Multiple Science Data-Oriented Technology Roadmapping Method

Yi Zhang[1,2], Hongshu Chen[1,2], Guangquan Zhang[1], Donghua Zhu[2], Jie Lu[1]
[1]Decision Systems & e-Service Intelligence research Lab, Centre for Quantum Computation & Intelligent Systems, Faculty of Engineering and Information Technology, University of Technology Sydney, Australia
[2]School of Management and Economics, Beijing Institute of Technology, Beijing, P. R. China

Abstract--**Since its first engagement with industry decades ago, Technology Roadmapping (TRM) is taking a more and more important role for technical intelligence in current R&D planning and innovation tracking. Important topics for both science policy and engineering management researchers involves with the approaches that refer to the real-world problems, explore value-added information from the complex data sets, fuse the analytic results and expert knowledge effectively and reasonable, and demonstrate to the decision makers visually and understandable. Moreover, the growing variety of science data sources in the Big Data Age increases these challenges and opportunities. Addressing these concerns, this paper proposes a TRM composing method with a clustering-based topic identification model, a multiple science data sources integration model, and a semi-automated fuzzy set-based TRM composing model with expert aid. We focus on a case study on computer science related R&D. Empirical data from the United States National Science Foundation Award data (innovative research ideas and proposals) and Derwent Innovation Index data source (patents emphasizing technical products) provide vantage points at two stages of the R&D process. The understanding gained will assist in description of computer science macro-trends for R&D decision makers.**

I. INTRODUCTION

The growing variety of data sources in the Big Data Age not only increases the challenges and opportunities for traditional competitive technical intelligence, but also leads the revolution of business management and decision making. Of significance is "data-driven," which uses Information Technology (IT) to support rigorous, constant experimentation that guides decisions and innovations [3], and has being characterized as one of the most competitive features in various R&D planning and business models [13]. Concentrated on Science, Technology, & Innovation (ST&I) research, IT and related internet techniques afford good possibilities to transfer traditional documentations into huge electric records, resulting in trouble analyzing these textual science data sources with different structures. Publication and patent data leads current ST&I textual analyses, but the boom of massive new data sources, e.g., Twitter, news, customer comments, R&D project proposals, product reports, etc., also matches the concept of Big Data perfectly.

The traditional Technology Roadmapping (TRM) approaches, described as a representative, prominent, and flexible instrument for long-range technological forecasting and strategic planning, make good sense to actively incorporate business data into planning procedures [14, 19]. However, standing on the new situation of the Big Data age, the way that the current TRM study used to transfer vague human thoughts to defined numerical values [9] is still unfavorable, and the combination of qualitative and quantitative methodologies is also not as smart as what we imagine. Meanwhile, the rapid engagement of multiple science data sources with difference formats and emphasis introduces new challenges for the current studies. At this stage, the emerging concerns are the approaches that refer to the real-world problems, explore value-added information from the complex data sets, fuse the analytic results and expert knowledge effectively and reasonable, and demonstrate to the decision makers visually and understandable.

Aiming to provide a solution for the question that how to construct the multi-layer TRM to revel multi-dimensional information from the emphasis-differed science data sources and to explore the insights for technical intelligence understandings, this paper develops a 3-step method that includes: 1) the clustering-based topic identification; 2) multiple science data source integration; and 3) fuzzy set-based semi-automatic TRM generation. On the one hand, we extend the existing K Means-based text clustering algorithm [21] for multiple science data sources, which is able to deal with million phrases and words on a high accuracy and to identify the emerging topics for technical intelligence studies. Based on the TRM composing model [19] and the special emphasis of different science data sources, this paper defines a multi-layer TRM model to arrange topics and related concepts (e.g., idea, technique, product, etc.) to explore the potential relations. In particular, aiming to seek a more efficient way to add expert knowledge for the quantum-based TRM, this paper introduces fuzzy set [18] and technology life cycle concept to identify the technology commercialization level, and proposes the semi-auto model for TRM generation. Moreover, the empirical study selects the United States (US) National Science Foundation (NSF) Award data (innovative research ideas and proposals) and the Derwent Innovation Index (DII) patent data source (technical products), which demonstrates the efficiency and feasibility of our methods and also provides vantage points at the top-bottom stages of the R&D process and assists in description of computer science macro-trends for decision makers.

The main contributions of this paper are: 1) we refine the K Means-based clustering approach to adapt the topic identification task of multiple emphasis-differed science data sources; 2) we engage the fuzzy set with the traditional TRM model, which is to combine the analytic results and the expert knowledge in a visual way; 3) the process in which we think

and solve problems emphasizes the combination of qualitative and quantitative methodologies, and is also adaptive and transferrable to related ST&I researches.

This paper is organized as follows: the Related Works section reviewed the previous studies on TRM and text clustering. In the Methodology section, we present the detailed research method for multiple science data-oriented TRM, involving with a clustering-based topic identification model, a multiple science textual data sources integration model, and a semi-automated fuzzy set-based TRM composing model with expert aid. The Empirical Study section follows, using the US NSF Awards and DII patent data as the case studies. Finally, we conclude our current research and outline future work.

II. RELATED WORKS

This section reviews the related literatures that include the Technology Roadmapping and the Text Clustering approaches, and then, summarizes the limitations of these previous works.

A. Technology Roadmapping

Based on the significant work of Phaal et al. [14], who summarized previous TRM methods and constructed an effective qualitative composing model, TRM research has already been extended from qualitative study only to a combination of qualitative and quantitative methodologies. Representatively, Huang et al. [7] introduced a bibliometrics technique-based four-dimensional TRM for the science and technology planning of China's solar cell industry; Zhang et al. [22] combined TRM model with Triple Helix innovation and Semantic TRIZ concepts, and presented an empirical study on China's dye sensitized solar cell industry; Lee et al. [11] proposed a scenario-based TRM that involved with a plan assessment map and an activity assessment map for organizational plans, which also engaged the Bayesian network for topology and a causal relations definition. Geum et al. [5] added association rule mining to TRM, which provides a useful approach to identify relations between the items in different layers.

