The International Conference on Innovative Methods for Innovation Management and Policy (IM2012)
May 21st to 24th, 2012

"Term Clumping" for Technical Intelligence: A Case Study on Dye-Sensitized Solar Cells

Yi Zhang[1], Alan L. Porter[2], Zhengyin Hu[3], Ying Guo[1], Nils C. Newman[4]

[1]Corresponding Author, School of Management and Economics, Beijing Institute of Technology, Beijing, China

[2]Technology Policy and Assessment Center, Georgia Institute of Technology, Atlanta GA 30332; and Search Technology, Inc., U.S.A.

[3]Chengdu Branch of the National Science Library, Chinese Academy of Sciences, Beijing, China

[4]Intelligent Information Systems Corporation (IISC), U.S.A.

Emails: yi.zhang.bit@gmail.com; alan.porter@isye.gatech.edu; huzy@clas.ac.cn; violet7376@gmail.com; newman@iisco.com;

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 "what?" is being pursued in the R&D activities addressed in such publication and patent abstract records. This paper explicitly presents some 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 are able to reduce some 90,980 terms & phrases to much more user-friendly sets as one progress through the clumping steps. The resulting phrases are much better suited to contribute usable technical intelligence. We engage seven persons knowledgeable about DSSCs to assess the resulting content. These empirical results advance development of a semi-automated term clumping process that can enable extraction of topical content intelligence.

HIGHLIGHTS

- Term clumping generates technical intelligence to facilitate "Tech Mining"
- Empirical results indicate what each of six term clumping steps accomplishes
- Clumping key terms and/or title & abstract phrases helps elicit topical emphases
- Resulting phrases can provide insight into newly emerging science & technologies

KEYWORDS

Term clumping; Dye-Sensitized Solar Cells; DSSCs; Tech mining; Technical intelligence; Text clustering; Text analytics;

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)

1 INTRODUCTION

Over the last twenty years, Georgia Tech's Technology Policy and Assessment Center has been pursuing variants of our "Tech Mining" approach for retrieving usable information on the prospects of particular technological innovations from Science Technology and Innovation (ST&I) resources.[1-4] We focus on processing search results from ST&I databases that typically range into thousands of records. Such searches provide terms that can indicate significant topics during the emergence of a technology.

However, those term sets, as in our case, can easily approach 100,000 items, making analysis challenging. Herein, we are trying to enable faster and better Tech Mining by processing that topical content.

In this paper, we focus on abstract record search results pertaining to a technology of interest as the resource from which 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. [5, 6] 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 or quantitative methods, we try out systematic software steps 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 -- studying how those consolidated topics (concepts) change over time and their patterns of interaction. We believe that can expedite generation of technical intelligence and advance efforts to Forecast Innovation Pathways. [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 Cells 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 built around bibliometric analyses of ST&I records over the past 60 or so years, see for instance.[8-10] DeBellis nicely summarizes many facets of the data and their analyses.[11] Our

group at Georgia Tech has pursued ST&I analyses aimed especially at generating Competitive Technical Intelligence (CTI) since the 1970's, with software development to facilitate mining of abstract records since 1993.[12, 1, 3] Our colleagues have explored ways to expedite such text analyses, c.f. [13, 2], as have others.[14] We increasingly turn toward extending such "research profiling" to aid in Forecasting Innovation Pathways (FIP).[7]

To state the obvious -- not all texts behave the same. 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 here. In other analyses, we extend to business and attendant popular press coverage of topics (e.g., the Factiva database) – for example, also concerning Dye-Sensitized Solar Cells (DSSCs).[15, 16] 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. It also seeks to retain chemical formulas.

2.2 Term Clumping

As Bookstein discusses, the concept of clumping is similar to that of clustering, while clumping concerns the objects appearing in 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 distinguish the content-bearing words. It can also consider statistical properties of the words or phrases, considering semantic connections among terms. [19]

Several of the "six or so"[1] 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 the most common modifiers provided analytical value in classifying British science. [20] He conceives of an inverted U shape that emphasizes analyzing moderately high frequency terms -- 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 several stopword lists of several hundred terms (including some stemming), and a common words in academic/scientific writing thesaurus of some 48,000 terms.[21] We are interested in whether removal of these enhances or, possibly, degrades further analytical possibilities.

A variety of statistical techniques have been brought to bear to help consolidate or cluster terms.[22] These offer the means to go well beyond consolidation of term variants, drawing upon semantic or syntactic associations. Latent Semantic Indexing (LSI) [23, 24] seeks to uncover the latent semantic structure in the data. Many related statistical methods [e.g., Principal Components Analysis (PCA), and Latent Dirichlet Allocation (LDA) [25] or Topic Modeling] are available. [26] 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, thereby going 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, 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, uses Singular Value Decomposition (SVD) to transform the basic terms by documents matrix to reduce ranks (i.e., to replace a large number of terms by a relatively small number of factors,

[1]We say about six steps because some of those contain multiple operations (e.g., we use both general and refined fuzzy matching algorithms), variations (e.g., multiple sets of common terms to be removed, sometimes), or applications (e.g., a general list clean-up routine being run more than once in the sequence of steps).

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.[12] This PCA routine generates a more balanced factor set than LSI (which extracts a largest variance explaining factor first; then a second that best explains 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 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). [27] 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,[28] an iterative scaling method, which gets higher precision of similarity measurement than SVD,[29] a Local Relevancy Ladder Weighted LSI (LRLW LSI) method, which improves text classification,[30] 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,[31] and Maletic & Marcus introduced LSI analysis to identify similarities and concept locations for program understanding.[32, 33] Variations on such text analytics can help get at concepts and relationships in various arenas, including web sites[34] and social media.[35]

We are comparing various term clumping and advanced statistical clumping techniques and combinations thereof. Elsewhere we consider topic modelling (LDA and variations) in more detail, and compare

treatment of a technical dataset (DSSCs) with a less technical one (concerning "management of technology"). [36] Newman et al [37] compare the efficacy of alternative text analytics on DSSC data.

3 DATA and METHODS

3.1 Data

This paper takes "term clumping" as the 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[38] 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 2001 (the inception of DSSC research [39]) through 2011 (not complete for this last year).

3.2 Term Clumping Framework

We construct a framework for "term clumping" (Figure 1), which 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, then, go into depth with empirical detail on the text cleaning and consolidation "six" steps. In the last section of the paper we touch on expert engagement and various extensions building on these basic term clumping operations.

Figure 1 Framework for Term Clumping

3.3 Records and Fields Treatment

Record selection is obviously essential to the analyses, but not our main interest 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, Figure 1 includes consideration of record attributes. For certain analyses one wants to focus on particular record selections. For instance, one might choose to analyse the most cited records to get at 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 year 2011 search set 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 special interest in recent R&D activity to get at "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[12] 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., "what?" topics are being pursued in the R&D activity described. In other analyses, we are keenly interested in "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 is researching what?

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. [Were one dealing with patent records, "Claims" are another important source of topical information.] Here, we consolidate the resulting fields to get **90980 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 one has options in the order in which to perform these operations, and that some may warrant repeat applications 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 (Figure 1 and discussed below). The second column relates the results of application of the steps to the DSSC data.

Table 1. Term Clumping Stepwise Results

Our starting list consists of 90980 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 to include. We begin here with a sequence of applying a number of thesauri containing various common term sets.

4.1.1 Step a. Applying Thesaurus for Common Term Removal

Stopword removal is the step where we remove some general verbs (e.g., am, is, are, do), prepositions (e.g., in, on, of), and articles (e.g., the). We first apply a basic stopwords thesaurus containing 279 terms. This thesaurus consolidates stopwords into one term with that label and also consolidates numbers into

one "NUMBERS" term -- both usually to be discarded. This operation reduces our list from 90980 to 89360 phrases. In terms of the distribution of the removed 1620 phrases:

- 6083stopword instances, occurring in 2635 records – these happen to be consolidated with a few occurrences of the term "numbers" itself (5 records and 6 instances);
- 801 number instances occurring in 464 records.

We next apply a general terms thesaurus. This consolidates some singular with plural nouns, uppercases with lowercases, and tags special characters and some academic terms for removal. As a result, we drop from 89360 to 87769 distinct phrases – a reduction of 1591. Some examples:

- "1-min" is consolidated to "1 min";
- "EFFICIENT SOLAR-CELLS," "solar cell efficiency," and "solar-cell efficiency" are consolidated to "efficient solar cells".

We also apply common academic/scientific terms thesaurus containing some 48,000 terms. [21] A common academic terms thesaurus reduces our list from 87769 to 87589 phrases by removing terms such as "manual," "similar," and "analysis." The 180 terms or phrases removed occurred in 3266 records, 7103 times.

Another common-terms thesaurus then drops the phrase count from 87589 to 85887. Sample terms removed include "claim," "undone," and "yearbooks." The 1702 terms removed occurred in 5966 records, 67333 times.

Next, we try a "Basic English" thesaurus (e.g. "race," "stay," "temperature,"), but it does not remove any additional terms. An XMLencoding thesaurus removes codes such as <inf>, </inf>, ^{, and} to facilitate consolidation with plain English terms. This reduces the list from 85887 to 77872 – a large drop of 8015 (over 9%).

We craft a thesaurus to treat some common DSSC terminology – e.g., "TiO2," "ZnO," "DSSCs" and "Solar Cell." Its application takes the list from 77872 to 75156 – a drop of 2716 phrases, including such consolidations as:

- "Titanium dioxide*" is consolidated to "TiO2";
- "DSSC*," "DSC*," "dye sensitized solar cell*," and similar terms will be consolidated to "DSSCs"

In this vein, we next apply a "DSSCDataFuzzyMatcher Results" thesaurus to our phrase list. This reduces it by 2629, from 75156 to 72527, via such combinations as:

- "rice grain-shaped TiO2 mesostructures" is consolidated to "rice grain-shaped TiO2";
- "semiconductor electrode" is consolidated to "semiconducting electrode"

Previous application of thesauri consolidated trash terms into several "noise terms" (shown in Table 2), but did not remove them completely. Thus, in the this step, we run several "Trash terms" thesauri to remove the special "terms," and also to remove some names of organizations, governments, and companies, such as "United States Abstract," "Chinese Chemical Society," "2009 Elsevier Ltd." Here, we reduce 72527 to 72091 phrases (a drop of 436).

Table 2. Trash Terms List

Presumably, almost all terms starting with non-alphabetic characters are meaningless to our research, such as "1.5 m/s," "1500 degree," etc. Therefore, we remove all of them, although there could be several meaningful terms thereby lost. We run the "NumPunctToSpace.the" thesaurus in VantagePoint, and reduce 72091 to 63812 phrases.

4.1.2 Step b. Fuzzy Matching

In addition to use of general and tailored thesauri, our other main computer-aided cleaning mechanism is fuzzy matching. 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 fuzzing 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 63812 to 58577. A tailored version of this (called "general-85cutoff-95fuzzywordmatch-1 exact.fuz") further drops the phrase set to 53718. So, together this effects a reduction of 4859 phrases removed.

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 could use "Combine Author 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 far fewer terms, but no increase in record count for existing terms, just more instances. Actually, the macro of CTN analysis will combine the low frequency terms to the high frequency terms (target terms) which appear in the same records. Sometimes, the target terms are meaningless for our research, especially for the 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 in the steps (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 53718 terms, we remove the top 8 terms, which appear in more than 1000 records, because they are really general in the DSSC domain. This also means 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. We check the unfinished results wherein the macro has reduced the terms to 52161. However, we also notice that the macro has grouped multiple word terms, including1-word to 8-word

terms, as shown in Table 3. In our experience, terms including 2, 3 or 4 words should be more meaningful than others. Thus, we group these three kinds of terms into a sub-list for further research, containing 37928 terms. We use these 37928 terms for our next steps.

Table 3. Multiply Word Terms Classification

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, we do so in this case after Step 16. Pruning here reduces the phrase count from 37928 to 15299. We then reapply the fuzzy match macro to those 15299 terms, thereby reducing to 14840 terms.

4.1.5 Step e. Screening

We run 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 is both high document and term frequency, while Part 2 is the medium level of document frequency. High frequency terms tend to be of interest. But do we also need to pay more attention to 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 most meaningful terms belong to Part 2. For another example, although 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 threshold of high, medium, and low frequency is also not strict and depends on the actual situation and desired outputs.

Figure 2.TFIDF Analysis Results Evaluation

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 (5,784 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 14840 terms and the group "All." Third, VantagePoint generates the TFIDF value for each term.

DSSCs represent an emerging technology, thus, as discussed in the beginning of this section, we prefer the Part 2 (medium document frequency) TFIDF terms. In this instance, we remove the Top 100 highest TFIDF value terms, such as "counter electrode," "photovoltaic performance," "electron transport," etc. We analyse the remaining 14740 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 14740 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 all of them are out of the top 1000 of the 14740 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 14740 terms, and reduce the terms 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 underlie these methods. That is, we consider terms that occur together in records more frequently than chance would indicate as associated. 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 VantagePoint's PCA analysis. Results are shown in Figure 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.

Figure 3.Factor Map of DSSCs (based on the Top 200 Terms)

Table 4.Clusters and Related Factors of DSSCs

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 term clumping, stopping short of clustering (as just illustrated using PCA).

Table 5.Final Topical Phrases List

4.2 Purposive Methods

For the topical analyses, which are the final 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.

Actually, in this paper, the term clumping steps for technical intelligence mostly belong to the inductive methods, and purposive methods are not the focal points for us. Thus, we will not apply those here. However, the comparison between inductive and purposive methods should be one of the most important topics in our further research. Moreover, some ideas also have been generated. For example, we are trying to pick up the valuable topical factors with "term clumping" processes and extract their nearby verbs, which could be considered as a kind of application, based on TRIZ. Also, we are able to locate each term's frequency of use over time, especially when it appears in the high frequency term list, and this can facilitate Technology Roadmapping.

4.3 Expert Assessment of the Clumped Terms and Clusters

How to engage an expert who is broadly knowledgeable over the domain in the most effective way is always challenging. Usually, experts are busy, difficult to invite into the surveys and workshops, and also occasionally cost much time or money. However, experts provide one 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 7 experts from Georgia Tech, Tsinghua University, Dalian University of Technology, IBM, and Booz Allen Hamilton, Inc., and asked for their judgments.

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 of them 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 one selected 183of 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 "interesting." So dividing each by 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 got at least 2 experts' acceptances; 8 of those got at least 3 experts' acceptances. 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 2nd and 3rd most highly ranked clusters are, respectively, 10th and 5th in their record coverage ranking. This shows that expert interest does not relate neatly to cluster generality.

Table 6.Comparisons with Expert Judgments and PCA Records Coverage

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%) got 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 interesting, 183 of 322 terms were still

judged interesting (57%) – a hearty acceptance rate. Alternatively, the other 6 raters gave 300 votes of term acceptance collectively, and those divided among 183 terms.]

2) Several top high frequency terms got a real low expert ranking. The highest frequency term (cell membrane) got no expert's acceptance; the 2nd term (electrochemical corrosion) got a single expert's acceptance; only the 3rd 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.

5 Result Comparisons

We compare the stepwise results in this section, and also, pick up one sample to show the changes step by step. In Table 7, 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 the "Applying Thesaurus for Common Term Removal" and "Fuzzy Matching" steps. Obviously, a big change with the Top 10 terms occurs after several thesauri are used to remove thousands of common terms, or consolidate term variations.

Table 7.Comparison with stepwise "term clumping" results

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 get 37928 terms, which are phrases containing 2, 3 or 4 words. After that, come the "pruning" and "screening" steps, 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 8 shows the terms starting with variations of "Electron/Electrons/Electronic/Electronics," just 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 around 100 in each step, and the "Fuzzy Matching" and "Combining" steps seem to be particular powerful.

Table 8.Stepwise Changes with "Electron/Electrons/Electronic/Electronics" Sample

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 9) on their ranking, "#R" and "#I." In this instance, there are 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 top 10 terms does not change much.
2) Before the "Screening" step, most changes resulted from "term consolidation," where similar terms were consolidated, thus, both "#R" and "#I" increase;
3) In the "Screening" step, extremely high TFIDF terms are removed, and then, low frequency terms are combined into high frequency terms, which appear in the same record. 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 be workable to concentrate a number of terms into several topics, and prepare for further topical analyses.

Table 9.Stepwise Changes for the Top 10 terms in the "After Screening" list of "Electron/Electrons/Electronic/Electronics" Phrases

6 Discussion

Recent attention to themes like "Big Data" and "MoneyBall" 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.[40] Term clumping, as presented here, offers an important tool set to help move toward 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 generation of a macro that would present the analyst with options as to which cleaning and clumping steps to run, 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 steps to Topic Modeling, and then trying to associate related phrases to help capture the thrust of each topic.

We see potential use of clumped terms and phrases in various text analyses. To mention two relating to competitive technical intelligence (CTI) and Future-oriented Technology Analyses (FTA):

- 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 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.

Acknowledgements

We acknowledge 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 Paul Frey, Jan Youtie, Donghua Zhu, and colleagues in the "Innovation Co-lab" of Georgia Institute of Technology, Beijing Institute of Technology, and Manchester University.

References

[1] A. L. Porter, M. J. Detampel, Technology opportunity analysis, Technol. Forecast. Soc. Change, 49 (1995) 237–255.

[2] D. Zhu, A. L. Porter, Automated extraction and visualization of information for technological intelligence and forecasting, Technol. Forecast. