

Please cite this article as: Zhang, Y., et al., Topic analysis and forecasting for science, technology and innovation: Methodology with a case study focusing on big data research, Technol. Forecast. Soc. Change (2016), <http://dx.doi.org/10.1016/j.techfore.2016.01.015>

Topic Analysis and Forecasting for Science, Technology and Innovation: Methodology and a Case Study

Yi Zhang^{1,2,*}, Hongshu Chen^{1,2}, Guangquan Zhang¹, Alan L. Porter³, Donghua Zhu², Jie Lu¹

¹Decision Systems & e-Service Intelligence research Lab, Centre for Quantum Computation & Intelligent Systems, Faculty of Engineering and Information Technology, University of Technology Sydney, Australia

²School of Management and Economics, Beijing Institute of Technology, Beijing, P. R. China

³Technology Policy and Assessment Centre, Georgia Institute of Technology, Atlanta, USA

Email Address: yizhang.bit@gmail.com; Hongshu.Chen@student.uts.edu.au; guangquan.zhang@uts.edu.au; alan.porter@isye.gatech.edu; zhudh111@bit.edu.cn; Jie.Lu@uts.edu.au;

Abstract The quantity and extent of current Science, Technology & Innovation (ST&I) 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 (2) identify changing topical emphases. Our results are carried forward to present mechanisms that forecast prospective developments using Technology Roadmapping. An empirical case study of Award 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.

Keywords Topic Analysis; Technological Forecasting; Text Mining; Text Clustering;

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 commercial models (Bughin et al. 2010; McAfee et al. 2012). Meanwhile, the researches addressing Science, Technology, & Innovation (ST&I) activities are widening into multiple perspectives (Bengisu 2003; Zhang et al. 2014a). Industry and national R&D efforts are beginning to track these trends in order to dominate global competition. However, the quantity 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 all of society.

ST&I data sources, involving with 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 help augment and amplify the capabilities of domain experts when dealing with real-world problems (Kostoff et al. 2001). Current ST&I text analysis focuses on emerging topics by combining both quantitative and qualitative methodologies and emphasizing automatic knowledge-based systems and bibliometric approaches. However, previous studies lack the macro scope to connect theoretical studies to real-world needs, and only concentrate on algorithm design and refinement, or the problem itself, ignoring possible quantitative improvements.

The concerns with recent research are:

- 1) The general text clustering algorithms are able to obtain perfect results on simulated datasets, but show bias on limited scopes and cannot be adapted to real-world data;
- 2) New approaches combined with old, unsolved issues have increased the confusion of feature selection, e.g. in which situation Term Frequency Inverse Document Frequency (TFIDF) analysis really benefits the text clustering process? Which one is better for text clustering, single words or phrases? etc.;
- 3) The gap between IT techniques and management applications still exists. It takes years, even decades, to apply a new IT technique to an appropriate empirical domain, and the dynamic change of the real world plus adaptability problems increase this difficulty.

Standing on these concerns, this paper attempts to address and predict the topic changes for technological forecasting and technical intelligence using a data-driven, but adaptive, methodology. First, we introduce a K-Means-based clustering approach for semi-supervised learning on semi-labelled ST&I records, which includes selection models for 1) phrases and single words, 2) TFIDF and normal Term Frequency (TF) value, and 3) assembled sets of features. We then use Technology Roadmapping (TRM) approaches in foresight studies to: combine quantitative evidence with expert knowledge; introduce visual models to represent innovation trends over a specified time period; and address concerns that may be raised in forecasting discussions. A case study on the United States (US) National Science Foundation (NSF) Awards is presented to demonstrate the feasibility of this approach.

The main contributions of this paper include: 1) the development of a data-driven, yet adaptive, K-Means-based clustering approach for topic analysis that uses a local K optimum model and a feature selection and weighting model to solve the K-value selection problem and adapt the specified data feature as needed; 2) application of the TRM approach for topic visualization, to effectively combine qualitative and quantitative methodologies as an efficient approach for forecasting studies; 3) experimental results that demonstrate the feasibility of our method for predicting trends in ST&I but also dives into the origins of the scientific innovation and draws reference for studies on other mature technologies or even emerging technologies.

The rest of this paper is organised according to the following structure. Related Works reviews previous studies including text clustering, topic analysis, and a comparison between our research and related works. In the Methodological Approach section, we present a detailed research method on ST&I textual data-oriented 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 study. This section identifies topics by clustering approaches, illustrates the development trend visually, and engages expert knowledge in foresight understanding. Finally, we conclude our current research and put forward possible directions for future work.

Related Works

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

Text Clustering

The purpose of clustering analysis is to explore the potential groups for a set of patterns, points, or objects (Jain 2009). 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. 2014b). Text clustering emphasizes the 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 statistic-based approaches are available for text clustering, e.g. Principal Components Analysis (PCA) (Zhu and Porter 2002), K-Means (Huang 2008; Jain 2009), Hierarchical Cluster (Cutting et al. 1992; Beil et al. 2002), etc. These approaches measure document similarity via a term-document matrix, in which the co-occurrence analysis is mostly evolved. On the other hand, the Topic Models approach, served as a springboard by Latent Dirichlet Allocation (LDA), 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).

Topic Analysis

It has been a long time since text clustering analysis was applied 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 of these topics is extended to the new sub-domain “topic analysis” including 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. 2014). 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 the events 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 2003), both of which perform value-added data analysis by extracting useful information from ST&I documents for a specified industry and identifying related component technologies, market stakeholders, and relations.

Comparison with the Related Works

Based on a 2.15-million-MEDLINE-publication dataset, Boyack et al. (2011) presented a comparison study on several text-based similarity approaches, e.g. TFIDF, Latent Semantic Analysis (LSA), 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 study only applied with single words, and one possible reason that the PMRA approach achieved the highest accuracy would be it was MEDLINE data-oriented. In addition, the TFIDF analysis was used as a separate similarity measure approach when a combination of other approaches would have provided more benefit.