The current TRM model has been combined with various concepts and methods, applying into real ST&I assessment, forecasting, and planning. According to previous studies, it is promising to conclude the benefits of the current TRM studies as follows: 1) TRM provides a visual model to present content, which enables easy understanding of both the macro and micro level [19]; 2) The hierarchical structure of TRM helps to indicate the potential relationships between items on different layers [5]; 3) The flexibility of TRM sensibly adds multiple-dimensional impact factors for consideration, e.g., time, science policy, market pull, and technique push [7, 10, 15]; and 4) the previous TRM studies have perfect adaptability for general publication, and patent data sources and match well to the requests from different industry domains.

However, one of the toughest tasks for the current TRM study is the approach transferring vague human thoughts to defined numerical values [9], and this issue will heavily influence the efficiency and accuracy of the TRM auto-generation process. In this context, fuzzy set could be considered as a helpful instrument to address these concerns, which minimizes the expert aid and time consumed, but maximizes the usage of expert knowledge. Lu et al. [12] constructed a novel group decision making method with fuzzy hierarchical criteria for theme-based comprehensive evaluation in new product development, which calculated the ranking results by fusing all assessment data from human beings and machines. Lee et al. [9] introduced the fuzzy analytic hierarchy process to the TRM model, which was definitely an innovative attempt for the combination of fuzzy concepts and TRM, although the empirical study only applied to a small-range data set (five sub-technologies of hydrogen energy) with expert ranking.

B. Text Clustering

As a fundamental instrument of the TRM model, the text mining technique has already been introduced into the process for decades via bibliometrics. Focusing on the text clustering approaches, for topic identification especially, it is common sense that many clustering approaches are available to address these concerns, but there is no one approach that exactly fits all requests and scopes. The Latent Dirichlet Allocation approach is effective for topic retrieval by its "Word – Document – Topic" structure [1] but the current algorithm is only able to focus on individual words, which have much weaker relations with each other than phrases [21]. The K-Means approach is easy to handle with a massive data set and has a passable accuracy for the clustering results, but how the K value can be worked out should be considered as a tough task for current studies [8]. The Hieratical Aggregative Clustering approach, described as a high-accuracy algorithm, is time-consuming and has no optimum function for generated results [4, 6], and, sometimes, it will also result in "big clusters" – a large number of items are grouped into one or two clusters. Moreover, focusing on the basic similarity measure approach, the Cosine function is the most basic algorithm, and comparably, Wu et al. [17] proposed a hieratical-case tree to calculate the similarity between business cases via a tree structure, which provides an innovative approach for record similarity measurement.

III. METHODOLOGY

On the purpose of providing solutions for constructing the multi-layer TRM to reveal multi-dimensional information from the emphasis-differed science data sources and to explore the insights for technical intelligence understandings, this paper constructs the multiple science data-oriented TRM method with the following three steps:

1) The clustering-based topic identification model – we group the term clumping process [20]-cleaned features of

science data sources into "general feature" and "specific feature," and refine the data-oriented, but adaptive text clustering model [21] to identify the hot research topics and key technologies;

2) The multiple science data source integration model – in respect of the technology life cycle and for the special emphasis of different science data sources, we re-arrange the hieratical structure of TRM composing model in [19] and enrich the relations between components on the multi-layer landscape to display detailed topic changing routes for assessment and foresight studies;

3) The fuzzy set-based semi-automatic TRM generation model – in order to combine the qualitative and quantitative methodologies, we engage experts to evaluate topics by removing meaningless topics, consolidating duplicate topics, and highlighting significant ones, and we also evaluate each topic and group them into specified meaningful fuzzy sets, after which we generate the TRM in an automatic manner.

