J. Eng. Technol. Manage. 32 (2014) 97–109

Contents lists available at ScienceDirect

Journal of Engineering and Technology Management

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

Comparing methods to extract technical content for technological intelligence

Nils C. Newman [a,*], Alan L. Porter [b,c], David Newman [d], Cherie Courseault Trumbach [e], Stephanie D. Bolan [c]

[a] IISC, P.O. Box 77691, Atlanta, GA 30357, USA
[b] Search Technology, Inc., Atlanta, GA 30332-0345, USA
[c] Georgia Institute of Technology, Atlanta, GA 30332-0345, USA
[d] University of California, Irvine, Irvine, CA 92697, USA
[e] University of New Orleans, 2000 Lakeshore Dr, New Orleans, LA 70148, USA

ARTICLE INFO

Article history:
Received 24 June 2013
Received in revised form 30 August 2013
Accepted 2 September 2013

JEL classification:
C
O

Keywords:
Tech mining
Topic modeling
Term clustering
Technological emergence
Dye-sensitized solar cells

ABSTRACT

We are developing indicators for the emergence of science and technology (S&T) topics. To do so, we extract information from various S&T information resources. This paper compares alternative ways of consolidating messy sets of key terms [e.g., using Natural Language Processing on abstracts and titles, together with various keyword sets]. Our process includes combinations of stopword removal, fuzzy term matching, association rules, and term commonality weighting. We compare topic modeling to Principal Components Analysis for a test set of 4104 abstract records on Dye-Sensitized Solar Cells. Results suggest potential to enhance understanding regarding technological topics to help track technological emergence.

Abbreviations: DSSCs, dye sensitized solar cells; LSA, latent semantic analysis; PCA, principal components analysis; TF/IDF, term frequency/inverse-document-frequency.

* Corresponding author. Tel.: +1 678 296 2287.
E-mail addresses: newman@iisco.com (N.C. Newman), aporter@searchtech.com (A.L. Porter), newman@uci.edu (D. Newman), ctrumbac@uno.edu (C.C. Trumbach).

0923-4748/$ – see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.jengtecman.2013.09.001

Introduction

Tracking technologies or trying to determine their state has always been a challenging task. The globalization of research adds to the difficulty. In the past, analysts primarily used expertise augmented by literature review to assess the state of development of a technology of interest. However, the increasing availability of electronic information about technology opens up new possibilities to facilitate this process. Since the early 1990s researchers at the Technology Policy and Assessment Center at the Georgia Institute of Technology have been investigating the use of text mining to aid in assessment and forecasting of technologies (e.g., Watts et al., 1997, 1998; Watts and Porter, 1999, 2003, 2007; Watts et al., 1999, 2004; Zhu et al., 1999; Zhu and Porter, 2002). This research is based on the premise that digital records (bibliographic journal abstracts, full text journal articles, conference proceedings, etc.) can be effectively text mined and that the results of that mining can help determine the state of a technology. This "Tech Mining" process is covered in detail in the book by Porter and Cunningham (2005).

The Tech Mining process combines bibliometrics and text analyses of Science, Technology and Innovation (STI) information resources. The rationale for pursuing this is the premise that Management of Technology (MOT) decision processes can benefit from empirical indicators to complement expertise. Porter and Cunningham (2005) identify 39 MOT questions that Tech Mining can help address, but a more succinct set are simply: Who, When, Where, and What? [The other two so-called "reporter's questions" – How and Why? – almost always require more human insight.] Who, when, and where interests are relatively straightforward to address by careful treatment of bibliographic record fields – e.g., who (authors, inventors, patent assignees), when (article publication or patent grant dates), and where (inventor or author address). Software can readily tally frequencies of such elements across a search set (e.g., patent abstract records concerning solar cells) to identify the leading organizations and trends. That is not to say that serious text analysis is not needed, it is – to extract organizational identities from address strings or to disambiguate author identities, for instance (Tang and Walsh, 2010).

The "what" question is far more challenging. Some fielded records contain helpful content, such as keywords in paper abstracts and classification codes in paper or patent abstract records. However, these tend to lag frontier developments as terminology emerges, so warrant enrichment to extract topical content, especially the noun phrases or words from titles, abstracts, claims, or full text. Additional approaches may introduce terms of special interest to ascertain their prevalence (over time; by key R&D players). Aims include identification of topics that show a marked upsurge in R&D attention in the most recent time period – i.e., "hot topics." Compilation of "new topics" in recent times can also help identify novel interests within a field by presenting them to field experts to scan for potentially emergent topics to pursue. The ultimate motivation is that such methods can be used to inform MOT judgments.

The process that evolved at Georgia Tech over 20+ years of development uses the output of text mining to good effect, but, overall, the techniques employed in the Tech Mining process still require a significant amount of analyst judgment, as well as expertise in text mining techniques (Porter and Cunningham, 2005). One key research question today is: Can recent advances in text analysis be leveraged to increase the level of automation in Tech Mining so the analyst can focus more on the question and less on the process? To this end, this paper looks at two techniques. The first is a sequence of "term clumping" steps to consolidate topical information. This technique represents a set of incremental engineering improvements on existing processes. The second approach uses Topic Modeling, which represents a more radical shift through the introduction of new algorithms. The study uses Dye-Sensitized Solar Cells as an example case. The comparison is carried out by two teams – the Tech Mining team at Georgia Tech applying Natural Language Processing (NLP) with Principal Components Analysis (PCA) and the Topic Modelers at UC Irvine.

Background

Over the years many techniques have been used to model the research indexed in technology databases. This is done to analyze the structure of technical domains and enable analysts to solve