Yau et al. (2014) used the LDA approach and its extensions to group labelled scientific publications from the Web of Science (WoS) data source, and they also compared their model with a basic K-Means approach in a clustering experiment and proved perfect 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 the WoS data, and then analysed the possible reasons and discussed the benefits of both approaches. They both considered combining the Term Clumping process (Zhang et al. 2014b) 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 underline to link each word of phrase; Additionally, the seven pre-set categories in Yau et al.’s experiment, e.g. solar cell, RNAi, tissue engineering, grapheme, etc., were low-coupled, while the DSSC data in Newman et al.’s research was very narrow and high-coupled, which required deeply evolved expert knowledge to make judgments. Moreover, Yau et al.’s 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 on large amounts of textual data. They also selected NSF granted proposals from the University of California as one data sample. This is one of the few existing 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.

Small et al. (2014) introduced the 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. Obviously, they contributed excellent work on identifying emerging topics and exploring the insights they hold. Comparing with document similarity measure, the citation and co-citation analysis was another clustering approach, and Small et al.’s empirical study stood on the entire science and technology domains and related more to the national R&D strategy and the science policy.

Methodological Approach

This study develops a methodology which contains a data pre-processing approach, a K-Means-based clustering analysis approach, and a trend analysis approach, and uses NSF Award 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 Figure 1, which has 3 main steps.

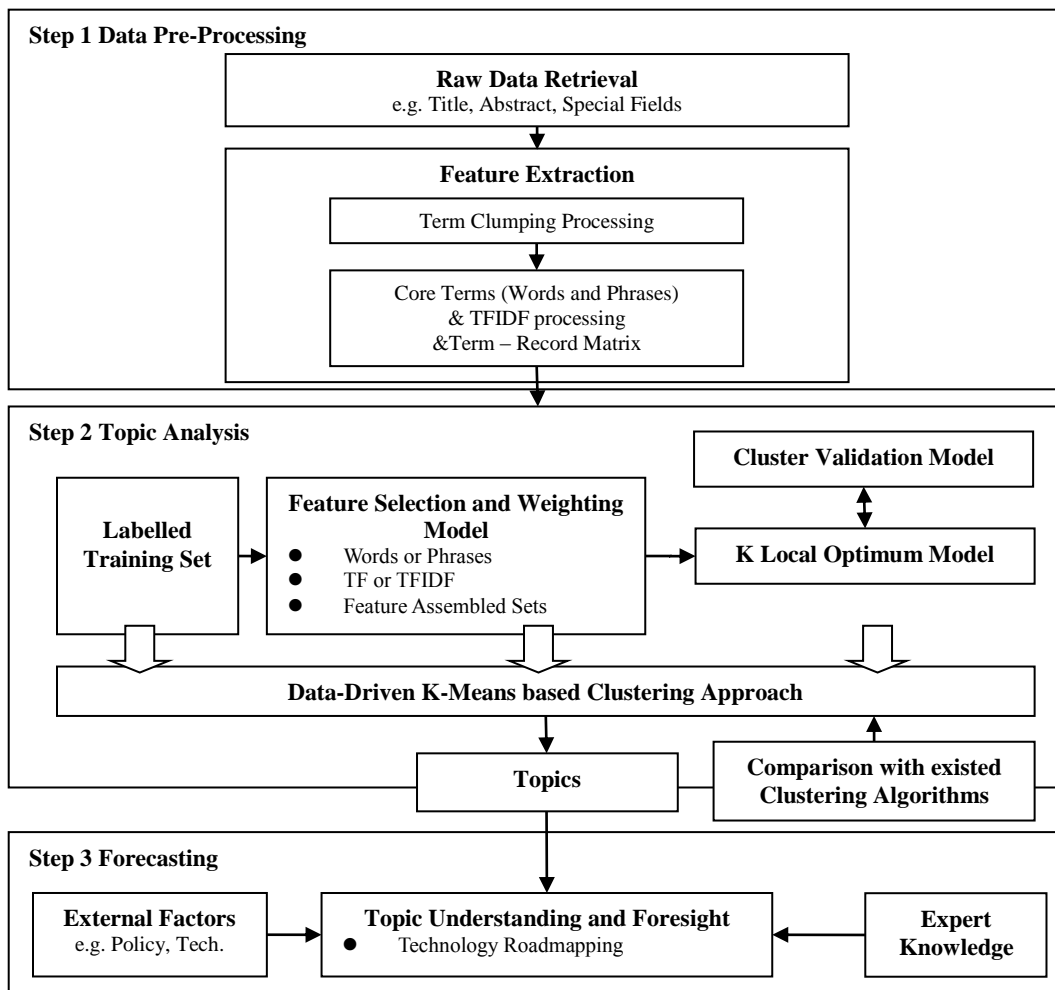


Figure 1. Research Framework for ST&I Textual Data-Oriented Topic Analysis and Forecasting

Step 1 Data Pre-Processing

Normally, ST&I textual data has common fields (e.g. Title, Abstract) and special ones (e.g. International Patent Classification in patent data, Program Element/Program Reference in NSF Awards data). Our purpose, in this step, is to remove meaningless data and retrieve relevant fields from the 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. 2014b). 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 also compare the TFIDF analysis with normal term frequency value to look into the benefits, and the classical formula (Boyack et al. 2011) of the TFIDF analysis is described as below:

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

Step 2 Topic Analysis

This step sets up a training set of labelled 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, we compose the cluster validation model with three target values: Recall; Precision; and F Measure, which are defined as follows:

$$\begin{aligned} \text{Recall} &= \frac{\text{Number of Relevant Records Clustered to the Category}}{\text{Total Number of Relevant Records of the Category}} \\ \text{Precision} &= \frac{\text{Number of Relevant Records Clustered to the Category}}{\text{Total Number of Records Clustered to the Category}} \\ \text{F Measure} &= \frac{\text{Recall} \times \text{Precision} \times 2}{\text{Recall} + \text{Precision}} \end{aligned}$$