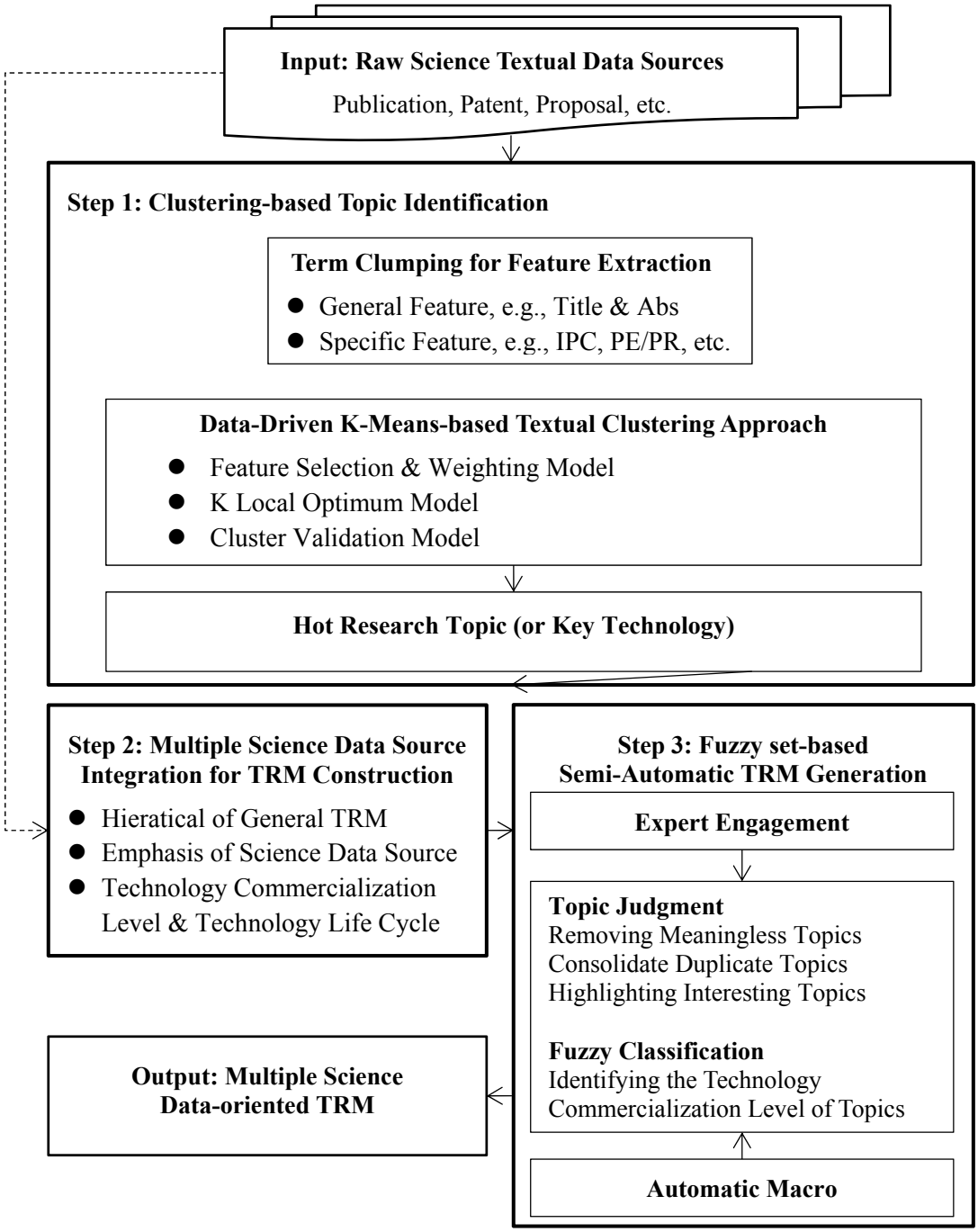

Fig. 1 Framework of Multiple Science Data-oriented TRM Method

In brief, the input of our method is the raw science textual data sources involving with some specified technical domains, the step 1 helps to retrieve value-added phrases and topics that reflect key technologies or important information, and based on the different emphases, we define the relations and the integration approach for the multiple science data sources. In the step 3, the expert knowledge is engaged to refine the topics and also to identify their technology commercialization levels, and then, we fuse all this information and generate the visualized TRM in a semi-auto way. At the end, the output is a graphic TRM for technical intelligence understanding on R&D plans and decision makings. The framework of the multiple science data-oriented TRM method is given in Fig. 1.

Step 1: Clustering-based Topic Identification

In modern society, various documentations related to science and technology could be considered as science data, including publications, patents, technical or product reports, proposals, news, etc., and, normally, textual data is a type of general format of science data. Therefore, our emphasis in this paper is science textual data, and text mining techniques will be used to handle the raw data for feature extraction. In [20], we developed a Term Clumping process to retrieve key terms from science textual data by term removal, consolidation, and clustering. However, the aim of the original Term Clumping is to remove common terms and select core terms; thus, we undertake modifications on this approach to meet our exact needs for feature extraction, to evolve it by adding a thesaurus for our specified technical domain, adjusting the existing Science and Academic Thesaurus, skipping the Term Frequency Inverse Document Frequency (TFIDF) approach for term removal, but applying it for feature weighting, and skipping the clustering-based term consolidation approaches (e.g., Term Clustering and Combine Term Network). Moreover, we categorize the features into two groups: "general feature" and "specific feature", which are used to describe the common fields (e.g., title, abstract, and full text) that exist in almost all science records, and the special fields (e.g., IPC in patent, UPC in United States Patent Trade Office patent, PE and PR code in NSF Awards, citation in publications and parts patents, and subject category in Web of Science, etc.) that belong to several specific data sources.

As already mentioned, the TFIDF approach is used to weight the features extracted by the Term Clumping process, and the general classical TFIDF formula [2] is described as below. The output from this section is a TFIDF-weighted term record matrix with categorized features.

$$\text{TFIDF} = \text{TF} \times \text{IDF} = \frac{\text{Frequency of Term } t_i}{\text{Total Instances of Terms in Record } D_j} \times log \frac{\text{Total Record Number in the Set}}{\text{Total Number of Records with Term } t_i}$$

The data-driven K-Means-based clustering approach presented in [21] involves with a cluster validation model, a K local optimum model, and a feature selection and weighting model. The benefits from this approach include: 1) the clustering approach works well with the core terms derived from the Term Clumping process; 2) The classical TFIDF value is effective for feature extraction by increasing the weighting of special technical terms and decreasing that of common terms; 3) Aiding with a training set, the K local optimum model could be used to get the most optimized K value in a selectable period. However, this approach concentrated on the NSF Award data, and the "PE Code" of NSF Awards was weighted as an important special feature after its training process. Considering this, we follow the main concept of the cluster validation and K local optimum model, but make modifications on the feature selection and weighting model.

As we categorize features into the two groups "general feature" and "specific feature", in the feature selection and weighting model we choose to set up two comparable assembled sets as detailed below, and the inverse ratio of term amount with general and specific features is used as the weighting approach.

- General Features + Specific Features
- General Features + Weighted Specific Features

Step 2: Multiple Science Data Source Integration

It is a general understanding that there are different emphases of current Science data sources. As shown in Fig. 2, emphasis of several main Science data sources is summarized below:

1) Academic proposal, e.g., NSF proposals, is usually granted by the national government to support academic institutions for basic research, the content of which focuses on new ideas, concepts, and any unrealized innovative actions, while the national R&D program proposal has a broad scope involving with the whole technology development stage.

2) Publication contributes to both basic research and application research, but, in detail, conference paper, e.g., IEEE paper, mostly presents draft research frameworks, experimental results, or mature ideas, while Web of Science paper, including SCI and SSCI data, and the EI Compendex paper emphasizes the fundamental research and application research, respectively.

3) Patent, e.g., DII patent, contributes to a mature application or product, the same as the detailed technical report or guidebook of a business service. In particular, patent terms should be vague and address legal effects (with patent barrier), and, comparably, academic terms are more clear and direct.

4) Business news, e.g., Factiva data highlights the social significance, where common technical terms and modifying adjectives will heavily influence the description. At the same time, text in social media is similar to that in news, but a more informal expression exists.

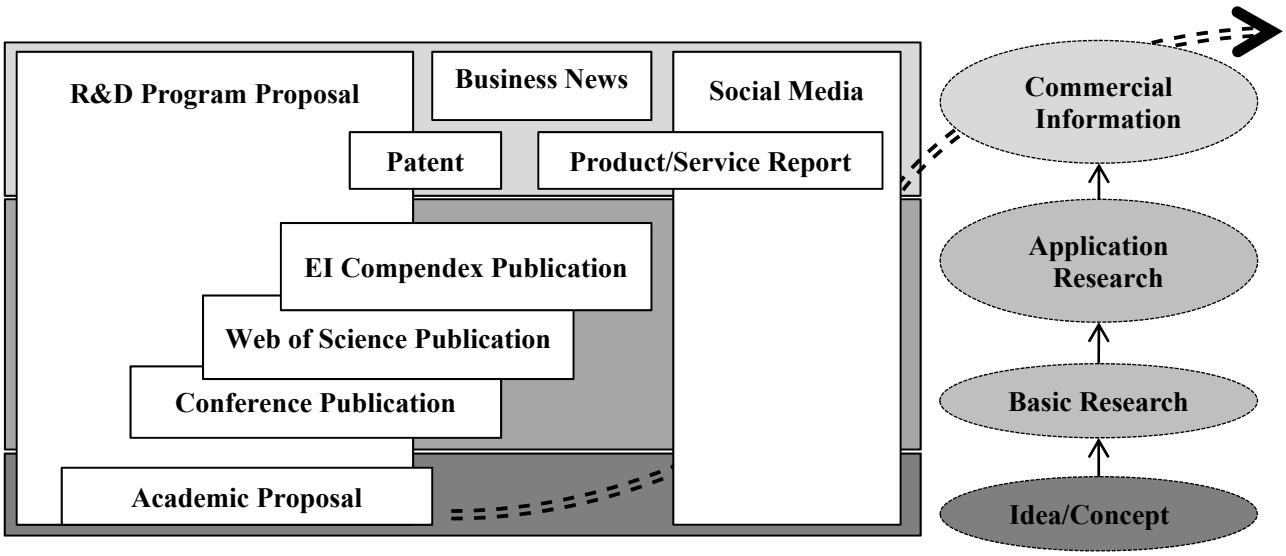

Fig. 2 Emphasis of Main Science Data Sources

Undoubtedly, a time gap always exists between different Science data sources. Normally, based on the technology life cycle concept, it takes approximately 3–5 years developing one technology from an idea to a patent, and may be shorter for an emerging technology. Addressing these concerns, a progressive comparison between different ST&I data sources with a different emphasis should make sense to better explore the detailed changing trends of existing technologies. In this context, based on the hierarchical landscape of TRM in [19] and the form of expression with TRM components in [23], we enrich the structure by: 1) distinguishing the scope of ST&I data sources with the shape of components, 2) defining linkages between components with multi-factors, e.g., semantic similarity, time, science policy, etc., and 3) identifying the universe "Technology Commercialization Level (TCL)" for specified fuzzy sets in which to engage the fuzzy concept. A sample of multiple science data-oriented TRM is given as Fig. 3.

In Fig. 3, Time and TCL is marked as the X and Y axis, respectively, while the X axis could be divided by year and several fuzzy sets of TCL compose the Y axis. On the mapping, different shaped components indicate topics derived from different science data sources, which also might be grouped on different layers because of their emphasis. In addition, experts will help to identify the linkages between components with their potential relations, and these linked components, named as a topic changing chain, could be used to describe the technology commercialization trends for technical intelligence understanding.

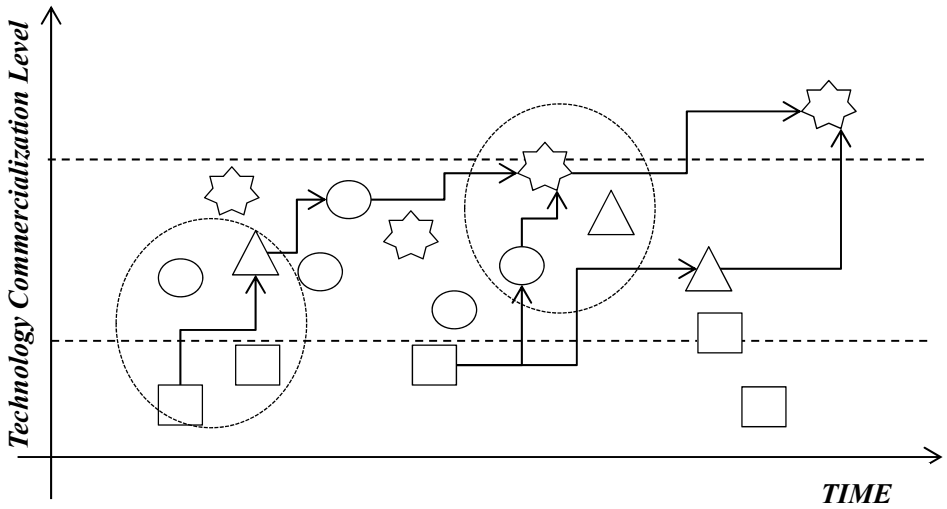

Fig. 3 The Sample of Multiple Science Data-Oriented TRM

Step 3: Fuzzy set-based Semi-Automatic TRM Generation

Although massive quantitative methods have already been used in the TRM model, in some sense, the engagement of experts is inevitable. As the common description for quantum-based TRM, experts help to refine the objective data evidence, to handle the allocation of components for TRM generation, and to understand the TRM for decision making. At this stage, we aim to minimize the aid of experts and to maximize the usage of expert knowledge in the limited time and scope, so that the fuzzy set can be considered as an effective tool to deal with this issue. Obviously, the definition of the technology commercialization level is a vague concept that depends on the personal understanding and benchmarks, which is definitely an important point related to the fuzzy concept.

