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.)

PCA, like LSI, which seeks to uncover the latent semantic structure in the data, uses Singular Value Decomposition (SVD) to transform the basic terms by document matrix to reduce ranks (i.e., to replace a large number of terms by a relatively small number of factors, capturing as much of the information value as possible). PCA eigen-decomposes a covariance matrix, whereas LSI does so on the term-document matrix. (See Wikipedia for basic statistical manipulations).

Herein, we use a special variant of PCA developed to facilitate ST&I text analyses [1]. This PCA routine generates a more balanced factor set than LSI (which first extracts a factor that explains the largest variance, then a second that best explains the remaining variance, etc.). The VantagePoint factor map routine applies a small-increment Kaiser Varimax Rotation (yielding more attractive results but running slower than SPSS PCA in developmental tests). Our colleague, Bob Watts of the U.S. Army, has led the development of a more automated version of PCA with an optimization routine to determine a best solution (maximizing inclusion of records with fewest factors) based on selected parameter settings—(Principal Components Decomposition—PCD) [30]. PCA is a basic form of factor analysis that allows terms to appear in multiple "factors" (we take the liberty to use that term in lieu of "principal components").

There are also several extended LSI methods, such as Probabilistic LSI, which constructs a statistical latent class model and is more principled [31]; an iterative scaling method, which gets higher precision of similarity measurement than SVD [32]; a Local Relevancy Ladder Weighted LSI (LRLW LSI) method, which improves text classification, [33] and so forth.

Researchers are combining term clumping with techniques such as PCA or LSI in order to retrieve synonymous terms from massive contents. For example, Xu et al. identified conceptual gene relationships from titles and abstracts with MEDLINE citations by LSI, [34] and Maletic & Marcus introduced LSI analysis to identify similarities and concept locations for program understanding [35,36]. Variations on such text analytics can help to identify concepts and relationships in various arenas, including web sites [37] and social media [38]. Shinmori et al. focused on terms' parts of speech for patent structure analysis and term explanation studies, [39] which strongly resemble what we did in the purposive method section of term clumping using semantic TRIZ and the Subject–Action–Object model [40].

We are comparing various term clumping and advanced statistical clumping techniques and combinations thereof. Elsewhere we consider topic modeling (LDA and variations) in more detail and compare treatment of a technical dataset (DSSCs) with a less technical one (concerning "management of technology") [41]. Newman et al. [42] compare the efficacy of alternative text analytics on DSSC data.

3 . Data and methods

3.1. Data

This study employs "term clumping" to take the necessary steps to clean and consolidate rich sets of topical phrases and terms. These steps are applied to a collection of documents relating to a topic of interest. In this case, we are addressing DSSCs. The present data derive from a multi-step Boolean search algorithm [43] adapted and applied via search interfaces to two leading, global ST&I databases—the Science Citation Index Expanded of Web of Science (WoS) and EI Compendex. Resulting abstract record sets were merged in VantagePoint, with duplicate records consolidated. The resulting 5784 publication abstracts are the focus of the present analyses. These cover the time span of 1991 (the inception of DSSC research) [44] through 2011 (not complete for this last year).

3.2 . Term clumping framework

We construct a framework for "term clumping" (Fig. 1), that includes record selection, field selection, text cleaning, consolidation of terms into informative topical factors, and expert engagement. We briefly treat record and field selection for the DSSC data and then elaborate with empirical detail on the term clumping steps for text cleaning and consolidation: Common Term Removal, Fuzzy Matching, Combining, Pruning, Screening, and Clustering. In the last section of the paper, we touch on expert engagement and various extensions building on these basic term clumping operations.

3.3 . Treatment of records and fields

Record selection is obviously essential to the analyses, but not our focus here. As mentioned, the present DSSC data derive from searches in WoS (a premier source of information on fundamental research) and Compendex (a leading R&D database emphasizing engineering and applied science). We have also searched and retrieved DSSC data from Derwent World Patent Index (DWPI) and from Factiva, but those are not addressed here.

For completeness, Fig. 1 includes consideration of record attributes. For certain analyses one wants to focus on particular record selections. For instance, one might choose to analyze the most cited records to recognize influential research. For extremely large domains, one may want to retrieve a sample—e.g., random or stratified. The present set is the full set resulting from the Boolean searches.

Time span is another dimension to consider. As noted, these records cover 1991–2011. Given that the search set for the year 2011 is not complete, one might choose to normalize the records to provide more interpretable trends (e.g., apply a correction factor to the most recent year counts). Often, we have a special interest in recent R&D activity to specifically examine "hot" topics.

The resulting document set consists of 5784 field-structured abstract records. That is, information is parsed into such fields as author, publication year, and abstract. Software, such as VantagePoint [1] used here, enables ready analysis of given record fields (e.g., to list the most prevalent authors) and to derive additional fields (e.g., to extract an author's country from an affiliation address field).

Our current attention is on topical content—i.e., which topics are currently pursued in the R&D activity described. In other analyses, we are keenly interested in finding out answers to the questions "who (i.e., organizations or individual researchers)?", "where (i.e., countries)?", and "when?" Especially valuable are analyses that address combinations of these elements—e.g., "who has recently been researching what?"

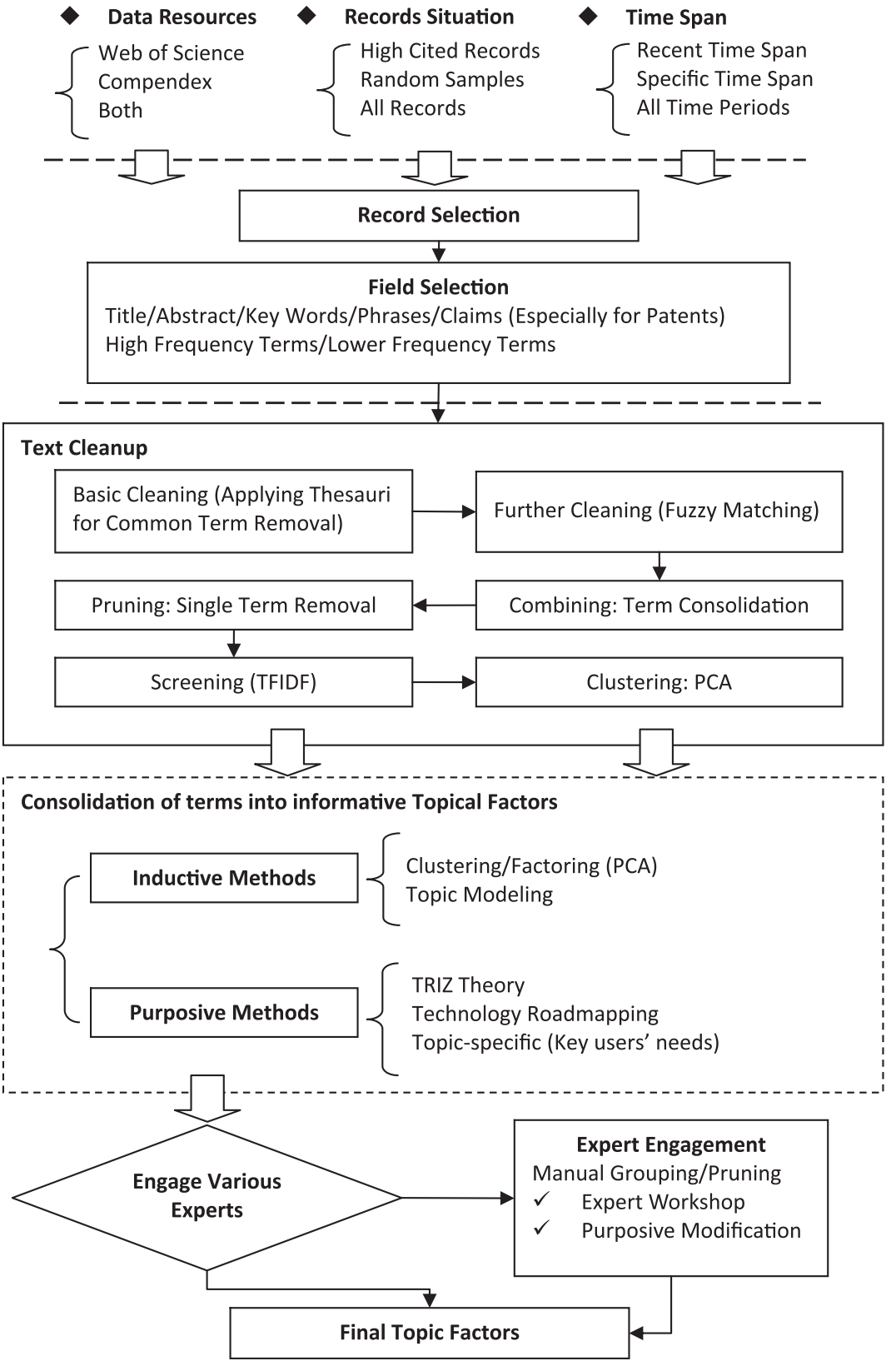

Fig. 1. Framework for term clumping.

Topical content is available in the WoS and Compendex variations of several fields:

- Title
- Abstract
- Keywords

Using VantagePoint's NLP routine, we extract noun phrases from the titles and from the abstracts. We also utilize the index terms (controlled vocabulary) from Compendex, and "Keywords Plus" from WoS. One could also utilize Compendex's classification codes. (If one were dealing with patent records, "Claims" would be another important source of topical information.) Here, we consolidate the resulting fields to obtain 90,980 terms and phrases in one merged field. Those provide the starting point for our term clumping steps.

4 . Stepwise results

4.1 . Text cleaning

We distinguish basic cleaning operations of common term removal and fuzzy term matching from later clumping operations. Note that the order in which to perform these operations is not set in stone and that some steps may need to be repeated in one form or another. Table 1 provides the stepwise tally of phrases in the merged topical fields undergoing term clumping. It is difficult to balance precision with clarity, so we hope this succeeds. The first column indicates which text analysis action was taken, corresponding to the list of steps (Fig. 1 and discussed below). The second column relates the results of application of the steps to the DSSC data.

Our starting list consists of 90,980 noun phrases and individual words (henceforth, usually called "phrases"). The noun phrases are an imperfect approximation as the automated NLP routine blends semantic and syntactic information to estimate where to separate term strings and which words to include. The NLP function has been modularized in VantagePoint; thus, we will not dive into this part too much and begin here with applying a number of thesauri containing various common term sets.

We have set up several "soft" rules in our term clumping steps, especially on how to determine the thresholds for many steps. For example, sometimes, we will pick up (or remove) the top 100 high-frequency terms for further analyses. Mostly, these actions are based on previous experience, experts' knowledge, or the software's best applicable requirements; thus, there is no strict threshold for these steps. We give more details as we follow the steps. In addition, we also repeat certain steps after some clumping; these seemingly redundant steps do add value (not too much, but they are easy to run).

4.1.1 . Step a. Applying thesaurus for common term removal

In this step, called "basic cleaning" in the framework, we apply thesauri for removal of common terms in three ways: 1) smooth the results derived from NLP, remove several HTML tags, and prepare for the cleaning; 2) remove (probably) meaningless terms starting with non-alphabetic characters; and 3) consolidate or remove the stopwords, common terms, and other trash terms.

First, we apply the XML encoding thesaurus, which removes codes such as <inf>, </inf>, ^{, and} to facilitate consolidation with plain English terms. This reduces the list from 90,980 to 82,746—a drop of 8234 (over 9%).

Presumably, almost all terms starting with non-alphabetic characters are meaningless to our research, such as "1.5 m/s," "1500°," etc. Therefore, we remove all of them; although several meaningful terms could thereby be lost. We run the "NumPunctToSpace.the" thesaurus in VantagePoint and reduce 82,746 to 72,406 phrases.

Table 1
Term clumping stepwise results.

DSSCs5784 records (WoS + Compendex), 2001–2010	
Field selection	Title & Abstract (NLP phrases + keywords)
Phrases with which we begin	90,980
Step a. Applying thesauri for common term removal	
01- XMLencoding.the	82,746
02- NumPunctToSpace.the	72,406
03- TermClumpingBasicCleaning.the	63,812
Step b. Fuzzy matching	
04- General.fuz	58,577
05- General-85cutoff-95fuzzywordmatch-1 exact.fuz	53,718
Step c. Combining	
06- Combine_Terms_Network.vpm (Optional)	Not Applied Here
07- Term_Clustering.vpm	52,161 to 37,928[a]
Step d. Pruning	
08- Remove single terms	15,299
09- General-85cutoff-95fuzzywordmatch-1 exact.fuz	14,840
Step e. Screening	
10- Term Frequency Inverse Document Frequency (TFIDF)	14,840 (with the Sequence of TFIDF) to 14,740[b]
11- Combine_Terms_Network.vpm (Optional)	8038
Step f. Clustering	
12- Principal Components Analysis (PCA)	11 Topical Clusters

[a] We ran an unfinished "Term_Clustering.vpm" in VantagePoint, and reduced the terms from 53,718 to 52,161, then, we selected the 37,928 terms which contain 2, 3 or 4 words.

[b] We ran the TFIDF analysis in VantagePoint, and got the TFIDF value for each term; then removed the top 100 highest TFIDF terms, then used the remaining 14,740 terms for the next steps.

Next we apply a big combined "TermClumpingBasic Cleaningthesaurus," involving "Stopword Removal," "Common Terms Removal," "General Academic Terms Removal," [23] "DSSC-related Terms Consolidation," and "Trash Terms Removal." Here, we reduce terms from 72,406 to 63,812. Some examples:

- Remove general verbs (e.g., am, is, are, and do), prepositions (e.g., in, on, and of), and articles (e.g., the);
- Remove common terms and general academic terms such as "year," "manual," "methodology," and "analysis";
- Consolidate common DSSC terminology—e.g., "DSSC*," "DSC*," "dye-sensitized solar cell*," and similar terms will be consolidated to "DSSCs";
- Remove "trash terms" generated by previous thesauri (shown in Table 2);
- Remove "trash terms"—names of organizations, governments, and companies, such as "United States Abstract," "Chinese Chemical Society," and "2009 Elsevier Ltd."

4.1.2 . Step b. Fuzzy matching

In addition to the use of general and tailored thesauri, our other main computer-aided cleaning mechanism is fuzzy matching. As named, the fuzzy matching function combines terms with similar structure based on stemming (e.g., technology and technological) and combines singular and plural forms of English words (e.g., method and methods). VantagePoint provides a general fuzzy matching routine, as well as routines tailored to match person names, organizations, and to coordinate British and American spelling. It also provides the capability to readily tune fuzzy parameters to consolidate particular types of matches. Fuzzy matching (called "List Cleanup" in VantagePoint) coordinates well with thesaurus operations. For example, one can run a fuzzy matching routine, check and tune the results, and then save the resulting pairings as a thesaurus for future applications.

We apply our main fuzzy matching routines in the fifth step. VantagePoint's general fuzzy routine reduces the 63,812 phrases to 58,577. A tailored version of this (called "general-85cutoff-95fuzzywordmatch-1 exact.fuz") further drops the phrase set to 53,718. Altogether this effects a reduction by 4859 phrases.

4.1.3 . Step c. Combining

In this section, we introduce two new approaches: "Combine Term Networks (CTN)" analysis and "Term Clustering" analysis. Typically, we might use "Combine Term Networks" analysis to consolidate authors and their main co-authors before we start author-associated analyses, because this consolidation helps us to find the core authors more easily. In this circumstance, we transfer the same idea from author consolidation to term consolidation, and it seems to work pretty well. In particular, a CTN macro in VantagePoint is able to consolidate related terms, which results in a reduced number of terms, but no increase in record count for existing terms, just more instances. Actually, the macro for CTN analysis will combine the low-frequency terms with the high-frequency terms (target terms) that appear in the same records. Sometimes, the target terms are meaningless for our research, especially for emerging technology studies. Thus, how to deal with CTN analysis is an option for the "term clumping" steps. In this paper, we focus on the "Term Clustering" analysis, and skip the CTN analysis (Table 1). However, we apply the CTN analysis in the Screening step, as a test, after the TFIDF analysis.

Before we try the "Term Clustering" macro for the 53,718 terms, based on the experts' suggestions, we remove the top terms appearing in more than 1000 records (i.e., DSSCs, solar cell, dye sensitive solar cell, photo electrochemical cells, efficient conversion, electrolyte, TiO2, and titanium dioxide). These eight terms are really general in the DSSC domain, and this also means that they will heavily influence the combining process. After that, we run the "Term Clustering" macro on a computer with substantial power and memory, but VantagePoint runs for 8 days and shuts down with an "out of memory" error. In the plan, the basic outline of the "Term Clustering" macro is as follows: [45]

- Remove hyphens, numbers, and punctuation.
- Remove common words.
- Clump phrases with four or more words in common into a new phrase.
- Rename the new phrase with the shortest possible phrase name.
- Calculate the prevalence of the remaining words.
- Clump phrases with three words in common into a new phrase.
- When a conflict arises, use a similarity measure to determine with which group of phrases the conflicted phrase will clump.
- Rename the new phrase with the phrase name with the highest prevalence score.
- Repeat the same steps for two-word matches.

Although we fail to obtain the final results expected, we still check the unfinished results wherein the macro has reduced the terms to 52,161. We notice the macro has grouped multiple word phrases, including 1-word to 8-word groupings (shown in Table 3). As we mentioned, we emphasize the more meaningful phrases including 2, 3 or 4 words; thus we group

Table 2
Trash terms list.

	# Records	# Instances	Abe + Ti phrases + keys
1	5966	67,333	**Remove**
2	3759	8411	Trash
3	3266	7103	ACAD COMMON
4	2835	13,844	
5	296	419	Basic English
6	259	289	Numbers

Table 3
Multiple word phrase classification.

Multiple word phrases	Number
1 word phrases	2680
2 word phrases	18,795
3 word phrases	13,962
4 word phrases	5171
5 word phrases	2430
6 word phrases	1187
7 word phrases	399
8 or more word phrases	201

those into a sub-list containing 37,928 terms for further research. We use these 37,928 terms for our next steps (We are working to streamline this macro).

4.1.4 . Step d. Pruning

Thousands of terms appear in a single record. As such they are useless in most analyses that depend on co-occurrence of terms across records. However, one wants to consolidate related terms to give multi-record compilations that can contribute to various analyses before "pruning"—i.e., discarding very low-frequency terms. As shown in Table 1, in the "pruning" step, we remove all the single terms that only appear in one record, in this case after Step 16. Pruning here reduces the phrase count from 37,928 to 15,299. We then reapply the fuzzy match macro to those 15,299 terms, thereby reducing to 14,840 terms.

4.1.5 . Step e. Screening

We run VantagePoint's Term Frequency Inverse Document Frequency (TFIDF) analysis in this step. As the name implies, it evaluates not only the frequency of the term but also the frequency of the records wherein the term appears. We have experience with the evaluation of TFIDF results (shown as Fig. 2). Focusing on the two parts in the figure, Part 1 depicts both high document and term frequency, while Part 2 shows the medium level of document frequency. An outstanding question concerns how much attention to focus on the high document frequency terms? The answer varies. For example, if we start TFIDF analysis after a "perfect text cleaning," Part 1 seems to be a good choice. If not, the terms of Part 1 are usually general ones, and the most meaningful terms belong to Part 2. For another example, even if we perform a "perfect text cleaning," Part 1 could be full of field-related common terms, which could be useful for macro-assessment but meaningless for emerging technology research. That is, for some uses, we want to focus on general DSSC concepts; for other uses, we care about specialized topics discussed in subsets of the DSSC records. In this instance, it is important for us to make the decision based on the intent of the study. In addition, the thresholds for high, medium, and low frequency depend on the actual situation and desired outputs.

The process of TFIDF can be considered in the following three steps. First, we create a key value field in VantagePoint for the whole set of DSSC records (5784 records). We then make a matrix with the key terms by publication year. Second, we add all key terms to a new group "All" for all publication years and create a matrix (using the TFIDF option in VantagePoint) with the 14,840 terms and the group "All." Third, VantagePoint generates the TFIDF value for each term.

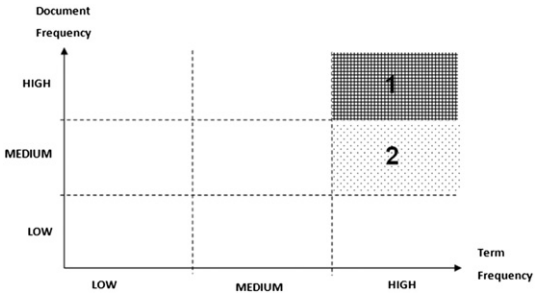

Fig. 2. TFIDF analysis evaluation.

DSSCs represent an emerging technology, thus, as discussed, we prefer the TFIDF terms derived from Part 2 (medium document frequency). Based on our previous experience, the top 1% of the total TFIDF terms are apt be really common within domain; thus we remove the Top 100 terms with the highest TFIDF value, such as "counter electrode," "photovoltaic performance," "electron transport," etc. Also, we mention that the top term removal here is a little different from the similar removal approach we perform in the combining step. In the combining step, the removed terms are high-frequency terms; meanwhile, here we remove terms with high TFIDF values. These high TFIDF value terms could occur with high frequency or low frequency because we introduce the document frequency in our weighting system, as we just discussed. We analyze the remaining 14,740 terms and select interesting terms with different positions on the two axes: TFIDF score and frequency of occurrence in records. For example:

A. Terms "solar hydrogen production" and "tandem cell system" only appear in 6 and 4 records, which are really low-level terms in the frequency-based term list. However, without the top 100 common terms, both of them rank in the top 50 of the 14,740 terms. These terms seem to make sense to us.
B. Terms "three dimensional," "hybrid material," and "building block" rank 364th, 584th, and 675th in the frequency-based term list, but none of them appear in the top 1000 of the 14,740 terms. We doubt that these terms are meaningful in analyzing emerging DSSC technologies;
C. We also notice that significant ranking changes occur before the top 3000 TFIDF terms, and the TFIDF terms out of the top 3000 also seem to fall in the low-frequency terms set.

Because we remove the top 100 TFIDF terms, the situation of the remaining terms seems to be more passable for the CTN analysis than for the Combining step. Thus, we apply the CTN macro to the 14,740 terms and reduce the term set to 8038.

4.1.6 . Step f. Clustering

As mentioned, the inductive method translates into a continuous process where we clean and consolidate terms step by step and then obtain the topical factors via statistical routines. Co-occurrence analyses teach us to consider terms as associated that occur together in records more frequently than chance would indicate. In our analyses, Principal Components Analysis (PCA) is usually applied to the clumped term set to reduce the number of items dramatically for further topical analyses.

In this paper, we select the top 200 terms of the "after CTN" 8038 terms as the high level terms, and generate a factor map via the VantagePoint's PCA analysis. The results are shown in Fig. 3 and also in Table 4. There are 11 clusters in the map, and most of them are not totally separated; several phrases that relate most closely to the cluster are listed. Especially, because the selected terms for PCA only cover 33% of the records (1924 records) from the whole dataset (5784 records), we report two kinds of coverage in Table 4.

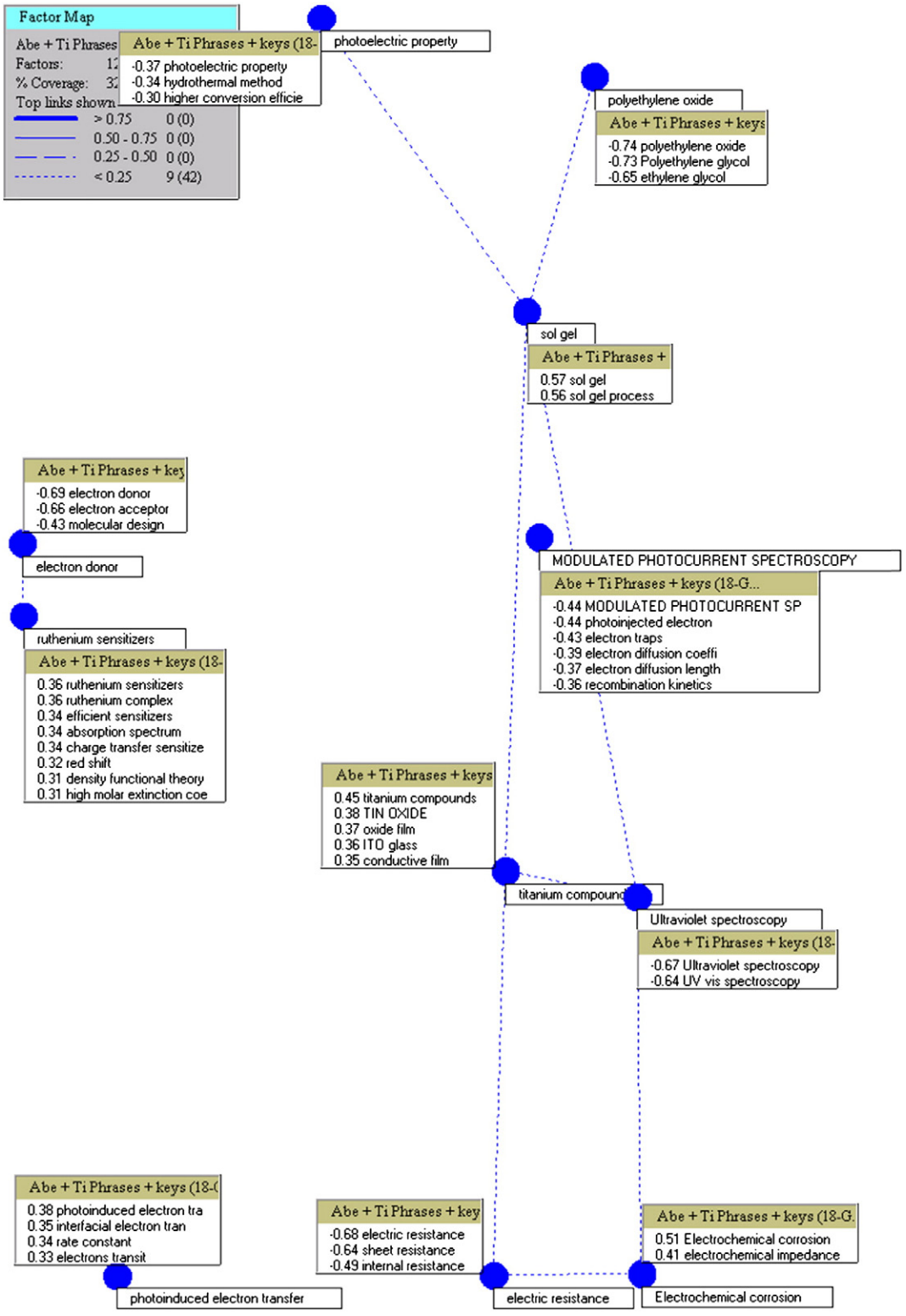

Fig. 3. Factor map of DSSCs (based on the top 200 terms).

4.1.7 . Results

After completing the "term clumping" steps, we scan the phrase set (prior to PCA) and nominate the following as particularly promising terms for further analyses. This provides an alternative output from the term clumping, stopping short of clustering (as just illustrated using PCA).

Based on more than 5 years of experience in analyzing DSSCs and experts' review, our "term clumping" processing is well adapted to the case of DSSCs. However, we also try to explore more opportunities to extend/modify our method for other uses. First, literature reviews and expert interviews are really important at the beginning of any case analysis. These can inform construction of a topical common terms list and tuning of the macros. Second, we define DSSCs as a kind of emerging technology, which has booming ST&I data resources for our "term clumping" studies. Thus, it should be easy to apply our method to another emerging technology (e.g., electric vehicle, nanotechnology, and etc.). However, if the case focuses on mature technology (e.g., machine tools, control systems), more efforts and modifications are warranted. We keep upgrading our term clumping method. However, we also must emphasize that our approach is attuned to examining a specific topic and particular analytical interests. Generalization of the term clumping process to other topics should work, but requires reflection on sensitivities.

4.2. Purposive methods

For the topical analyses, which are the purpose of "term clumping," we plan to explore the relative advantages of two approaches to generate interpretable, informative topical factors. The first is an inductive method, emphasized herein, where we work to consolidate terms into topical factors. This works from the dataset without a priori criteria to target particular terms. The second is a purposive method that comes to the given text compilation with pre-conceived key terms. We are exploring the relative efficacy of such approaches.

In this paper, the term clumping steps for technical intelligence mostly belong to the inductive approach. We will not apply the deductive approach here. However, in past years, we have contributed efforts toward purposive methods. On the one hand, we introduce the semantic TRIZ tool to bridge the term clumping results and Technology Roadmapping by extending our understanding from "topical factors (phrases and terms)" to "Problem & Solution patterns (Subject + Action + Object)" [39]. On the other hand, we also package "Term Clumping," "Semantic TRIZ," and "Technology Roadmapping" for the study of "Triple Helix" relations—the interplay of the Academy, Government, and Private Enterprise for technology and policy assessment researches [46].

4.3 . Expert assessment of the clumped terms and clusters

How to most effectively engage an expert who is broadly knowledgeable over the domain is challenging. Usually, experts are busy, difficult to invite into surveys and workshops, and also occasionally cost much time or money. However, experts provide a critical means of assessing the resulting terms and clusters. For this paper, we sent the clusters in Table 4, combined with another 10 Hierarchical Dirichlet Process (HDP) and Hierarchical Latent Dirichlet Allocation (HLDA) clusters, and 322 terms after the term clumping steps (including the top 60 terms in Table 5) to seven experts from Georgia Tech, Tsinghua University, Dalian University of Technology, IBM, and Booz Allen Hamilton, Inc. and asked for their judgments. We have been compiling our DSSC expert contacts since we started this case more than five years ago and keep updating this list. Considering areas of specialization and balance between Chinese and American experts, we chose the seven experts as mentioned.

Before we present the expert feedback, we calculate the correlations among the experts' judgments. Several experts appear to have quite similar research interests, but ratings are highly independent. From our knowledge of their backgrounds, some of them may focus on specific DSSC sub-fields, while some focus on a larger domain (e.g., solar cells in general). The highest inter-rater correlation on terms was 0.18 for a PhD student and her advisor. These two were also relatively highly correlated in their rating of 33 term clusters (including topic models and the PCA factors) at 0.31, with another two pairs of experts correlating a little higher (0.39, 0.37). But, overall, the experts' cluster judgments correlate at only 0.09.

The experts varied in the number of terms or clusters they found to be of interest. For terms, one selected 183 out of 322 as interesting; the others ranged from 36 to 67 terms selected. So we choose to weight their responses as a fraction of their overall selections. For the term clusters, we score "interesting" as 1.0 and "possibly interesting" as 0.5. For terms, we only asked for "interesting" judgments and score those as 1.0. We then add up all the ratings by a given expert and divide his/her individual item ratings by that value. For instance, one expert rated 50 of the 322 terms as "interesting." So dividing each by

Table 4
Clusters and related factors of DSSCs.

Clusters	Coverage/1924	Coverage/5784	Factors[a]
1	8.42%	2.80%	**Photoelectric property**, hydrothermal method, and higher conversion efficiency
2	7.80%	2.59%	**Polyethylene oxide**, polyethylene glycol, and ethylene glycol
3	10.34%	3.44%	**Sol gel**, sol gel process
4	7.33%	2.44%	**Electron donor**, electron acceptor, and molecular design
5	22.40%	7.45%	**Ruthenium sensitizers**, ruthenium complex, efficient sensitizers, absorption spectrum, charge transfer sensitizer, red shift, density functional theory, and high molar extinction coefficient
6	5.87%	1.95%	**Electric resistance**, sheet resistance, and internal resistance
7	13.15%	4.37%	**Modulated photocurrent spectroscopy**, electron diffusion coefficient, electron traps, recombination kinetics, photo-injected electron, and electron diffusion length
8	13.98%	4.65%	**Photo-induced electron transfer**, electrons transit, interfacial electron transfer, rate constant
9	9.77%	3.25%	**Electrochemical corrosion**, electrochemical impedance spectra
10	4.11%	1.37%	**Ultraviolet spectroscopy**, UV vis spectroscopy
11	17.52%	5.83%	**Titanium compounds**, oxide film, tin oxide, ITO glass, and conductive film

[a] The bolded terms in the Factors column are the factor names suggested by VantagePoint, based mainly on the phrase that loads most highly on the resulting factor.

Table 5
Final topical phrases list.

	Terms		Terms
1	Cell membrane	31	Raman spectroscopy
2	Electrochemical corrosion	32	Interfacial electron transfer
3	Electron mobility	33	Conjugated polymers
4	Titanium compounds	34	Crystalline materials
5	Nanocrystalline material	35	Ionization of liquids
6	Electrochemical electrodes	36	Nanotube arrays
7	Ruthenium sensitizers	37	High molar extinction coefficient
8	Sol gel process	38	Transparent conductive oxide
9	Temperature molten salt	39	Charge transfer sensitizer
10	Semiconducting zinc compounds	40	Conducting glass substrate
11	Ruthenium complex	41	Photoelectrochemical performance
12	Oxide film	42	Electron injection efficiency
13	Impedance spectroscopy	43	Absorption spectrum
14	Mesoporous material	44	Electrochemical impedance spectra
15	Polyethylene oxide	45	Photoelectrochemical solar cell
16	Tin oxide	46	Spectral sensitivity
17	Organic polymer	47	Electrophoretic deposition
18	Polyethylene glycol	48	Semiconducting electrode
19	Semiconductor material	49	Ultraviolet spectroscopy
20	Semiconductor film	50	Electron donor
21	Chemical vapor deposition	51	Fourier transform infrared spectroscopy
22	Conductive polymer	52	Hydrothermal synthesis
23	Organic solvent	53	Solid state solar cell
24	Solid electrolyte	54	Differential scanning calorimetry
25	Short circuit photocurrent	55	Modulated photocurrent spectroscopy
26	Electron diffusion coefficient	56	Dye-sensitized photoelectrochemical cell
27	Organic sensitizers	57	Nanocrystalline titanium dioxide
28	Molecular design	58	Organic hole transport material
29	Solid state device	59	Transient absorption spectroscopy
30	Photocatalytic activity	60	Cathodicelectrodeposition

50 gives a score of 0.02 for each item he tagged. (Equivalently, we are giving each expert 100 points to divide among the items, so the fractional score is like a percentage of their vote.)

Addressing the clusters:

1) 10 out of 11 PCA clusters received at least 2 experts' acceptance; 8 of those obtained at least 3 experts' acceptance. One cluster (ranked 9th in record coverage) was not selected by any of our seven experts as interesting for further analyses.
2) Table 6 arrays 4 of the 11 PCA clusters based on their record coverage. The second and third most highly ranked clusters are, respectively, tenth and fifth in their record coverage ranking. This shows that expert interest does not relate neatly to cluster generality.

Addressing terms (phrases):

1) We sent 322 terms to the DSSC experts and asked for expert judgments as to which are interesting—249 terms (77.3%) received at least 1 expert's indication of interest. This suggests that the term clumping process is producing high interest outputs for further analyses.
(If we exclude our extreme rater who judged 183 terms as interesting, 183 of 322 terms were still judged as interesting (57%)—a hearty acceptance rate. Alternatively, the other 6 raters gave 300 votes of term acceptance collectively, and those were divided among 183 terms.)
2) Several top high-frequency terms obtained a very low expert ranking. The highest-frequency term (cell membrane) did not obtain a single expert's acceptance; the second term (electrochemical corrosion) received a single expert's acceptance; only the third most frequent term (electron mobility) is ranked highly (16th) in the experts' rankings.
3) Also, the top term in the experts' ranking (diffusion length) does not appear in the top 60 terms based on frequency of occurrence in the record set. This suggests that it seems to be a specialty area within the overall research field. This makes sense – some topics are apt to be quite general – appearing in some form in many of the abstract records – and some would be expected to be quite specific. For further analyses, sorting topics into “general” and “specific” could be helpful.

Table 6
Comparisons of expert judgments and PCA record coverage.

PCA cluster label	Record coverage	Experts' choice (of 11)
Ruthenium sensitizers	22.4%	4
Titanium compounds	17.52%	9
Photo-induced electron transfer	13.98%	1
Modulated photocurrent spectroscopy	13.15%	6

Table 7
Comparison of the top 30 terms BEFORE and AFTER term clumping steps.

	Original top 30 terms	Final top 30 terms
1	Dye sensitive solar cell	Cell membrane
2	Solar cell	Electrochemical corrosion
3	Rights reserved	Electron mobility
4	Photoelectrochemical cell	Titanium compounds
5	Dye	Electrons transit
6	Conversion efficiency	Sol gel
7	Film	Nanocrystalline material
8	Dye-sensitized solar cells	Electrochemical electrodes
9	Electrolyte	Ion exchange
10	Titanium dioxide	Redox reaction
11	Electrode	Switching circuits
12	Efficient	Ruthenium sensitizers
13	Photovoltaic cell	Sol gel process
14	TiO2	Digital camera
15	Results	Photocurrent density J sc
16	Conversion	Electric potential
17	Light	Overall energy conversion efficiency
18	Effect	Experimental result
19	Efficiency	Temperature molten salt
20	Photocurrent	Ruthenium compound
21	Cell	Short circuit currents
22	Application	Efficient cell
23	Recombination	Prepared film
24	DSSC	Semiconducting zinc compounds
25	Thin film	First time
26	Fabrication	Redox couple
27	Performance	Ruthenium complex
28	Dye-sensitized solar cells DSSCs	Visible light
29	TiO2 film	Oxide film
30	Energy conversion	Impedance spectroscopy

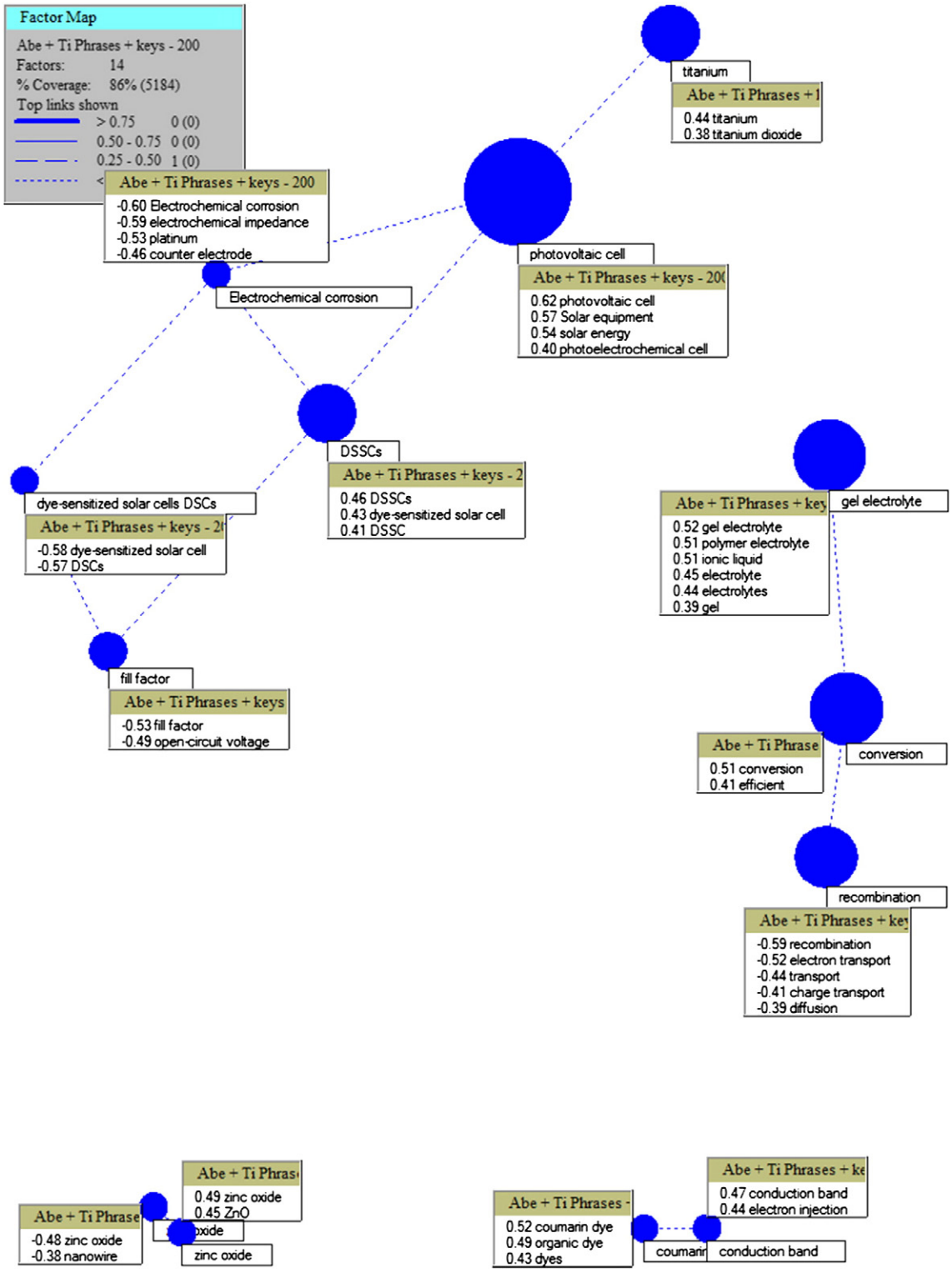

Fig. 4. Factor map of DSSCs (based on the top original 200 terms without term clumping).

5 . Result comparisons

In order to assess the utility of our term clumping steps, we compare the stepwise results in this section: 1) we compare the top 30 terms before and after the term clumping steps; 2) we generate the factor map based on the top 200 terms from the original list without term clumping and compare with Fig. 3; 3) we pick up several

typical terms and one sample to show the changes step by step.

In Table 7, we take the top 30 terms from the original term list and the top 30 terms from the final term list. Obviously, the difference is significant. All top 30 terms from the original term list are removed in our term clumping steps; these include common terms (e.g., film, efficient, results, etc.), redundancies (e.g., dye sensitive solar cell, dye-sensitized solar cell and DSSC), and general terms for this dataset (e.g., solar cell, electrolyte, DSSC).

Aiming to discover the significance of the term clumping steps visually, we generate the factor map with the original top 200 terms (shown as Fig. 4). There are 13 topical factors in the map, but most of them are redundant (e.g., 2 zinc oxide and 2 DSSCs) or meaningless (e.g., titanium and conversion). Although we could also find one important topical factor (e.g., gel electrolyte) in Fig. 4, considering all the 13 factors, it is easier to judge the factor map (Fig. 4) after the term clumping steps.

In Table 8, we take the top 10 terms from the original term list (#1–10) and another 8 interesting terms (A–H) and compare the changes after "Applying Thesaurus for Common Term Removal" and "Fuzzy Matching." Obviously, a big change within the top 10 terms has occurred after several thesauri are used to remove thousands of common terms or consolidate term variations.

Notice that the top terms do not change much when we apply the fuzzy matching routines. The next several steps similarly reduce the total amount of terms sharply, but the top terms remain in the same sequence. Considering the top 8 general DSSC terms, we remove them, and start the "Combining" step. Also, because of the unfinished "Term Cluster" macro, we are left with 37,928 terms, which are phrases containing 2, 3, or 4 words. After that, the "pruning" and "screening" steps follow, where thousands of single or low-frequency terms are removed or consolidated. In this instance, we pick up a special sample to show the BIG changes among these steps.

Table 9
Stepwise changes with "Electron/Electrons/Electronic/Electronics" sample.

	Step	Number
1	Original List	756
2	After the "Applying Thesaurus for Common Term Removal" step	640
3	After the "Fuzzy Matching" step	452
4	After the "Combining" step	388
5	After the "Pruning" step	223
6	After the "Screening" step	137

Table 9 shows the terms starting with variations of "Electron/Electrons/Electronic/Electronics," as a case illustration. Compare the changes in the total number of these terms step by step. It is obvious that the amount is reduced by around 100 terms in each step, and the "Fuzzy Matching" and "Combining" steps seem to be particularly powerful.

At the same time, to track the changes in detail, we choose the top 10 terms "After Screening," and compare the changes in different steps (shown in Table 10) on their ranking, "#R" and "#I" (#R = Number of Records containing that term; #I = Number of Instances—how many times the terms appear, counting multiple occurrences in a record.). This comparison gives us several interesting discoveries:

1) The sequence of top 10 terms is always changing. Comparing the difference between the top 10 terms from the original term list to the "After Screening" term list, only 2 terms are the same. However, after the "Applying Thesaurus for Common Term Removal" step, the sequence of the top 10 terms does not change much.
2) Before the "Screening" step, most changes result from "term consolidation," where similar terms were consolidated, thus, both "#R" and "#I" increase.
3) In the "Screening" step, terms with extremely high TFIDF values are removed. Then low-frequency terms are combined into high-frequency terms that appear in the same record.

Table 8
Comparison of stepwise "term clumping" results.

	Terms	Original			After Thesauri for Common Term Removal			After Fuzzy Matching		
		Rank	#R	#I	Rank	#R	#I	Rank	#R	#I
1	Dye sensitive solar cell	1	2780	4045	3	2780	4045	3	2882	4240
2	Solar cell	2	2408	2823	2	3171	5117	2	3171	5117
3	Rights reserved	3	1608	2092	–	–	–	–	–	–
4	Photoelectrochemical cell	4	1605	1692	4	1623	1727	4	1630	1740
5	Dye	5	1326	1844	–	–	–	–	–	–
6	Conversion efficiency	6	1301	1691	–	–	–	–	–	–
7	Film	7	1133	1285	–	–	–	–	–	–
8	Dye-sensitized solar cells	8	1126	1610	–	–	–	–	–	–
9	Electrolyte	9	1117	1911	6	1190	2251	6	1190	2251
10	Titanium dioxide	10	1073	1150	8	1073	1150	8	1073	1150
A	Photovoltaic cell	13	935	978	–	–	–	–	–	–
B	TiO2	14	926	1273	7	1173	1692	7	1173	1692
C	DSSCs	32	534	1118	1	4672	13,509	1	4672	13,509
D	Open-circuit voltage	175	147	196	18	547	848	18	547	848
E	X-ray diffraction	341	89	113	58	209	269	57	211	274
F	Efficient conversion	–	–	–	5	1319	1737	5	1319	1737
G	Applicator	–	–	–	10	777	1165	10	777	1165
H	Material nanostructure	–	–	–	20	525	537	20	525	538

Rank is based on the #R.

Table 10
Stepwise changes for the top 10 terms in the "After Screening" list of "Electron/Electrons/Electronic/Electronics" phrases.

	Terms	Original			After thesauri for common term removal			After fuzzy matching			After pruning			After screening		
		Rank	#R	#I	Rank	#R	#I	Rank	#R	#I	Rank	#R	#I	Rank	#R	#I
1	Electron mobility	6	140	149	5	143	154	5	143	154	6	143	154	1	143	236
2	Electrons transit*	7*	129	129	7*	133	134	7	139	141	7	139	141	2	139	195
3	Electron diffusion coefficient	14	50	73	12	51	75	10	65	102	10	65	102	3	70	306
4	Electron traps	17	46	54	14	50	60	11	59	90	11	59	90	4	59	128
5	Electron acceptor	16	46	72	10	56	95	12	58	98	12	58	98	5	58	160
6	Electron injection efficiency	24	26	37	13	51	73	13	55	79	13	55	79	6	58	168
7	Electron donor	18	44	67	11	53	86	14	53	88	14	53	88	7	53	153
8	Electrons recombine**	15**	49	64	15**	49	64	15	53	68	15	53	68	8	53	124
9	Electron diffusion length	21	33	59	16	38	66	16	41	69	16	41	69	9	43	105
10	Electron energy levels	20	34	34	19	34	34	17	37	41	17	37	41	10	37	41

Before "Fuzzy Matching," *"electrons transit" is named "electrons transition," **"electrons recombine" is named "electrons recombination."

However, this combination only increases the "#I" and does not change the "#R."

4) Based on the "Electron/Electrons/Electronic/Electronics" sample, our efforts seem to work to concentrate a number of terms into several topics and prepare for further topical analyses.

6 . Disscussion

Recent attention to themes like "Big Data" and "Money Ball" draw attention to the potential in deriving usable intelligence from information resources. We have noted the potential for transformative gains, and some potential unintended consequences of exploiting information resources [47]. Term clumping, as presented here, offers an important tool set to approach real improvements in identifying, tracking, and forecasting emerging technologies and their potential applications.

Desirable features in such text analytics include:

- Transparency of actions—not black box;
- Evaluation opportunities—we see value in comparing routines on datasets to ascertain what works better; we recognize that no one sequence of operations will be ideal for all text analytics.

We are pointing toward the generation of a macro that would present the analyst with options as to which cleaning and clumping steps to run and in what order; however, we also hope to come up with a default routine that works well to consolidate topical terms and phrases for further analyses.

Some future research interests have been noted. We are particularly interested in processing unigrams (single words), because of the potential in such approaches to work with multiple languages. On the other hand, we appreciate the value of phrases to convey thematic structure. Possibilities include processing single words through a sequence of topic model steps and then trying to associate related phrases to help capture the thrust of each topic.

We see a potential use for clumped terms and phrases in various text analyses. Two relating to competitive technical intelligence (CTI) and Future-oriented Technology Analyses (FTA) would be as follows:

- Combining empirical with expert analyses is highly desirable in CTI and FTA—clumped phrases can be further screened to provide digestible input for expert review to point out key topics and technologies for further scrutiny.
- Clumped phrases and/or PCA factors can provide appropriate level content for Technology RoadMapping (TRM)—for instance, to be located on a temporal plot.

We recognize a considerable interplay among text content types as well. This poses various cleaning issues in conjunction with co-occurrence of topical terms with time periods, authors, organizations, and class codes. We look forward to exploring ways to use clumped terms and phrases to generate valuable CTI.

Acknowledgments

We acknowledge the support from the US National Science Foundation (Award #1064146—"Revealing Innovation Pathways: Hybrid Science Maps for Technology Assessment and Foresight"). The findings and observations contained in this paper are those of the authors and do not necessarily reflect the views of the National Science Foundation.

We thank Stephen Carley, Jan Youtie, Donghua Zhu, Philip Shapira, and colleagues in the "Innovation Co-lab" of Georgia Institute of Technology, Beijing Institute of Technology, and Manchester University, for their advice and feedback.

References

[1] VantagePoint, www.theVantagePoint.com (accessed 5 August 2013) also see Thomson Data Analyzer (TDA), http://thomsonreuters.com/thomson-data-analyzer/.
[2] A.L. Porter, M.J. Detampel, Technology opportunity analysis, Technol. Forecast. Soc. Change 49 (1995) 237–255.
[3] A.L. Porter, S.W. Cunningham, Tech Mining: Exploiting New Technologies for Competitive Advantage, Wiley, New York, NY, 2005.
[4] R.J. Watts, A.L. Porter, S.W. Cunningham, D. Zhu, TOAS intelligence mining, an analysis of NLP and computational linguistics, Lect. Notes Comput. Sci. 1997 (1263) (1997) 323–334.
[5] D. Zhu, A.L. Porter, Automated extraction and visualization of information for technological intelligence and forecasting, Technol. Forecast. Soc. Change 69 (2002) 495–506.

[6] P. Losiewicz, D.W. Oard, R.N. Kostoff, Textual data mining to support science and technology management, J. Intell. Inf. Syst. 15 (2) (2002) 99–119.
[7] D. K. R. Robinson, L. Huang, Y. Guo, A. L. Porter, Forecasting innovation pathways for new and emerging science & technologies, Technological Forecasting & Social Change, to appear.
[8] Y. Kim, Y. Tian, Y. Jeong, J. Ryu, S. Myaeng, Automatic discovery of technology trends from patent text, Proceedings of the 2009 ACM symposium on Applied Computing, Hawaii, USA, 2009.
[9] M. Verbitsky, Semantic TRIZ, TRIZ J. (Feb 2004)(http://www.triz-journal.com/archives/2004/ (accessed 20 May 2012)).
[10] Y. Zhang, Y. Guo, X. Wang, D. Zhu, A.L. Porter, A hybrid visualization model for technology roadmapping: bibliometrics, qualitative methodology, and empirical study, Tech. Anal. Syst. Manag. 25 (6) (2013) 707–724.
[11] D.S. Price, Little Science, Big Science and Beyond, Columbia University Press, New York, NY, 1986.
[12] E. Garfield, M. Malin, H. Small, Citation data as science indicators, in: Y. Elkana, et al., (Eds.), The Metric of Science: the Advent of Science Indicators, Wiley, New York, NY, 1978.
[13] A.F.J. van Raan, Advanced bibliometric methods to assess research performance and scientific development: basic principles and recent practical applications, Res. Eval. 3 (3) (1992) 151–166.
[14] N. De Bellis, Bibliometrics and Citation Analysis, The Scarecrow Press, Lanham, MD, 2009.
[15] T. Ma, A.L. Porter, J. Ready, C. Xu, L. Gao, W. Wang, Y. Guo, A technology opportunities analysis model: applied to dye-sensitized solar cells for China, 4th International Seville Conference on "Future-oriented Technology Analysis (FTA)," Spain, 2011.
[16] Y. Guo, C. Xu, L. Huang, A.L. Porter, Empirically informing a technology delivery system model for an emerging technology: illustrated for dye-sensitized solar cells, R&D Manag. 42 (2) (2012) 133–149.
[17] A. Bookstein, T. Klein, T. Raita, Clumping properties of content-bearing words, J. Am. Soc. Inf. Sci. 49 (2) (1998) 102–114.
[18] A. Bookstein, T. Raita, Discovering term occurrence structure in text, J. Am. Soc. Inf. Sci. Technol. 52 (6) (2000) 476–486.
[19] A. Bookstein, K. Vladimir, T. Raita, N. John, Adapting measures of clumping strength to assess term-term similarity, J. Am. Soc. Inf. Sci. Technol. 54 (7) (2003) 611–620.
[20] J. Allan, J. Carbonell, G. Doddington, J. Yamron, Y. Yang, Topic detection and tracking pilot study final report, Proceedings of the DARPA Broadcast News Transcription and Understanding Workshop, 1998.
[21] R. Nallapati, Semantic language models for topic detection and tracking, Proceedings of the conference of the North American chapter of the Association for Computational Linguistics on Human Language Technology, 2003, pp. 1–6.
[22] S.W. Cunningham, The content evaluation of British scientific research, D.Phil. Thesis Science Policy Research Unit, University of Sussex, Brighton, United Kingdom, 1996.
[23] Haywood, S. Academic Vocabulary, Nottingham University.http://www.nottingham.ac.uk/~alzsh3/acvocab/wordlists.htm (accessed 26 May, 2012)
[24] I.K. Fodor, A survey of dimension reduction techniques, U.S. Department of Energy, Lawrence Livermore National Lab, 2002. (https://e-reports-ext.llnl.gov/pdf/240921.pdf (accessed 22 May 2012)).
[25] S. Deerwester, S.T. Dumais, G.W. Furnas, T.K. Landauer, R.A. Harshman, Indexing by latent semantic analysis, J. Soc. Inf. Sci. 41 (6) (1990) 391–407.
[26] In: T.K. Landauer, D.S. McNamara, S. Denis, W. Kintsch (Eds.), Handbook of Latent Semantic Analysis, Erlbaum Associates, Mahwah, NJ, 2007.
[27] D. Blei, A. Ng, M. Jordan, Latent Dirichlet allocation, J. Mach. Learn. Res. 3 (2003) 993–1022.
[28] M.W. Berry, M. Castellanos, Survey of text mining II: clustering, classification, and retrieval, Springer, New York, NY, 2008.
[29] W. Pottenger, T. Yang, Detecting emerging concepts in textual data mining, Comput. Inf. Retr. (2001) 1–17.
[30] R.J. Watts, A.L. Porter, Mining foreign language information resources, Proceedings of Portland International Conference on Management of Engineering and Technology, Portland, OR, USA, 1999.
[31] H. Thomas, Probabilistic latent semantic indexing, Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in, information retrieval, 1999, pp. 50–57.
[32] R.K. Ando, Latent semantic space: iterative scaling improves precision of inter-document similarity measurement, Proceedings of the 23rd annual international ACM SIGIR conference on Research and development in information retrieval, 2000, pp. 216–223.
[33] W. Ding, S.N. Yu, S.Q. Yu, W. Wei, Q. Wang, LRLW-LSI: an improved latent semantic indexing (LSI) text classifier, Proceedings of the 3rd international conference on Rough sets and knowledge technology, 2008, pp. 483–490.
[34] L. Xu, N. Furlotte, Y. Lin, K. Heinrich, M.W. Berry, E.O. George, R. Homayouni, Functional cohesion of gene sets determined by latent semantic indexing of PubMed abstracts, PLoS One 6 (4) (2011) 1–9.
[35] J.I. Maletic, A. Marcus, Using latent semantic analysis to identify similarities in source code to support program understanding, 12th IEEE International Conference on Tools with Artificial Intelligence, 2001, pp. 46–53.
[36] J.I. Maletic, A. Marcus, Supporting program comprehension using semantic and structural information, Proceedings of the 23rd International Conference on Software Engineering, 2001, pp. 103–112.
[37] Q. Mei, C. Zhai, A mixture model for contextual text mining, Proceedings of the 12th ACM SIGKDD international conference on Knowledge Discovery and Data Mining, 2006, pp. 649–655.
[38] H. Sayyadi, M. Hurst, A. Maykov, Event detection and story tracking in social streams, Proceeding of 3rd Int'l AAAI Conference on Weblogs and Social Media, San Jose, California, USA, 2009.
[39] A. Shinmori, M. Okumura, Y. Marukawa, M. Iwayama, Patent claim processing for readability: structure analysis and term explanation, Proceedings of the ACL-03 workshop on patent corpus processing, 2003, pp. 56–65.
[40] Y. Zhang, X. Zhou, A.L. Porter, J.M.V. Gomila, How to combine term clumping and technology roadmapping for newly emerging science & technology competitive intelligence: the semantic TRIZ tool and case study, 14th International Society of Scientometrics and Informetrics Conference, Vienna, Austria, 2013.
[41] In: A.L. Porter, Y. Zhang, S. Sakurai (Eds.), Text Clumping for Technical Intelligence, Theory and Applications for Advanced Text Mining, Tech Publishing, Croatia, 2012.
[42] N.C. Newman, A.L. Porter, D. Newman, C. Courseault-Trumbach, S.D. Bolan, Comparing Methods to Extract Technical Content for Technological Intelligence, Portland International Conference on Management of Engineering and Technology, Vancouver, Canada, 2012.
[43] L. Huang, Y. Guo, T. Ma, A. L. Porter, Text mining of information resources to inform forecasting of innovation pathways, Technology Analysis & Strategic Management, to appear.
[44] B. O'Regan, M. Grätzel, A low-cost, high efficiency solar-cell based on dye-sensitized colloidal TiO2 films, Nature 353 (6346) (1991) 737–740.
[45] C.C. Trumbach, D. Payne, Identifying synonymous concepts in preparation for technology mining, J. Inf. Sci. 33 (6) (2007) 660–677.
[46] Y. Zhang, X. Zhou, A. L. Porter, J. M. V. Gomila, Triple helix innovation in China's dye-sensitized solar cell industry: hybrid methods with semantic TRIZ and technology roadmapping, scientometrics, DOI: 10.1007/s11192-013-1090-9
[47] In: A.L. Porter, W. Read (Eds.), The Information Revolution: Current and Future Consequences, JAI/Ablex, Westport, CT, 1998.

Yi Zhang is a Ph.D. Candidate in the School of Management and Economics, Beijing Institute of Technology, China. His specialty is data mining and Science Technology and Innovation (ST&I) management, especially the study of technology forecasting and assessment. Current researches emphasize the "Text Mining", TRIZ and Technology Roadmapping topics.

Alan Porter is Professor Emeritus of Industrial & Systems Engineering, and of Public Policy, at Georgia Tech, where he remains Co-director of the Technology Policy and Assessment Center. He is also Director of R&D for Search Technology, Inc., Norcross, GA. He is author of some 220 articles and books. Current research emphasizes measuring, mapping, and forecasting ST&I knowledge diffusion patterns.

Zhengyin Hu is a faculty member in Chengdu branch of National Science Library, Chinese Academy of Science, China. His specialty is text mining and Knowledge Organization System (KOS), especially the study of data mining based on KOS. Current researches focus on "Topic Model", TRIZ and Science & Technology Knowledge Organization System (STKOS).

Ying Guo is a faculty member in the School of Management and Economics, Beijing Institute of Technology, China. Her specialty is science and technology management, particularly the study of technology forecasting and assessment. She is currently focusing on how to forecast the innovation pathways for emerging science and technology topics.

Nils C. Newman is an Economics Ph.D. Candidate with UNU-MERIT at the University of Maastricht in the Netherlands. He is also the President of Intelligent Information Services Corporation in Atlanta, Georgia. For nearly two decades, Mr. Newman has worked on the development of analytical tools and processes related to the management of technology. His efforts focus on the use of bibliographic and patent information in research evaluation, competitive intelligence, and strategic planning.