Generally, Recall denotes the fraction of the records that are relevant to the query that are successfully retrieved. Precision indicates the fraction of retrieved documents that are relevant to the find. F Measure combines both as the harmonic mean. Since the Recall value for the whole dataset is meaningless (all records have been clustered), we only calculate the Total Precision to evaluate the total number of correctly grouped records. In addition, we also calculate the Recall, Precision, and F Measure for each Category, and use the Average F Measure as another main target value.

$$\begin{aligned} \text{Average F Measure} &= \frac{\sum \text{F Measure of each Category}}{\text{Total Number of Categories}} \\ \text{Total Precision} &= \frac{\text{Number of Records clustered to the correct Category}}{\text{Total Number of Records in Training Set}} \end{aligned}$$

It is also necessary to mention that, in the Cluster Validation Model, we label all records with the real category of the Centroid that they are grouped into. Therefore, the selection of the Centroid is one sensitive issue that will influence the cluster validation process.

2) K-Local Optimum Model

The 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 the specified interval, and decides the best K-value in the interval based on its F Measure.

The main concept of K-Means is described as follows:

- A. Initialization: Select the top K records with the highest Euclid Length as the Centroid of K Clusters;

$$\begin{aligned} &\text{Let } tf_{in} \text{ as the frequency of term } t_i \text{ in Record } D_n \\ \text{Record - Term Vector: } &V(D_n) = \{tf_{1n}, tf_{2n}, \dots, tf_{(i-1)n}, tf_{in}\} \end{aligned}$$

$$\text{Euclid Length of Record } D_n: ELEN(D_n) = \frac{1}{\sqrt{\sum tf_{in}^2}}$$

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

$$\text{Similarity Value: } S(D_n, D_m) = \text{Cos}(V(D_n), V(D_m))$$

C. Centroid Refine: Calculate the Similarity between record and its cluster, set the record with the highest Similarity Value as the Centroid of this cluster;

Let Cluster $C = \{D_1, D_2, \dots, D_{l-1}, D_l\}$

$$\text{Similarity Value: } S(D_n, C) = \frac{\sum_{k=1}^l S(D_n, D_k)}{l}$$

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.

3) Feature Selection and Weighting Model

Considering the specified fields of the NSF Award data in our case study, this step will use NSF Awards as the sample to present our method. Title and Abstract (described as Narration in NSF Awards) are the most common fields used by text analysis, and, for NSF Award data, we also introduce the Program Element (PE) Code and Program Reference (PR) Code to 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, but sometimes they mislead to other direction (e.g. in the case of PR code, most codes 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 term, PE code, and PR code. 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 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 Code, and PR code 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 weight terms below:

$$V(D_n) = VN(D_n) + VT(D_n)$$

$VN(D_n)$ is the Term – Record Vector with only Narration Terms, while $VT(D_n)$ with only Title Terms

Let $T_N = VN(D_n) \cap VN(D_m)$ and $T_T = VT(D_n) \cap VT(D_m)$

$$\omega_N = \frac{T_T}{T_N + T_T}, \omega_T = \frac{T_N}{T_N + T_T}$$

$$\text{Weighted Similarity Value: } S_w(D_n, D_m) = \omega_N \times S(VN(D_n), VN(D_m)) + \omega_T \times S(VT(D_n), VT(D_m))$$

This model also attempts to compare another two topics that always attract researcher's interests in text analysis: the clustering accuracy of words and phrases, normal TF and TFIDF value.

According to common sense, comparing with single word, phrase is more specific and would help create a more accurate cluster, since the relations between phrases are more meaningful than what individual words have.

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, will be generated at the end of this step. We will also compare our clustering approach with two popular mainstream text clustering algorithms: LDA and Hierarchical Aggregative Clustering (HAC).

Step 3 Forecasting

In the past, based on terms, we proposed a semi-automatic Technology Roadmapping composing model (Zhang et al. 2013), and this paper will use topics to take the place of terms and locate them in the time series as topic trends. 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 only considered as objective evidence to assist the decision making of 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 remove distinct duplicate topics manually;
- 2) Send the topic list to domain experts for assessment. Mark a topic 1 for “interesting topic at that time”, 0 for “not interesting at that time”, and 0.5 for “not sure”;
- 3) Calculate the marks for each topic and obtain the ranking list;
- 4) Remove low-ranked items and meaningless topics, consolidate similar topics, and classify topics into their appropriate technology development levels, with the help of experts;
- 5) Locate topics on the visual map and address the understanding gained regarding relationships, development trends, and foresights.

Empirical Study

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

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 the most important government agency in the US for funding 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 (US NSF Website, see Reference). Understanding of NSF Award 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 works. 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 Award database is open access, and all data can be downloaded on the NSF’s website (US NSF Website, see Reference). All awards are classified according to a 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 Award data is labelled 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, etc.). Statistically, less than half of the NSF

Award data is labelled in detail or with any kind of “usable” classification, while others have a common or meaningless labels or no label at all. As a result, we have treated the NSF Award data as semi-labelled.

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 5010000 and 5090000. 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 Award data, we removed awards granting support for travel, summer school, and further education to arrive at a final total of 9,274 records. We then applied the Term Clumping steps (Zhang et al. 2014b) 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.

Table 1. Steps of Term Clumping Processing

	Step	# N.* Terms	# T.* Terms
1	9274 Records, with 9274 Titles and 8975 Narrations	-	-
2	Natural Language Processing via VantagePoint (see Reference)	254992	17859
3	Basic Cleaning with thesaurus	214172	16208
4	Fuzzy Matching	184767	15309
5	Pruning (Remove terms appearing only in one record)	42819	2470
6	Extra Fuzzy Matching	40179	2395
7	Computer Science based Common Term Cleaning	38487	2311
8	Deep Cleaning: Expert-aided Screening*	30037	-

*N. = Narration, *T. = Title;

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

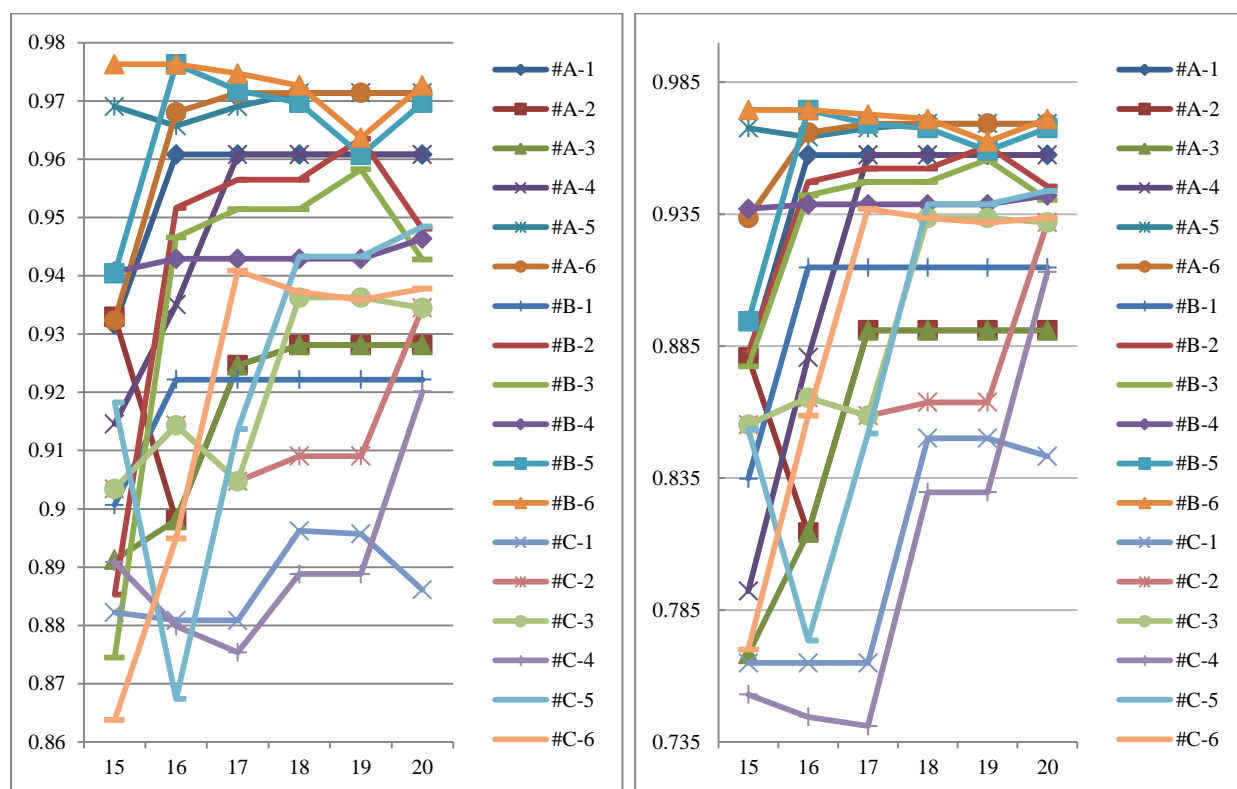
Before further processing, we dealt with the training set first. Given the NSF Award data is semi-labelled, we screened all 1,124 records in 2009, chose 10 categories (shown in Table 2) associated with 587 records, and established the training set, which included 369 Title Terms and 2,161 Narration Terms. 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 high-coupling, which is generally considered a big challenge for existing text clustering approaches (Yau et al. 2014).

Table 2. List of Ten Selected Categories

No.	Category	Record Num.	Note
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 the 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 Figure 2, and we also listed the maximum and mean of the Average F Measure and Total Precision of six Assemblies against Word & TFIDF, Phrase & TF, and Phrase & TFIDF in Table 3.



*#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.

Figure 2. Average F Measure (L) and Total Precision (R) of 6 Assemblies with TF and TFIDF

Table 3. Max and Avg. Value of F Measure and Total Precision of six Assemblies with Word & TFIDF, Phrase & TF, and Phrase & TFIDF

			#1	#2	#3	#4	#5	#6
#A PHRASE & TF	Average F Measure	Max	0.9608	0.9329	0.9281	0.9608	0.9714	0.9714
		Avg.	0.9246	0.9023	0.8681	0.9214	0.9526	0.9397
	Total Precision	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	Average F Measure	Max	0.9222	0.9635	0.9583	0.9463	0.9763	0.9763
		Avg.	0.9186	0.9436	0.9375	0.9431	0.9648	0.9727
	Total	Max	0.9148	0.9608	0.9557	0.9421	0.9744	0.9744

	Precision	Avg.	0.9015	0.9401	0.9350	0.9390	0.9554	0.9710
#C WORD & TFIDF	Average F Measure	Max	0.8962	0.9345	0.9362	0.9202	0.9484	0.9408
		Avg.	0.8870	0.9125	0.9216	0.8906	0.9224	0.9184
	Total Precision	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

Before looking into Table 3 and Figure 2, one undertaking 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 average target values, like Average F Feature and Total Precision, are 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. (2014b), 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 assemblies, and the phrase-based ones worked better than those with words as shown in both Table 3 and Figure 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 the phrase and the single word

The 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 Award 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 the highest values of both the Average F Measure and the Total Precision. This helps to claim that the TFIDF analysis makes good sense for feature extraction, which weights the common terms as lower value and is positive on cutting down the noise terms. However, an interesting exception still exists in #1 and #4, the only two assemblies without PE/PR code. We will discuss the significance of the PE/PR code 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 into 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 between the six assembled sets of feature combinations

PE and PR codes can be treated as the keywords of publications. They are special and meaningful, but far fewer terms originate from them than those from the narration and title. Thus, it is beneficial to engage with 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 the difference between them, and discovered that the PE code acted as the main keyword for a proposal, while the PR Code contained a large amount of noise information which might obfuscate the relationship between 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”, “Science of Science Policy”. This may explain the reason that #2 and #A-5 is slightly ahead of #3 and #A-6 respectively. Additionally, if we also consider the TFIDF, 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 result of #A-5.1 was that the highest F Measure was 0.91436 and the highest Total Precision was 0.873935, both of which are 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.

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 assembles.

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.

Comparison with the Two Text Clustering Algorithms

Aiming to demonstrate the efficiency and adaptability of our approach with NSF Award data, we used a Cluster Validation Model to compare our results with those derived from the LDA approach and the HAC approach.

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 the single word, 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 4.

Table 4. The Avg. F Measure and Total Precision of the Results in the LDA Approach

Topic Number	Term Clumping-cleaned Phrase						Raw Content of Title and Narration						
	#15	#16	#17	#18	#19	#20	#15	#16	#17	#18	#19	#20	#50
Avg. F Measure	0.4697	0.4928	0.4816	0.524	0.486	0.5661	0.5154	0.5703	0.5164	0.5157	0.4315	0.5666	0.4963
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

We concluded that the LDA approach, which effective for text clustering, has increased efficacy with single words, large data sets and low-coupling domains. The LDA approach is single-word based and therefore term frequency is its most important factor, however, removing and consolidating terms during the Term Clumping process drastically reduces the term frequency and renders the cleaning process pointless. Our training set only contained 587 records with approximately 3,000 single words, which just met the bottom-line of the LDA’s requirement on the data amount. As we mentioned above, the training set involved with 10 high-coupling sub categories in the computer science-related domain, which also increased 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 in 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.

- Initialization: to set each record as a cluster;

- **Iteration:** to calculate the similarity between each cluster and group the two clusters with the highest similarity value;
- **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 is shown in Table 5, and it is obvious that the HAC approach had better efficacy with large topic number 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 about more than hours running this basic and raw HAC approach for the small 587-record data set while our approach only took several minutes to generate the results.

Table 5. The Avg. F Measure and Total Precision of the Results in the HAC Approach

	Raw Phrase			Term Clumping-cleaned Phrase		
	0.1	0.05	0.035	0.1	0.05	0.035
Topic Number	58	29	20	58	29	20
Avg. F Measure	0.7375	0.7289	0.6840	0.8415	0.8277	0.8021
Total Precision	0.7462	0.7428	0.7172	0.8556	0.8391	0.8187

Step 3 Forecasting

We applied our method to NSF Award datasets from 2009 to 2013, numbering approximately 1000 each year and 4,847 in total. After the K-Means clustering approach, we obtained 75 topics, consolidated six duplicate topics, and retrieved 69 topics for further processing. As mentioned in the Methodology section, we sought to combine quantitative and qualitative methods for forecasting studies and treated our auto-generated results as objective evidence for decision-making. 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 School of Software, University of Technology Sydney, Australia, were invited as our panel. Based on their research experience and deep academic understanding, they helped us to consolidate similar topics and confirm whether the topics generated by our Text Clustering analysis were interesting or not.

We used an inverse ratio to weight the 4-Researcher Group and 5-PhD-Candidate Group, then removed all topics marked below ‘not sure’ (Rank 0.5) resulting in 54 remaining topics. Referring to the Technology Roadmapping composing model (Zhang et al. 2013), two experts from the 5-PhD-Candidate Group then helped us to classify these 54 topics into three levels of Technology Development Phases: Basic Research, Assistant Instrument, and System and Product. We listed parts of final topics (all 11 topics in 2009, and 10 topics in 2013) in Table 6. It is necessary to mention that assessments on the level of the technology development phase are subjective. In an attempt to reduce this kind of bias, we set up a 0.5 amendatory value between nearby levels. For example, 2.5 means the topic is based on “Assistant Instrument” but is also close to “System and Product”. This kind of correction is reflected as the location of the topic on the Y axis. In addition, in 2013, we consolidated four Big-Data-related topics and packaged them into one, considering the significance of this topic.

The visual Technology Roadmapping for trend analysis is shown in Figure 3.

Table 6. Interesting Topics of 2009 and 2013 Selected by Experts

Year	Topic	Topic Description	Rank	Level
2009	Adaptive Grasping	Adaptive Grasping, Automatic Speech Recognition, Empirical Mechanism Design, Hierarchical Visual Categorization	0.9306	2.0
	Behavior Modeling	Behavior Modeling, Human Centered Computing, Citizen Science, Dynamic Environments	0.8861	1.5
	Online Social Networks	Large Scale, Online Social Networks, Applications, Measurement	0.8417	2.5
	High Performance Computer	High Performance Computer, Major Research Instrument, Certifiably Dependable Software, Consortium,	0.7722	3.0

	Multi-Dimensional Brain Imaging Data	Comparative Analysis, Multi-Dimensional Brain Imaging Data, Coordinated Visualization, Multi Measure	0.7278	1.0
	Bayesian Model Computation	Bayesian Model Computation, Graphical Models, High Dimensional Data Sets, Information Integration	0.7278	1.0
	Cyber Physical Device	Cyber Physical Device, Cyber Physical System, Research Instrument, Service Attacks, Trust Worthy Cyber	0.7528	2.5
	Multi-Core Processors	Accelerator, Multi-Core Processors, Algorithms	0.7472	2.5
	Large Scale Data Centers	Large Scale Data Centers, Management, Parallel Programs	0.7722	1.0
	ad Hoc Wireless Networks	ad Hoc Wireless Networks, Cooperative Beam Forming, Cross Layer Optimization, Data Centers	0.6139	2.0
	Biological Networks	Biological Networks, Algorithm Foundation, Communication Networks, Cryptography	0.5694	2.0
2013	Big Data	Big Data, Computer Science Principles, Machine Learning, Large Scale Neural Networks, Complex Time Series Data	1	2.5
	Robotic Intelligence	Robotic Intelligence , High End Computer Users, Robotics Engineering, Sustainable Future, Software Needs	1	3.0
	Ethical Complexities	Ethical Complexities, Future Computer Professionals, Cyber Security, Diverse World	0.9556	2.0
	Joint Source Channel Codes	Joint Source Channel Codes, Two Dimensional Recording Systems, Topological Interference Management	0.7722	2.5
	Large Scale Hydrodynamic Brownian Simulations	Large Scale Hydrodynamic Brownian Simulations, Parallel Structured Adaptive Mesh Refinement Calculations, Coarray Fortran	0.8861	2.5
	Automatic Graphical Analysis	Automatic Graphical Analysis, Intuition Wall, Matrix Free Algorithms, Program New Computer Architectures, Predictable Datacenters	0.8167	2.0
	Supporting Knowledge Discovery	Supporting Knowledge Discovery, Scientific Visualization Language, Learning Causal Structure	0.7028	2.0
	Asynchronous Learning Experiences	Asynchronous Learning Experiences, Earthquake, Learning via Architectural Design	0.7278	2.5
	Scalable System Software	Scalable System Software, Optimal, Engineering Practice, Information	0.7972	2.5
	NSF Smart Health	NSF Smart Health, Health Influences, Learning Fine, Social Media, Data Centers	0.6833	2.5

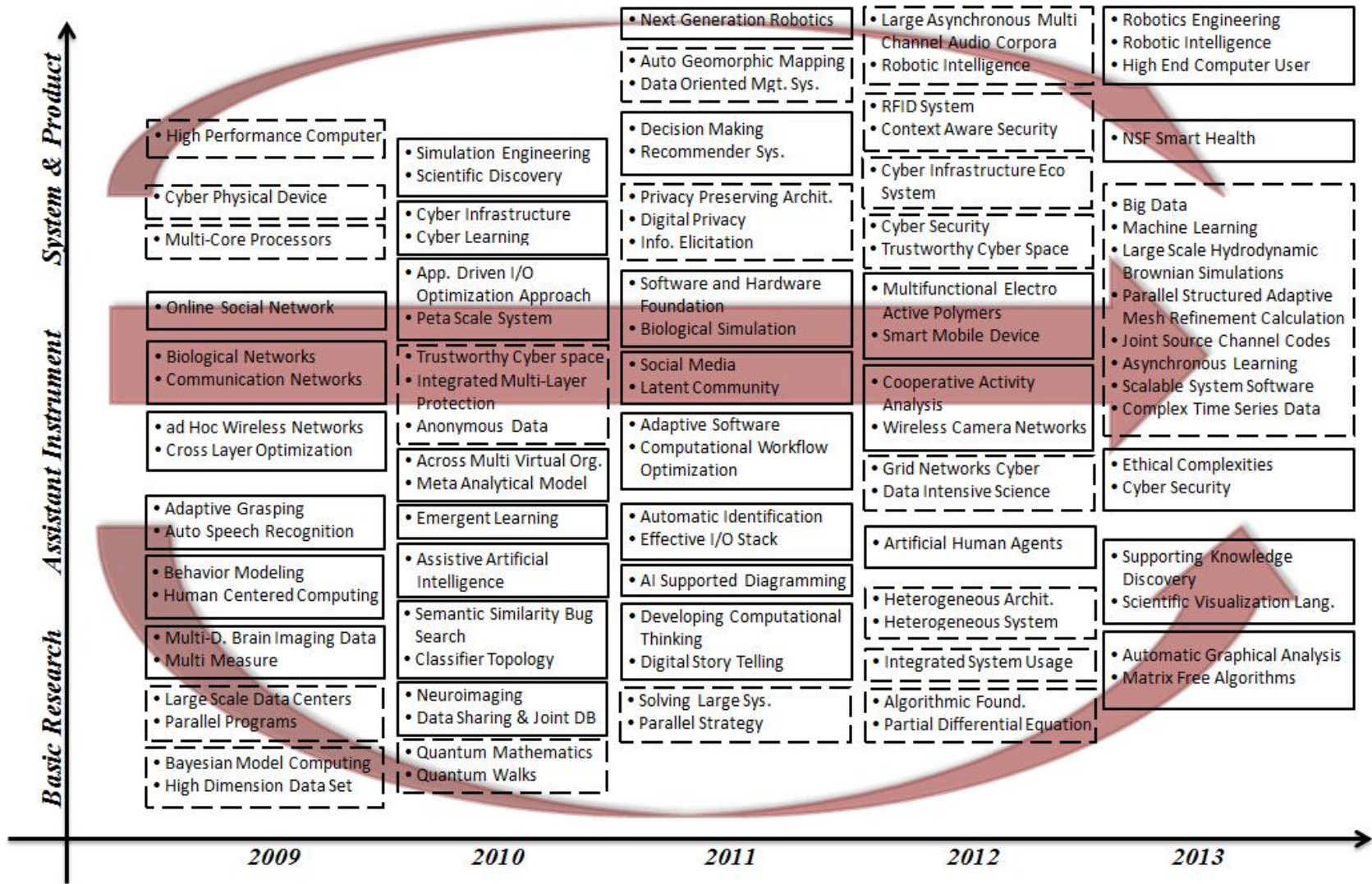


Figure 3. Technology Roadmapping for Computer Science (based on NSF Award Data)

Discussion and Implications

Currently considered one of the hottest topics, it is interesting and promising to explore and discuss the topic of ‘Big Data’ in more detail. On March 29, 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 US federal departments and agencies announced more than \$200 million US dollars to launch the initiative: the NSF being one of them. Figure 3 shows ‘Big Data’ was a hot topic in 2013, having evolved with various kinds of new techniques and concepts, but it is also easy to link these new ideas to their original ones (as indicated by solid boxes in Figure 3). “Large Scale” related concepts, algorithms, and systems have been generated since 2009, e.g. ‘Large Scale Data Centers (2009),’ ‘Solving Large System and Parallel Strategy (2011),’ ‘Large Asynchronous Multi Channel Audio Corpora (2012).’ So Big Data is not a total invention, but a kind of evolution from previous techniques and a solution for real-world problems. Almost all components can be traced back, but looking forward must be more valuable to us. Concerning the general technology development pathway and the situation of ‘Big Data’, we leave several comments here for the purpose of forecasting studies.

1) Outwardly, ‘Trustworthy Cyber Space (2010 and 2012),’ ‘Privacy Preserving Architecture and Digital Privacy (2011),’ and ‘Ethical Complexities (2013)’ seem to have no direct relationship to ‘Big Data’, especially its techniques, but in May 2014 the White House announced another report *Big Data: Seizing Opportunities Preserving Values* (The White House, 2014), which involved the relationships between government, citizens, businesses, and consumers and 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. Therefore, it is reasonable to imagine that, in the near future, ‘Privacy in Big Data’ and the corresponding privacy protecting techniques should be a big concern for both government and citizens in policy development and legal domains.

2) Another set of topics that attracts our eye is ‘Across Multi Virtual Organization (2010),’ ‘Automatic Identification (2011),’ ‘Supporting Knowledge Discovery (2013),’ and also those real application related topics, including ‘Biological Networks (2009),’ ‘Cyber Infrastructure (2010),’ ‘Wireless Camera Networks (2012),’ and ‘RFID System (2012).’ As the most powerful competitor of the US, China has issued ‘Internet of Things’ as its Top 5 Emerging Industry, announced firstly in the Speech *Let Science and Technology Lead China’s Sustainable Development* by the then current Premier 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 its related techniques and cyber security issues, must be another hot research topic in the coming decades.

3) It has been a long time since people started to imagine intelligent robots. Although these topics are not new, and appear several times in Figure 3, e.g. ‘Next Generation Robotics (2011),’ ‘Robotic Intelligence (2012),’ and ‘Robotics Engineering (2013),’ we still address positive foresight on robotic techniques, which must be able to gain more intelligence from Big Data and upgrade into a smarter format.

4) As part of Obama Administration’s Big Data program, the NSF started its *NSF Smart Health and Wellbeing* program in 2012, which “seeks improvements in safe, effective, efficient, equitable, and patient-centred health and wellness services through innovations in computer and information science and engineering” (NSF Website, see Reference). Although there is no direct topic related to ‘health and 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 must be an emerging industry for a long time to come.

Conclusion and Further Study

In the current Age of Big Data, it is common sense to transfer traditional method-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) the introduction of a local K optimum model and a feature selection

and weighting model to the classical K-Means model to solve the K-value problem and adapt the data feature assemblies; 2) we apply our method to NSF Award data and focus on computer science-related techniques for topic analysis and forecasting studies. This empirical study dives into the origins of scientific innovation and draws reference for studies on other mature technologies or even emerging technologies.

We anticipate further study in three directions: 1) to continue to improve our clustering algorithm by comparing it with other text clustering approaches, and to make it more operable, adaptable, and effective; 2) to extend the empirical study to cover multiple data sources, like publications and patents, and to extend our scope to broader innovation processes; and 3) to introduce a smarter quantitative approach for an automatic or semi-automatic TRM generation model. We believe this will hold interest for government, industry, and researchers.

Acknowledgement

This work is supported by the Australian Research Council (ARC) under discovery grant DP140101366 and the National High Technology Research and Development Program of China (Grant No. 2014AA015105).

References

- Aizawa, A. (2003). An information-theoretic perspective of tf-idf measures. *Information Processing & Management*, 39(1), 45-65.
- Allan, J., Carbonell, J., Doddington, G., Yamron, J., & Yang, Y. (1998). Topic Detection and Tracking Pilot Study Final Report, in *Proceedings of the DARPA Broadcast News Transcription and Understanding Workshop*.
- 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*. Cambridge, MA: MIT Press.
- Begelman, G., Keller, P., & Smadja, F. (2006). Automated tag clustering: Improving search and exploration in the tag space. In *Collaborative Web Tagging Workshop at WWW2006* (pp. 15-33). Edinburgh, Scotland.
- Beil, F., Ester, M., & Xu, X. (2002). Frequent term-based text clustering. In *Proceedings Of The Eighth ACM SIGKDD International Conference On Knowledge Discovery and Data Mining* (pp. 436-442). ACM SIGKDD. Edmonton, Canada.
- Bengisu, M. (2003). Critical and emerging technologies in materials, manufacturing, and industrial engineering: A study for priority setting. *Scientometrics*, 48(3), 473-487.
- Blei, D. M. (2012). Probabilistic topic models. *Communications of the ACM*, 55(4), 77-84.
- Blei, D. M., & Lafferty, J. D. (2006). Dynamic topic models. In *Proceedings Of The 23rd International Conference On Machine Learning* (pp. 113-120). ACM ICML. Pittsburgh, USA.
- Boyack, K. W., Newman, D., Duhon, R. J., Klavans, R., Patek, M., Biberstine, J. R., et al. (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 Quarterly*, 56(1), 75-86.
- Cataldi, M., Di Caro, L., & Schifanella, C. (2010). Emerging topic detection on twitter based on temporal and social terms evaluation. In *Proceedings of the Tenth International Workshop on Multimedia Data Mining* (p. 4). ACM MDM/KDD. Washington DC, USA.
- Chang C. H., & Hsu, C. C. (1997). Customizable multi-engine search tool with clustering. *Computer Networks and ISDN Systems*, 29(8), 1217-1224.
- Chen, H., Zhang, G., & Lu, J. (2013). A time-series-based technology intelligence framework by trend prediction functionality. In *Systems, Man, and Cybernetics*, 2013 IEEE International Conference on (pp. 3477-3482). IEEE SMC. Manchester, UK.
- Chen, H., Zhang, G., Lu, J., & Zhu, D. (2014). A Two-Step Agglomerative Hierarchical Clustering Method for Patent Time-Dependent Data. *Foundations and Applications of Intelligent Systems* (pp. 111-121). Berlin Heidelberg: Springer.
- Cutting, D. R., Karger, D. R., Pedersen, J. O., & Tukey, J. W. (1992). Scatter/gather: A cluster-based approach to browsing large document collections. In *Proceedings Of The 15th Annual International ACM SIGIR Conference On Research And Development In Information Retrieval* (pp. 318-329). ACM SIGIR. Copenhagen, Denmark.

- Dai, X. Y., Chen, Q. C., Wang, X. L., & Xu, J. (2010). Online topic detection and tracking of financial news based on hierarchical clustering. In *Machine Learning and Cybernetics*, 2010 International Conference on (Vol. 6, pp. 3341-3346). IEEE ICMLC. Bangalore, India.
- 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 Transactions on Intelligent Systems and Technology (TIST)*, 3(2), 23.
- Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., & Witten, I. H. (2009). The WEKA data mining software: an update. *ACM SIGKDD Explorations Newsletter*, 11(1), 10-18.
- Huang, A. (2008). Similarity measures for text document clustering. In *Proceedings Of The Sixth New Zealand Computer Science Research Student Conference* (pp. 49-56). NZCSRSC2008. Christchurch, New Zealand.
- 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. *Technological Forecasting and Social Change*, 81, 39-48.
- Jain, A. K. (2010). Data Clustering: 50 Years beyond K-Means. *Pattern Recognition Letters*, 31(8), 651-666.
- Kontostathis, A., Galitsky, L. M., Pottenger, W. M., Roy, S., & Phelps, D. J. (2004). A survey of emerging trend detection in textual data mining. In *Survey of Text Mining* (pp. 185-224). New York: Springer.
- Kostoff, R. N., Toothman, D. R., Eberhart, H. J., & Humenik, J. A. (2001). Text mining using database tomography and bibliometrics: A review. *Technological Forecasting and Social Change*, 68(3), 223-253.
- Lu, N., Zhang, G., & Lu, J. (2014). Concept drift detection via competence models. *Artificial Intelligence*, 209, 11-28.
- McAfee, A., & Brynjolfsson, E. (2012). Big data: the management revolution. *Harvard business review*, (90), 60-6.
- Newman, N. C., Porter, A. L., Newman, D., Trumbach, C. C., & Bolan, S. D. (2014). Comparing methods to extract technical content for technological intelligence. *Journal of Engineering and Technology Management*, 32, 97-109.
- Porter, A. L., & Cunningham, S. W. (2004). *Tech mining: exploiting new technologies for competitive advantage*. New York: John Wiley & Sons.
- Porter, A. L., & Detampel, M. J. (1995). Technology opportunities analysis. *Technological Forecasting and Social Change*, 49(3), 237-255.
- Small, H., Boyack, K. W., & Klavans, R. (2014). Identifying emerging topics in science and technology. *Research Policy*, 43(8), 1450-1467.
- The White House. (2012). *Big data is a big deal*. <http://www.whitehouse.gov/blog/2012/03/29/big-data-big-deal>. Accessed October 12, 2014.
- The White House. (2014). *Big Data: Seizing Opportunities Preserving Values*. http://www.whitehouse.gov/sites/default/files/docs/big_data_privacy_report_may_1_2014.pdf. Accessed October 24, 2014.
- United States National Science Foundation, <http://www.nsf.gov/>. Accessed October 12, 2014.
- VantagePoint, <http://www.theVantagePoint.com>. Accessed October 12, 2014.
- Voorhees, E. M. (1986). Implementing agglomerative hierarchic clustering algorithms for use in document retrieval. *Information Processing & Management*, 22(6), 465-476.
- Wen, J. (2009). *Let Science and Technology Lead China's Sustainable Development*. <http://www.chinanews.com/gn/news/2009/11-23/1979809.shtml>. Accessed October 24, 2014.
- Wu, H. C., Luk, R. W. P., Wong, K. F., & Kwok, K. L. (2008). Interpreting tf-idf term weights as making relevance decisions. *ACM Transactions on Information Systems (TOIS)*, 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. In *Proceedings of the 22nd ACM international conference on Conference on information & knowledge management* (pp. 99-108). ACM CIKM. San Francisco, USA.
- Yau, C., Porter, A., Newman, N. (2014). Clustering scientific documents with topic modelling. *Scientometrics*, 100,767-786.
- 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. *Technology Analysis & Strategic Management*, 25(6), 707-724.
- Zhang, Y., Zhou, X., Porter, A. L., Gomila, J. M. V., & Yan, A. (2014a). 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., Porter, A. L., Hu, Z., Guo, Y., & Newman, N. (2014b). "Term clumping" for technical intelligence: A case study on dye-sensitized solar cells. *Technological Forecasting and Social Change*, 85, 26-39.

Zhu, D., & Porter, A. L. (2002). Automated extraction and visualization of information for technological intelligence and forecasting. *Technology Forecasting & Social Change*, 69, 495–506.