis
162 • Trustworthy Cyberspace • Secure Protocols
81 • Supporting Knowledge Discovery • Social Media & Mobile Device
50 • Cyber Physical System • Power System
47 • NSF Smart Health • Signal Processing
24 • Virtual Organization • Socio Technical System
Forecasting
Mobile Device
Smart Phone
Wireless Network
Cyber Security
Data Privacy
Internet Ethics
Internet of Thing
RFID System
Cyber/Software Infrastructure
Robotic Intelligence
Human Computer Interaction
Computation Engine
Big Data
Large Scale Data Management – e.g. Meta Data, Social Media
Analytic Capability – e.g. Real Time, Prediction
Analytic Scope – e.g. Business Intelligence
Techniques & Algorithms – e.g. Artificial Intelligence
Software Tools – e.g. Hadoop, MapReduce, Cloud
Pattern Recognition
Statistics
Machine Learning
Artificial Intelligence
Cloud Computing
Data Mining
Optimization
2009
2010
2011
2012
2013
After 2014

Fig. 3. Technology Roadmapping for big data research (based on NSF Awards data).

H. Free discussion helped us select seven domains of computer science and applied mathematics as highly-related techniques to Big Data. This understanding was totally based on the personal research background and subjective insights of the experts, and we blended them in Fig. 3 to better understand the relationships between Big Data techniques and existing knowledge.

The final output of our case study — the NSF Awards data-based Technology Roadmapping for computer science is shown in Fig. 3.

4.4. Discussion and implications

Currently considered one of the hottest topics, it is interesting and promising to explore and discuss Big Data in more detail. In March, 2012, the Obama Administration announced the *Big Data Research and Development Initiative* (The White House, 2012) to improve the ability to extract knowledge and insights from large and complex collections of digital data and help accelerate the pace of discovery in science and engineering, strengthen national security, and transform teaching and learning. Six Federal departments and agencies announced more than $200 million to launch the initiative — the NSF being one of them. Concerning the general technology development pathway and the situation of Big Data, we first discussed the origins of Big Data techniques, and evidence was discerned from Fig. 3. Then, we summarized several possible directions to foresee the development of Big Data techniques. Such insights were mostly derived from the analytic enhanced by the expert knowledge derived from the interviews and workshop. We attempted to extend the topic analysis to forecasting studies in such manner, and hope it anticipates a capability to adapt changing requirements for multiple applications.

1) Big Data is not a pure invention, but a kind of evolution from previous techniques and a solution for real-word problems.

It is easy to link Big Data with its “3V” features — volume, velocity, and variety (McAfee et al., 2012), and “large scale,” “real time,” and “unstructured data” are listed as the hottest terms in Big Data-related records. However, there is no entire invention here. In Fig. 3 we are able to track the original techniques and models of these “new” concepts. On the one hand, seven subdomains of computer science and applied mathematics were identified as the foundation of Big Data techniques — artificial intelligence, machine learning, data mining, optimization, pattern recognition, statistics, and cloud computing, and we could easily locate related topics of these seven subdomains in Fig. 3 or even in the data before 2009. On the other hand, large scale and real time related concepts, algorithms, and systems have been generated continuously since 2009 — e.g., “peta scale data [2009]”, “quantum mathematics and real time [2009]”, “large scale software tool [2010]”, and “large scale data collection [2012].” At this stage, we assert that the main concept of Big Data is not new, but rather the combination of previously existing, but also emerging, technologies, theories and techniques. This package plays active roles in dealing with real-world problems, and its applications have been and still will be one emergent direction of the computer science domain in the following decade.

2) Cyber security and internet of things are two hot topics relating to Big Data.