This paper denotes "Technology Commercialization Level" as the universe $X(x) = [0, 1]$ and the power set $F(X) = \{A_1, A_2, \dots, A_{i-1}, A_i\}$, where A_i is a fuzzy set of X and $A_i(x)$ is its member function. In this context, it is reasonable to understand A_i as the index of a specific phase of TCL and to identify $A_i(x)$ with the real requests. The steps are outlined below:

1) Considering the cases, to identify $F(X)$, A_i, and $A_i(x)$;
2) Referring to expert knowledge, to evaluate TCL as $X(x)$ for each topic;
3) Based on $A_i(x)$, to calculate the topic's membership grade to each fuzzy set and to locate a topic onto the exact phase of TCL by its classified fuzzy set;
4) To generate TRM automatically via macros.

After our 3-step method, we identify the key topics, integrate the multiple science data sources, fuse the analytic results with expert knowledge, and generate the graphic TRM of technology commercialization as our final output. Definitely, the findings that the TRM indicates provide an outstanding support for R&D plans and decision makings.

IV. EMPIRICAL STUDY

Computer science might not be still considered as an emerging technology as what we did decades ago, but it has been integrated with information technologies and various engineering applications and become a fundamental instrument for multidisciplinary researches. The study on the technology commercialization pathways of the computer science might be able to draw a macro-landscape for the complex technology fusion and evolution and to hold great interests for R&D planning and technology management. In particular, if our attempt is promising for such a fundamental technical domain, it is comprehensive to imagine the form that we think and the process that we construct would be adaptable for other fundamental technology or even emerging technology. Since the limited condition for time and data sources, this paper only choose the NSF Award data and the DII patent data for empirical study. Nevertheless, the two data sources concentrate on the innovative ideas and mature technical products respectively, and the contrast on the technology commercialization level would be better to indicate the importance of the information fusion for multiple science data sources and also to demonstrate the benefits of our TRM method.

Step 1: The Clustering-based Topic Identification

This paper chooses two distinguished science data sources for the empirical study: NSF Award data and DII patent data. One of the most important reasons for this selection is to identify the technology commercialization trend via comparison between the academic proposals and the patents, emphasizing the bottom and top phases of TCL, respectively. In particular, as the NSF Awards are the significant and national academic proposal data in the US, we concentrate the DII patent data by setting the fields "basic patent country" and "priority country" as "US." At the same time, considering our background, we select Computer Science-related records from 2009 to 2013 in both data sources, from which 12,915 granted proposals under the Division of Computer and Communication Foundation in the NSF Awards [21], and 177,974 patents with topic and subject category "computer science" in DII, are obtained. We then apply Term Clumping steps for feature extraction, the process of which is given in Table 1.

TABLE 1 STEPS OF TERM CLUMPING PROCESSING

	Step	*NSF Awards*		*DII**	
		#T	*#A*	*#T*	*#A*
1	Raw Record	12,915		44,141	
2	Natural Language Processing via VantagePoint [16]	254992	17859	706,739	154,791
3	Basic Cleaning with thesaurus	214172	16208	679,736	131,577
4	Fuzzy Matching	184767	15309	-	-
5	Pruning*	42819	2470	19,926	19,930
6	Extra Fuzzy Matching	40179	2395	15,510*	16,603
7	Computer Science based Common Term Cleaning	30015	2311	14,029	16,529

* Considering the larger amount of DII data, we divide it by year for the Term Clumping process, and therefore, we only present DII data in 2013 as a sample; also, we apply Pruning to DII first, and then, Fuzzy Matching.

* #T = Number of Title Terms and #A = Number of Abstract Terms;

* In the Pruning process, we remove terms appearing in only one record in the NSF Awards and DII Titles, but remove terms appearing in less than five records in DII Abstracts.

After the learning process with both training sets, we set "Phrases, TFIDF value, Abstract Terms + Weighted (Title Terms and PE Code), K=16" for NSF Awards and "Phrases, TFIDF value, Abstract + Title Terms, K=12" for DII patents, the accuracy of which is 0.9710 and 0.9573, respectively. Compared with the same NSF case in [21] (K = 18 and Precision is 0.9893), we improve step 7 in Table 1 via a further cleaning with more Computer Science-related thesaurus and engage in TFIDF value, after which we achieve better accuracy and less K value. In this context, after a topic removal and consolidation process, we acquire 54 topics from the NSF Awards and 44 from the DII patents.

Step 2&3: Multiple Science Data Source Integration and Fuzzy Set-based Semi-Automatic TRM Generation

In this experiment, depending on Fig. 2 and the emphasis of NSF Awards and DII patents, we let the power set $F(X) = \{A_1, A_2, A_3\}$, indicating three fuzzy sets on "basic research," "application research," and "products", as the stages of technology commercialization. Furthermore, based on expert knowledge and our experience, we introduce Gaussian distribution to identify member functions, which also could be trained by machine learning techniques in the future. The three member functions are provided below and the distribution curves are shown in Fig. 4.

With the help of nine experts (Lecturers, Researchers, and PhD Candidates) from the Centre for Quantum Computation & Intelligent Systems, School of Software, University of Technology Sydney, Australia, we mark all topics with $X(x)$ for the TCL, and calculate the $A_i(x)$ for each member function. As a sample, we list parts of the marked topics in Table 2, and the generated TRM is given in Fig. 5.

$$A_1(x): X \sim N\left(0, {}^{1}/_{2\pi}\right), x \in [0,1]$$
$$A_2(x): X \sim N\left({}^{1}/_{2}, {}^{1}/_{2\pi}\right), x \in [0,1]$$
$$A_3(x): X \sim N\left(1, {}^{1}/_{2\pi}\right), x \in [0,1]$$

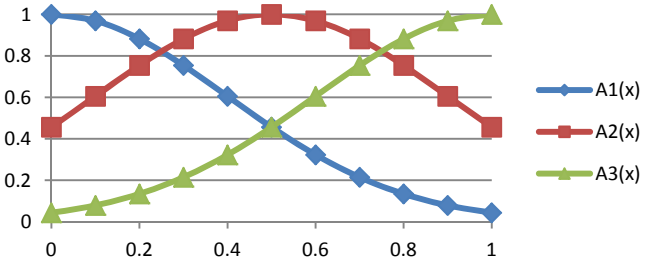

Fig. 4 Distribution Curves of Member Functions

Based on the X(x) and $(A_1(x), A_2(x), A_3(x))$ as we calculated in Table 2, we classify these 98 topics with the following rules: 1) to classify the topic to the fuzzy set as the First Preference with the largest membership grade, if the same membership grades occur, we prefer to choose the fuzzy set for the lower level of TCL; 2) If the second largest membership grade >= 0.7, we set the related fuzzy set as the Second Preference. In this consideration, we find 51 topics in $A_2(x)$ (42 NSF topics) and 47 topics in $A_3(x)$ (35 DII topics) with First Preference, and three topics in $A_1(x)$, all of which belong to NSF Awards, 13 topics in $A_2(x)$ (8 NSF topics), and 20 topics in $A_3(x)$ (13 NSF topics) with Second Preference.

TABLE 2 BIG DATA-RELATED TOPICS WITH MEMBERSHIP GRADES OF THREE FUZZY SETS

Year	*Topic*	*Topic Description*	*Data*	$X(x)$	$(A_1(x), A_2(x), A_3(x))$
2009	Adaptive Grasping	Adaptive Grasping, Automatic Speech Recognition, Empirical Mechanism Design, Hierarchical Visual Categorization, Infinite Bayesian Networks	NSF	0.63	(0.29, 0.95, 0.65)
2010	Reading Data	Reading Data, RFID Tag, Tag Memory, Configuration Data, Service Provider	DII	0.88	(0.09, 0.64, 0.96)
2011	Solving Large Systems	Linear Equations, Parallel Strategy, Recursive Divide, Solving Large Systems	NSF	0.67	(0.24, 0.91, 0.71)
2011	Remote Location	Remote Location, Retail Establishment, Source Code, Information Source, Navigation Database, Road Sign	DII	0.74	(0.18, 0.83, 0.81)
2012	Real Time	Real Time, Telecommunication Network, Advertisement Server, Mobile Communication Facility Data, Monetization Platform	DII	0.67	(0.24, 0.91, 0.71)
2012	Large Asynchronous Multi Channel Audio Corpora	Large Asynchronous Multi Channel Audio Corpora, Novel Speech Processing Advancements, Robotic Intelligence	NSF	0.43	(0.56, 0.98, 0.36)
2013	Video Frames	Video Frames, Encoding Video Frame, Improving Video Quality, Live Multicast System, Severe Degradation	DII	0.76	(0.16, 0.81, 0.83)
2013	Big Data	Algorithm Foundation, Big Data, Parsimonious Model, Mathematical Problems	NSF	0.5	(0.46, 1, 0.46)

Findings for the Commercialization of the Computer Science-related Technologies

In our previous studies, the single science data-based TRM takes active role in technical intelligence studies, which holds more benefits on exploring the inner features of the related technologies and identifying the technology development chains. However, the integration of multiple science data sources makes possible to stand on a higher macro-level to understand the technology commercialization pathways and to discover the gaps between academic research and commercial events. Addressing the concerns on Fig. 5 for reviewing the topic changing chains, linking related topics, and comparing the advantages and disadvantages of current technologies on the commercialization process, our findings are given as below:

1) The "mobile device" and the related techniques were, are, and still will be the hot commercial targets in the near future.

Obviously, the "mobile device" and the related techniques (marked as the blue solid box) keep being identified as the hot topics from 2009 to 2013 in the DII patents, which means keen competitions occur or will occur in this filed and the inventors (including the commercial firms) are seeking the intellectual property protection from the patents. At the same time, not only the mature products in the DII patents, but also some undertaking researches in the NSF researches (marked as the blue dashed box) could be addressed in Fig. 5, e.g., "mobile video processing" and "wireless smart camera networks." Therefore, it is reasonable to conjecture that the innovations and advanced techniques will also be engaged and transferred as a strong technical support for the follow-up developments.

2) Big Data is not a creation, but a result of technology evolution and fusion, all related techniques of which are able to track down the origins.

Big Data is an unavoidable topic in recent years. The social media, e.g., Twitter, Facebook, etc., is more popular than any other periods in the history, and the boom of various new techniques, e.g., MapReduce, Hadoop, etc., also illustrates the revolutionary changes. In this situation, the voice that highlights the new creation of the Big Data-related techniques would be enough to have its supporters. However, as shown in Fig. 5, it is definite to declare that the Big Data could be considered as the results of a kind of technology evolution or fusion, and all related techniques are able to track down the origins. Extended the discussion in [21], the social media-leaded online social network and web data (marked as the orange dotted box) constitute parts of the foundation of Big Data, and the coming Big Data age also increases the concerns on the information security (marked as the orange dashed box) rapidly. On the other hand, the efforts on the improvement of the existed algorithms (marked as the orange solid box) have never been stopped, which compose the mainstream techniques of the Big Data Age.

In addition, if we narrow down the scope to the technology commercialization, the current Big Data-related research still concentrate on the NSF Awards and stand at the fundamental stages that include constructing the concepts (e.g., Trust Worth Cyberspace, Real Time) and the algorithms (e.g., Bayesian Network Computing, Large Asynchronous Multi Channel Audio Corpora, Large Scale Hydrodynamic Brownian Simulations) and collecting the data and applying it to the experiments, while the Big Data-related business models and applications are crude, even there are no direct related topics in the DII patents. Thus, we should imagine that it will take time to transfer the new technique to the commercial practices, and this kind of attempts would be an obvious trend in the near future.

3) The process of the technology commercialization is much faster than that several years ago.

It is a common sense that the NSF interests include various fundamental researches that hold potential capabilities on further innovation, and the DII topics only concentrates on the applicable techniques with any commercial benefits, e.g., software or hardware techniques. However, considering the three fuzzy sets for the TCL studies, only few topics belong to the set "basic research," and most of them are on the medium level between basic research and products. The possible explanation could be that the current process of the technology commercialization is much faster than that several years ago, and new techniques could be used to solve the real-world problems in a short time, or as we say, the experimental time is engaging into the commercialization process. In addition, more and more innovations are originated from the real-world needs, which might be another strong driving force.

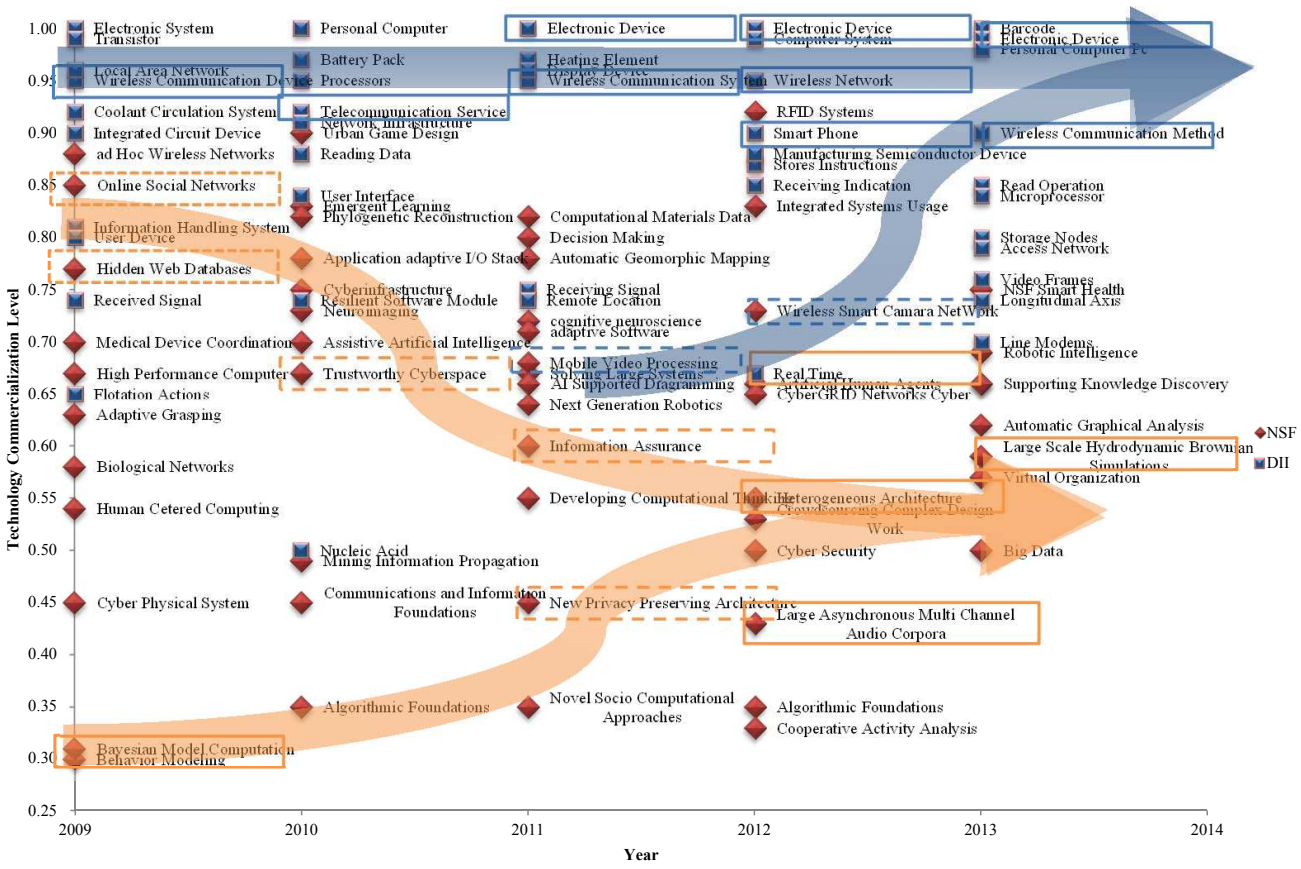

Fig. 5 Technology Roadmapping for Computer Science from 2009 to 2013 (based on NSF Awards and DII Patents)

V. DISCUSSION AND CONCLUSION

Highlighting the real-world needs and the engagement of the emphasis-differed multiple science data sources, this paper proposes an effective method to 1) explore value-added information from the complex data sets via a data-oriented but adaptive text clustering model, 2) introduce the fuzzy set to fuse the analytic results and expert knowledge smoothly, and 3) improve the traditional TRM approach to match the needs of multiple science data sources and address the commercial understandings for technical intelligence via a semi-auto TRM generation model. The thinking that combines qualitative and quantitative methodologies runs through the whole paper, the attempts on which provide great potential for related expert systems or decision making processes.

We anticipate further study to look into the following directions: 1) to enrich the approaches that identify relations between topics and detect detailed topic changing chains via more effective algorithms, e.g., concept drift; and 2) referring to Fig. 2, to continue to extend the scope of multiple science data sources engagement, e.g., adding publication data sources to retrieve details at the middle stage of technology commercialization to link with both innovative ideas and mature products. In addition, we will also consider the influences resulting from the different empirical domains, e.g., emerging technology, social science, mixed data with multidisciplinary, and address the concerns with more experiments.

ACKNOWLEDGEMENTS

This work is supported by the Australian Research Council (ARC) under discovery grant DP140101366.

REFERENCES

[1] Blei, D. M. and J. D. Lafferty, "Dynamic topic models," in *Proceedings of the 23rd international conference on Machine learning*, pp. 113-120, Pittsburgh, PA: ICML, June 2006.

[2] Boyack, K. W., D. Newman, R. J. Duhon, R. Klavans and M. Patek, "Clustering More than Two Million Biomedical Publications: Comparing the Accuracies of Nine Text-Based Similarity Approaches," *PLoS ONE*, vol. 6(3), e18029. doi:10.1371/journal.pone.0018029, 2011.

[3] Bughin, J., M. Chui and J. Manyika, "Clouds, big data, and smart assets: Ten tech-enabled business trends to watch," *McKinsey Quarterly*, vol. 56(1), pp. 75-86, 2010.

[4] Chen, H., G. Zhang, J. Lu and D. Zhu, "A Two-Step Agglomerative Hierarchical Clustering Method for Patent Time-Dependent Data," in *Foundations and Applications of Intelligent Systems*, pp. 111-121. Berlin Heidelberg: Springer, 2014.

[5] Geum, Y., H. Lee, Y. Lee and Y. Park, "Development of data-driven technology roadmap considering dependency: An ARM-based technology roadmapping," *Technological Forecasting and Social Change*, vol. 91, pp. 264–279, 2015.

[6] Han, J. and M. Kamber, *Data mining: concepts and techniques*. San Francisco, CA: Morgan Kaufmann, 2001.

[7] Huang, L., Y. Zhang, Y. Guo, D. Zhu and A. L. Porter, "Four dimensional Science and Technology planning: A new approach based on bibliometrics and technology roadmapping," *Technological Forecasting and Social Change*, vol. 81, pp. 39-48, 2014.

[8] Jain, A. K.; "Data Clustering: 50 Years beyond K-Means," *Pattern Recognition Letters*, vol. 31(8), pp. 651-666, 2010.

[9] Lee, S., G. Mogi, S. Lee and J. Kim, "Prioritizing the weights of hydrogen energy technologies in the sector of the hydrogen economy by using a fuzzy AHP approach," *International Journal of Hydrogen Energy*, vol. 36(2), pp. 1897-1902, 2011.

[10] Lee, S. and Y. Park, "Customization of technology roadmaps according to roadmapping purposes: Overall process and detailed modules," *Technological Forecasting and Social Change*, vol. 72(5), pp. 567-583, 2005.

[11] Lee, C., B. Song and Y. Park, "An instrument for scenario-based technology roadmapping: How to assess the impacts of future changes on organisational plans," *Technological Forecasting and Social Change*, vol. 90, pp. 285-301, 2015.

[12] Lu, J., J. Ma, G. Zhang, Y. Zhu, X. Zeng and L. Koehl, "Theme-based comprehensive evaluation in new product development using fuzzy hierarchical criteria group decision-making method Industrial Electronics," *IEEE Transactions*, vol. 58(6), pp. 2236-2246, 2011.

[13] McAfee, A., E. Brynjolfsson and T. H. Davenport, "Big Data. The management revolution," *Harvard Business Review*, vol. 90(10), pp. 61-67, 2012.

[14] Phaal, R., C. J. P. Farrukh and D. R. Probert, "Technology roadmapping—a planning framework for evolution and revolution," *Technological forecasting and social change*, vol. 71(1), pp. 5-26, 2004.

[15] Robinson, D. K. R. and T. Propp, "Multi-path mapping for alignment strategies in emerging science and technologies," *Technological Forecasting and Social Change*, vol. 75(4), pp. 517-538, 2008.

[16] VantagePoint, Retrieved 10/12/14 www.theVantagePoint.com

[17] Wu, D., J. Lu and G. Zhang, "Similarity measure models and algorithms for hierarchical cases," *Expert Systems with Applications*, vol. 38(12), pp. 15049-15056, 2011.

[18] Zadeh, L. A.; "Fuzzy sets," *Information and control*, vol. 8(3), pp. 338-353, 1965.

[19] Zhang, Y., Y. Guo, X. Wang, D. Zhu and A. L. Porter, "A hybrid visualisation model for technology roadmapping: bibliometrics, qualitative methodology and empirical study," *Technology Analysis & Strategic Management*, vol. 25(6), pp. 707-724, 2013.

[20] Zhang, Y., A. L. Porter, Z. Hu, Y. Guo and N. Newman, "'Term clumping' for technical intelligence: A case study on dye-sensitized solar cells," *Technological Forecasting and Social Change*, vol. 85, pp. 26-39, 2014.

[21] Zhang, Y., G. Zhang, A. L. Porter, D. Zhu and J. Lu, "Science, Technology & Innovation Textual Data-Oriented Topic Analysis and Forecasting: Methodology and a Case Study," in *Proceedings of 5th International Conference on Future-Oriented Technology Analysis*, Brussels: FTA, November 2014.

[22] Zhang, Y., X. Zhou, A. L. Porter, J. M. V. Gomila and A. Yan, "Triple Helix innovation in China's dye-sensitized solar cell industry: Hybrid methods with semantic TRIZ and technology roadmapping," *Scientometrics*, vol. 99(1), pp. 55-75, 2014.

[23] Zhou, X., Y. Zhang, A. L. Porter, Y. Guo and D. Zhu, "A patent analysis method to trace technology evolutionary pathways," *Scientometrics*, vol. 100(3), pp. 705-721, 2014.