Technological Forecasting & Social Change 85 (2014) 26–39

Contents lists available at ScienceDirect

Technological Forecasting & Social Change

"Term clumping" for technical intelligence: A case study on dye-sensitized solar cells

Yi Zhang [a], Alan L. Porter [b,c], Zhengyin Hu [d], Ying Guo [a,*], Nils C. Newman [e]

[a] School of Management and Economics, Beijing Institute of Technology, Beijing, China
[b] Technology Policy and Assessment Center, Georgia Institute of Technology, Atlanta, GA 30332, USA
[c] Search Technology, Inc., Norcross, GA 30092, USA
[d] Chengdu Branch of the National Science Library, Chinese Academy of Sciences, Beijing, China
[e] Intelligent Information Systems Corporation (IISC), Norcross, GA 30092, USA

ARTICLE INFO

Article history:
Received 6 September 2012
Received in revised form 20 December 2013
Accepted 27 December 2013
Available online 28 January 2014

Keywords:
Term clumping
Dye-sensitized solar cells
DSSCs
Tech mining
Technical intelligence
Text clustering
Text analytics

ABSTRACT

Tech Mining seeks to extract intelligence from Science, Technology & Innovation information record sets on a subject of interest. A key set of Tech Mining interests concerns which R&D activities are addressed in the publication and patent abstract records under study. This paper presents six "term clumping" steps that can clean and consolidate topical content in such text sources. It examines how each step changes the content, potentially to facilitate extraction of usable intelligence as the end goal. We illustrate for an emerging technology, dye-sensitized solar cells. In this case we were able to reduce some 90,980 terms & phrases to more user-friendly sets through the clumping steps as one indicator of success. The resulting phrases are better suited to contributing usable technical intelligence than the original results. We engaged seven persons knowledgeable about dye-sensitized solar cells (DSSCs) to assess the resulting content. These empirical results advanced the development of a semi-automated term clumping process that can enable extraction of topical content intelligence.

1. Introduction

Over the last twenty years, Georgia Tech's Technology Policy and Assessment Center has been pursuing the development of variants of our "Tech Mining" approach to retrieving usable information on the prospects of particular technological innovations from Science Technology and Innovation (ST&I) resources. We have conducted ST&I analyses aimed especially to generate competitive technical intelligence (CTI) since the 1970s and have included software development to facilitate mining of abstract records in our research since 1993 [1–3]. Our colleagues have explored ways to expedite such text analyses, c.f. [4,5], as have others [6]. We increasingly turn toward extending such "research profiling" to aid in Forecasting Innovation Pathways (FIP) [7].

We focus on processing search results from ST&I databases that typically yield thousands of records. Such searches provide terms that can indicate significant topics during the emergence of a technology. However, those term sets (about 5000 publications), as in our case, can easily approach 100,000 items after Natural Language Processing (NLP) to extract noun phrases, making analysis challenging. Herein, we are trying to enable faster and better Tech Mining by processing that topical content. We attempt to construct a term clumping model for term cleaning, consolidation, and clustering. Different from existing approaches (we will discuss previous work in the literature review), our approach emphasizes the construction of term clumping steps from term cleaning to term consolidation and then to term clustering. We further extend traditional term clumping concepts with "Combine Terms Network," "Term

Abbreviations: CTI, Competitive technical intelligence; DSSCs, Dye-sensitized solar cells; LSI, Latent Semantic Indexing; NLP, Natural Language Processing; PCA, Principal Components Analysis; ST&I, Science, Technology & Innovation; WoS, Web of Science (including Science Citation Index).

* Corresponding author.
E-mail addresses: yi.zhang.bit@gmail.com (Y. Zhang), alan.porter@isye.gatech.edu (A.L. Porter), huzy@clas.ac.cn (Z. Hu), violet7376@gmail.com (Y. Guo), newman@iisco.com (N.C. Newman).

0040-1625/$ – see front matter © 2014 Elsevier Inc. All rights reserved.
http://dx.doi.org/10.1016/j.techfore.2013.12.019

Frequency Inverse Document Frequency (TFIDF) Analysis," and other purposive approaches [e.g., TRIZ (a concept for inventive problem solving that will be combined with semantic studies and bibliometric methods for system component understanding) and Technology Roadmapping (a graph to visually describe technology development trends along the time axis)]. We also pay attention to the use of automated macros in VantagePoint [1] for term clumping.