Outwardly, “trustworthy cyber space [2009, 2010 and 2013],” “privacy preserving architecture and data privacy [2011],” and “cyber security [2012]” seem to have no direct relationship to Big Data, but in May 2014 the White House announced another report, *Big Data*: *Seizing Opportunities Preserving Values* (The While House, 2014), which involved relationships among government, citizens, businesses, and consumers. It focused on how the public and private sectors can maximize the benefits of Big Data while minimizing its risks. Clearly, cyber security is considered a great risk in the Age of Big Data, and it is also obvious that in Fig. 3 the cyber security topic kept dominating a large proportion of papers from 2009 to 2013. Therefore, it is reasonable to imagine that, in the near future, privacy of Big Data and the corresponding privacy protecting techniques should be a big concern for both government and citizens in policy development and legal domains.

Another set of topics that attracts our eye is “adaptive grasping [2009],” “AI supported diagramming [2011],” “supporting knowledge discovery [2013],” and also application-related topics, including “ad hoc wireless network [2009],” “sensor network [2010],” “access control [2010],” “cyber infrastructure [2010],” and “RFID system [2012].” As the most powerful competitor of the US, China has identified the internet of things in its top 5 emerging industries, announced in the speech, *Let Science and Technology Lead China's Sustainable Development*, by then Premier Jiabao Wen (2009). Not uniquely, in the 2014 White House report mentioned above, internet of things is highlighted as the ability of devices to communicate with each other using embedded sensors that are linked through wired and wireless networks. This is also linked with Big Data. Thus, internet of things, including mobile device-related techniques, is likely to be another hot research topic in the coming years.

3) Big Data-oriented data centers or systems to apply Big Data techniques for real-world requirements.

As part of the Obama Administration's Big Data program, the NSF started its *NSF* “*Smart Health and Wellbeing*” program in 2012. This “seeks improvements in safe, effective, efficient, equitable, and patient-centered health and wellness services through innovations in computer and information science and engineering” (United States National Science Foundation, 2014). Although there is no direct topic related to health & wellbeing, various data analytic techniques, systems, and software would likely be the foundation of this program, and “NSF smart health” rose exponentially as a hot topic in 2013. With the push of the NSF program and the enormous pull of wellbeing requests in modern society, the application of computer techniques in health and wellness services promises to be an emerging industry for a long time to come. Although there was no other representative example in Fig. 3, the development of data analytic techniques and software tools is continuously increasing the capability for large-scale data collection (e.g., EHR — Electronic Health Records), processing, storage, and analysis, and we foresee that growth of Big Data-oriented data centers and systems will likely be a trend for both academia and industry in the near future.

4) The combination of robotic intelligence and Big Data would be a predictable direction for both engineering and IT techniques.

It has been a long time since people started to imagine intelligent robots. Although these topics are not new, and appear several times in Fig. 3, e.g., “next generation robotics [2011]” and “robotic intelligence [2012 and 2013]”, we foresee on-going robotic intelligence gains via Big Data advances.

As shown in the forecasting part of Fig. 3, Big Data derives from quite a few mature subjects of computer science and mathematics, where artificial intelligence would be considered as one of the underlying ones. Rapid change in the modern world and the increasing needs promote exploring insights from large scale data, especially unstructured data, e.g., audio, and video, resulting in interdisciplinary fusion, in some sense. MapReduce and Hadoop, then, lead the revolution of analytic software, perhaps likened to “the butterfly effect” of the Big Data Age. On the one hand, Big Data techniques are required for knowledge discovery, decision making, and even prediction, and business intelligence attracts the eyes of both governments and industries, where data-driven is highly praised, rather than objective. On the other hand, “real time” is identified as another landmark of Big Data, where researchers pursue the accuracy and efficiency of data analysis, and cloud computing becomes a necessary tool for Big Data. Briefly, Big Data is a management revolution, which builds upon the data and then results in a technical revolution, and the evolution of the techniques themselves would be considered as underlying stories.

5. Conclusions and further study

In the current Age of Big Data, it is common sense to transfer traditional objective-driven research into a data-driven empirical study, and this paper could be considered as this kind of an attempt. We focus on NSF Awards, propose a clustering approach for topic retrieval, and then engage expert knowledge to identify developmental patterns. A combination of quantitative and qualitative methods provides a promising approach to forecast potential advances. The main contributions of this paper include: 1) an NSF data-driven K-Means-based clustering approach with high accuracy and a local K optimum; 2) a similarity measure function for relationship identification of TRM components and a creative TRM model for visualizing both objective results and expert knowledge-based qualitative discussions. The empirical study to dive into the roadmapping of computer science provided a quick means to obtain relevant technical intelligence pertaining both to academia and industry. This served to identify core technology, trace technology evolutionary pathways, and help forecast techniques and products in the next generation. These possible advanced computing applications for specified industrial issues present capabilities for R&D planning. Such TRM modeling also generates Competitive Technical Intelligence (CTI) to inform strategic management.

There are also several limitations of this paper requiring more detailed and specific discussions. On the IT technical side, we emphasized accuracy more than efficiency and scalability, and the current approach would be time-consuming if dealing with larger record sets. Also, we stopped pressing forward when we generated ST&I topics and simply identified their relationships. We see promise in use of intelligent techniques (e.g., concept drift detecting technique) to semi-automatically train an algorithm to retrieve multi-dimensional relationships for further trend analyses. On the technology management side, we engaged experts for topic understanding and forecasting studies, but a systematic technology foresight process would be able to improve the efficiency of qualitative approaches. At this stage, we anticipate further study in four directions: 1) to continue to improve our clustering algorithm in extended dimensions – i.e., efficiency, robustness, and scalability; 2) to introduce smarter IT techniques for relationship identification among ST&I topics; 3) to introduce a systematic quantitative approach to weight/rank the analytic results – i.e., topics – to support expert-based decision making in further steps; and 4) to extend the empirical study to address multiple ST&I data sources and to take into account external environmental factors (e.g., science policies and market forces).

Acknowledgments

This work is supported by the Australian Research Council (ARC) under discovery grant DP140101366, the National High Technology Research and Development Program of China (Grant No. 2014AA015105), and the US National Science Foundation (Award #1527370 — Forecasting Innovation Pathways of Big Data & Analytics).

References

Aizawa, A., 2003. An information-theoretic perspective of tf–idf measures. Inf. Process. Manag. 39 (1), 45–65.

Allan, J., Carbonell, J.G., Doddington, G., Yamron, J., Yang, Y., 1998. Topic Detection and Tracking Pilot Study Final Report.

Allison, G., Blackwill, R.D., Wyne, A., Kissinger, H.A., 2013. Lee Kuan Yew: The Grand Master's Insights on China, the United States, and the World. MIT Press.

Begelman, G., Keller, P., Smadja, F., 2006. Automated tag clustering: improving search and exploration in the tag space. Paper Presented at the Collaborative Web Tagging Workshop at WWW2006 (Edinburgh, Scotland).

Beil, F., Ester, M., Xu, X., 2002. Frequent term-based text clustering. Paper Presented at the Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining.

Bengisu, M., 2003. Critical and emerging technologies in materials, manufacturing, and industrial engineering: a study for priority setting. Scientometrics 58 (3), 473–487.

Blei, D.M., 2012. Probabilistic topic models. Commun. ACM 55 (4), 77–84.

Blei, D.M., Lafferty, J.D., 2006. Dynamic topic models. Paper Presented at the Proceedings of the 23rd International Conference on Machine Learning.

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

Bughin, J., Chui, M., Manyika, J., 2010. Clouds, big data, and smart assets: ten tech-enabled business trends to watch. McKinsey Q. 56 (1), 75–86.

Cataldi, M., Di Caro, L., Schifanella, C., 2010. Emerging topic detection on twitter based on temporal and social terms evaluation. Paper Presented at the Proceedings of the Tenth International Workshop on Multimedia Data Mining.

Chang, C.-H., Hsu, C.-C., 1997. Customizable multi-engine search tool with clustering. Comput. Netw. ISDN syst. 29 (8), 1217–1224.

Chen, C., 2006. CiteSpace II: detecting and visualizing emerging trends and transient patterns in scientific literature. J. Am. Soc. Inf. Sci. Technol. 57 (3), 359–377.

Chen, H., Zhang, G., Lu, J., 2013. A time-series-based technology intelligence framework by trend prediction functionality. Paper Presented at the Systems, Man, and Cybernetics (SMC), 2013 IEEE International Conference on.

Choi, C., Park, Y., 2009. Monitoring the organic structure of technology based on the patent development paths. Technol. Forecast. Soc. Chang. 76 (6), 754–768.

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.

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 (1), 52–74.

Cutting, D.R., Karger, D.R., Pedersen, J.O., Tukey, J.W., 1992. Scatter/gather: a cluster-based approach to browsing large document collections. Paper Presented at the Proceedings of the 15th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval.

Dai, X.-Y., Chen, Q.-C., Wang, X.-L., Xu, J., 2010. Online topic detection and tracking of financial news based on hierarchical clustering. Paper Presented at the Machine Learning and Cybernetics (ICMLC), 2010 International Conference on.

Garcia, M.L., Bray, O.H., 1997. Fundamentals of Technology Roadmapping: Sandia National Laboratories Albuquerque, NM.

Gretarsson, B., O'donovan, J., Bostandjiev, S., Höllerer, T., Asuncion, A., Newman, D., Smyth, P., 2012. Topicnets: visual analysis of large text corpora with topic modeling. ACM Trans. Intell. Syst. Technol. 3 (2), 23.

Guo, Y., Ma, T., Porter, A.L., Huang, L., 2012. Text mining of information resources to inform forecasting innovation pathways. *Tech. Anal. Strat. Manag.* 24 (8), 843–861.

Huang, A., 2008. Similarity measures for text document clustering. Paper Presented at the Proceedings of the Sixth New Zealand Computer Science Research Student Conference (NZCSRSC2008), Christchurch, New Zealand.

Huang, L., Guo, Y., Porter, A.L., Youtie, J., Robinson, D.K., 2012. Visualising potential innovation pathways in a workshop setting: the case of nano-enabled biosensors. *Tech. Anal. Strat. Manag.* 24 (5), 527–542.

Huang, L., Zhang, Y., Guo, Y., Zhu, D., Porter, A.L., 2014. Four dimensional science and technology planning: a new approach based on bibliometrics and technology roadmapping. Technol. Forecast. Soc. Chang. 81, 39–48.

Jain, A.K., 2010. Data clustering: 50 years beyond K-means. Pattern Recogn. Lett. 31 (8), 651–666.

Jeong, Y., Yoon, B., 2015. Development of patent roadmap based on technology roadmap by analyzing patterns of patent development. Technovation 39, 37–52.

Kontostathis, A., Galitsky, L.M., Pottenger, W.M., Roy, S., Phelps, D.J., 2004. A Survey of Emerging Trend Detection in Textual Data Mining Survey of Text Mining. Springer, pp. 185–224.

Kostoff, R.N., Toothman, D.R., Eberhart, H.J., Humenik, J.A., 2001. Text mining using database tomography and bibliometrics: a review. Technol. Forecast. Soc. Chang. 68 (3), 223–253.

Lee, S., Lee, S., Seol, H., Park, Y., 2008. Using patent information for designing new product and technology: keyword based technology roadmapping. *R&D Manag.* 38 (2), 169–188.

Lee, S., Yoon, B., Lee, C., Park, J., 2009a. Business planning based on technological capabilities: patent analysis for technology-driven roadmapping. Technol. Forecast. Soc. Chang. 76 (6), 769–786.

Lee, S., Yoon, B., Park, Y., 2009b. An approach to discovering new technology opportunities: keyword-based patent map approach. Technovation 29 (6), 481–497.

Lu, N., Zhang, G., Lu, J., 2014. Concept drift detection via competence models. Artif. Intell. 209, 11–28.

McAfee, A., Brynjolfsson, E., Davenport, T.H., Patil, D., Barton, D., 2012. Big data: the management revolution. Harv. Bus. Rev. 90 (10), 61–67.

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.

Nichols, L.G., 2014. A topic model approach to measuring interdisciplinarity at the National Science Foundation. Scientometrics 100 (3), 741–754.

Phaal, R., Farrukh, C.J., Probert, D.R., 2004. Technology roadmapping—a planning framework for evolution and revolution. Technol. Forecast. Soc. Chang. 71 (1), 5–26.

Phaal, R., Farrukh, C.J., Probert, D.R., 2006. Technology management tools: concept, development and application. Technovation 26 (3), 336–344.

Porter, A.L., Cunningham, S.W., 2004. Tech Mining: Exploiting New Technologies for Competitive Advantage vol. 29. John Wiley & Sons.

Porter, A.L., Detampel, M.J., 1995. Technology opportunities analysis. Technol. Forecast. Soc. Chang. 49 (3), 237–255.

Porter, A.L., Guo, Y., Huang, L., Robinson, D.K., 2010. Forecasting innovation pathways: the case of nano-enhanced solar cells. Paper Presented at the International Conference on Technological Innovation and Competitive Technical Intelligence.

Porter, A.L., Cunningham, S.W., Sanz, A., 2013. Extending the FIP (Forecasting Innovation Pathways) approach through an automotive case analysis. Paper Presented at the Technology Management in the IT-Driven Services (PICMET), 2013 Proceedings of PICMET'13.

Robinson, D.K., Propp, T., 2008. Multi-path mapping for alignment strategies in emerging science and technologies. Technol. Forecast. Soc. Chang. 75 (4), 517–538.

Robinson, D.K., Huang, L., Guo, Y., Porter, A.L., 2013. Forecasting Innovation Pathways (FIP) for new and emerging science and technologies. Technol. Forecast. Soc. Chang. 80 (2), 267–285.

Salton, G., Buckley, C., 1988. Term-weighting approaches in automatic text retrieval. Inf. Process. Manag. 24 (5), 513–523.

Small, H., Boyack, K.W., Klavans, R., 2014. Identifying emerging topics in science and technology. Res. Policy 43 (8), 1450–1467.

The While House, 2014. Big Data: seizing opportunities preserving values. Retrieved October 24, 2014, from http://www.whitehouse.gov/sites/default/files/docs/big_data_privacy_report_may_1_2014.pdf.

The White House, 2012. Big data is a big deal. Retrieved October 12, 2014, from http://www.whitehouse.gov/blog/2012/03/29/big-data-big-deal.

Tran, T.A., Daim, T., 2008. A taxonomic review of methods and tools applied in technology assessment. Technol. Forecast. Soc. Chang. 75 (9), 1396–1405.

United States National Science Foundation, 2014. United States National Science Foundation. Retrieved October 12, 2014, from http://www.nsf.gov/.

VantagePoint. (2015). VantagePoint. Retrieved April 10, 2015, from www.theVantagePoint.com

Voorhees, E.M., 1986. Implementing agglomerative hierarchic clustering algorithms for use in document retrieval. Inf. Process. Manag. 22 (6), 465–476.

Walsh, S.T., 2004. Roadmapping a disruptive technology: a case study: the emerging microsystems and top-down nanosystems industry. Technol. Forecast. Soc. Chang. 71 (1), 161–185.

Waltman, L., van Eck, N.J., Noyons, E.C., 2010. A unified approach to mapping and clustering of bibliometric networks. J. Inf. 4 (4), 629–635.

Wen, J., 2009. Let science and technology lead China's sustainable development. Retrieved October 14, 2014, from http://www.chinanews.com/gn/news/2009/11-23/1979809.shtml.

Winebrake, J.J., 2004. Alternate Energy: Assessment and Implementation Reference Book: The Fairmont Press, Inc.

Wu, H.C., Luk, R.W.P., Wong, K.F., Kwok, K.L., 2008. Interpreting tf–idf term weights as making relevance decisions. ACM Trans. Inf. Syst. 26 (3), 13.

Yang, L., Qiu, M., Gottipati, S., Zhu, F., Jiang, J., Sun, H., Chen, Z., 2013. Cqarank: jointly model topics and expertise in community question answering. Paper Presented at the Proceedings of the 22nd ACM International Conference on Conference on Information & Knowledge Management.

Yau, C.-K., Porter, A., Newman, N., Suominen, A., 2014. Clustering scientific documents with topic modeling. Scientometrics 100 (3), 767–786.

Yoon, B., Park, Y., 2005. A systematic approach for identifying technology opportunities: keyword-based morphology analysis. Technol. Forecast. Soc. Chang. 72 (2), 145–160.

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 (6), 707–724.

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.

Zhang, Y., Zhou, X., Porter, A.L., Gomila, J.M.V., 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 (2), 1375–1389.

Zhang, Y., Zhou, X., Porter, A.L., Gomila, J.M.V., Yan, A., 2014c. Triple helix innovation in China's dye-sensitized solar cell industry: hybrid methods with semantic TRIZ and technology roadmapping. Scientometrics 99 (1), 55–75.

Zhang, Y., Robinson, D., Porter, A.L., Zhu, D., Zhang, G., Lu, J., 2015. Technology roadmapping for competitive technical Intelligence. Technol. Forecast. Soc. Chang. http://dx.doi.org/10.1016/j.techfore.2015.11.029.

Zhu, D., Porter, A.L., 2002. Automated extraction and visualization of information for technological intelligence and forecasting. Technol. Forecast. Soc. Chang. 69 (5), 495–506.

Yi Zhang is a dual-degree Ph.D. candidate at the Decision Systems and e-Service Intelligence Research Laboratory, Centre for Quantum Computation and Intelligent System, Faculty of Engineering and Information Technology, University of Technology Sydney, Australia, and the School of Management and Economics, Beijing Institute of Technology, China. His specialty is text mining and technology innovation management, especially technological assessment and forecasting with the combination of qualitative and quantitative methodologies.

Guangquan Zhang received his Ph.D. degree in applied mathematics from Curtin University of Technology, Perth, Australia, in 2001. He is currently an Associate Professor in the Faculty of Engineering and Information Technology, and the Co-director of the Decision Systems and e-Service Intelligence Research Laboratory, Centre for Quantum Computation and Intelligent Systems, at University of Technology Sydney. His main research interests include multi-objective and group decision making, decision support system tools, fuzzy measure and optimization, and uncertain information processing.

Hongshu Chen is a dual-degree Ph.D. candidate at the Decision Systems and e-Service Intelligence Research Laboratory, the Centre for Quantum Computation and Intelligent System, Faculty of Engineering and Information Technology, University of Technology Sydney, Australia, and the School of Management and Economics, Beijing Institute of Technology, China. Her main research interests include text mining, trend estimation, technology intelligence and information processing.

Alan L Porter is Professor Emeritus of Industrial & Systems Engineering, and of Public Policy, at Georgia Institute of Technology, where he remains 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 of some 220 articles and books. Current research emphasizes measuring, mapping, and forecasting ST&I knowledge diffusion patterns.

Donghua Zhu is currently a Professor and the Associate Dean of the School of Management and Economics, and the Director of the Knowledge Management and Data Analysis Laboratory, at Beijing Institute of Technology, China. His main academic research fields include science and technology data mining, technology innovation management, technology forecasting and management. His current research emphasizes big data analytics.

Jie Lu received the Ph.D. from Curtin University of Technology, Perth, Australia, in 2000. She is currently a Professor and the Associate Dean Research (Acting) in the Faculty of Engineering and Information Technology, and the Director of the Decision Systems and e-Service Intelligence Research Laboratory, Centre for Quantum Computation and Intelligent Systems, at University of Technology Sydney, Australia. Her main research interests include decision making modeling, decision support system tools, uncertain information processing, recommender systems, and e-Government and e-Service intelligence.