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

Contents lists available at ScienceDirect

Technological Forecasting & Social Change

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

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

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

ARTICLE INFO

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

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

ABSTRACT

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

1. Introduction

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

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

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

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

* Corresponding author at: Decision Systems & e-Service Intelligence Research Lab, Centre for Quantum Computation & Intelligent Systems, Faculty of Engineering and Information Technology, University of Technology Sydney, Australia.
E-mail addresses: yizhang.bit@gmail.com (Y. Zhang), guangquan.zhang@uts.edu.au (G. Zhang), Hongshu.Chen@student.uts.edu.au (H. Chen), alan.porter@isye.gatech.edu (A.L. Porter), zhudh111@bit.edu.cn (D. Zhu), Jie.Lu@uts.edu.au (J. Lu).

http://dx.doi.org/10.1016/j.techfore.2016.01.015
0040-1625/

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

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

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

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

2. Related works

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

2.1. Text clustering

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

2.2. Topic analysis

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

2.3. Technology Roadmapping

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

2.4. Comparison with related work

Based on a 2.15-million-MEDLINE-publication dataset, Boyack et al. (2011) presented an outstanding comparison study on several text-based similarity approaches, e.g., TFIDF, Latent Semantic Analysis, Topic Models, BM25, and PubMed's own Related Articles (PMRA) approach. The study covered almost all mainstream text clustering algorithms and included a detailed discussion summarizing the advantages