In this paper, we focus on abstract record search results that pertain to a particular technology of interest and will serve as source to profile R&D and forecast potential innovation paths. Drawing on text mining and bibliometric methods, this paper approaches "term clumping" as an inductive method; we are also interested in deductive approaches wherein we import target terms—e.g., using TRIZ to identify innovation prospects [8,9]. The aim here is to explore the methods of cleaning and consolidating large sets of topical phrases in order to generate better topical phrases for further analyses. In particular, compared with single qualitative (e.g., expert interview or workshop) or quantitative (e.g., statistical analysis) methods, we try out systematic software steps (e.g., VantagePoint; alternatively Thomson Data Analyzer provides similar functionality [1]) with varying degrees of human intervention. The human intervention can entail analyst data treatment (e.g., removing obvious noise) and/or topical expertise, but our aim is to devise a term clumping process that minimizes human effort. We want to concentrate analyst and expert attention on high-value activities, such as studying how those consolidated topics (concepts) change over time and their patterns of interaction. We believe such progress could expedite the generation of technical intelligence and advance efforts at Technology Roadmapping [10] (or FIP [7]).

This paper is organized as follows: Section 2 summarizes key literature, emphasizing ST&I analyses and term clumping. Section 3 describes our dye-sensitized solar cell data and inductive methods for "term clumping." Stepwise results are given to verify the practical value of this model in Section 4. Section 5 compares the top terms in different steps and also displays several selected samples to open up more "term clumping" stepwise details. Finally, we present expert assessment and conclusions in Section 6.

2 . Literature review

2.1 . ST&I text analyses

A research community has grown around bibliometric analyses of ST&I records over the past 60 years or so [11–13]. De Bellis has nicely summarized many facets of the data and their analyses [14]. To state the obvious—not all texts behave the same way. The language of the text and the venue for the discourse, with its norms, affect usage. Text mining needs to take such facets into consideration. In particular, we focus on ST&I literature and patent abstracts regarding it. In other analyses, we extend our analysis to business press and attendant popular press coverage of topics (e.g., Factiva or ABI Inform databases)—for example, also concerning dye-sensitized solar cells (DSSCs) [15,16,44]. English ST&I writing differs somewhat from "normal" English in structure and content. For instance, scientific discourse tends to include technical phrases that should be retained, not parsed into separate terms by Natural Language Processing (NLP). The VantagePoint NLP routine [1] applied here strives to do that and furthermore seeks to retain chemical formulas.

2.2 . Term clumping

As Bookstein discussed, the concept of clumping is similar to that of clustering, but clumping further concerns the objects' sequence and their adjacency properties [17]. He also classified term clumping into condensation measures and linear measures to evaluate "clumping strength" [18,19]. These approaches are based on statistical models of language use, such as term condensation, distribution over textual units, etc. Term clumping can help to distinguish the content-bearing words. It can also treat statistical properties of the words or phrases, considering semantic connections among terms [19]. Significantly, the Topic Detection and Tracking (TDT) model, defined by Allan et al., intends to explore techniques for detecting the appearance of new topics and for tracking their reappearance and evolution [20]. In research on extension of this model, Nallpati proposed a semantic language modeling approach that uses probabilistic methods for TDT with news stories [21].

Several of the term clumping steps that we treat here are basic. Removal of "stopwords" needs little theoretical framing. However, it does pose some interesting analytical possibilities. For instance, Cunningham found that common modifiers provided analytical value in classifying British science [22]. He conceives of an inverted U-shape that emphasizes analyzing terms of moderately high frequency—excluding both the very high frequency (stopwords and commonly used scientific words that provide high recall of records but low precision) and low-frequency words (suffering from low recall due to weak coverage but high precision). Pursuing this notion of culling common scientific words, we can remove "common words." In our analyses we apply a number of stopword lists of several hundred terms (including some stemming), and a thesaurus containing common words in academic/scientific writing consisting of some 48,000 terms [23]. We are interested in whether removal of these terms enhances or possibly degrades further analytical possibilities.

A variety of statistical techniques have been brought to bear on consolidating or clustering terms [24]. These offer the means to go well beyond consolidation of term variants, drawing upon semantic or syntactic associations. Various statistical methods [e.g., Principal Components Analysis (PCA), Latent Semantic Indexing (LSI), [25,26] and Latent Dirichlet Allocation (LDA) [27] or Topic Model] are available [28]. Contrasted with statistical methods, Pottenger and Yang introduced a neural network model to calculate the relations within the results of term co-occurrence analysis for emerging concepts detection [29]. However, all of these draw upon the pattern of co-occurrence of terms in records of the data set under scrutiny. In so doing, one seeks to group related concepts and thereby goes beyond the basic term clumping of like terms or phrases (e.g., those with shared words or slight spelling variations). In this paper, we focus on those basic term clumping operations and only then introduce PCA to further group related terms or phrases. (Note that other statistical approaches attempt the converse—seeking to group *records* [documents] based on commonalities in their term patterns.)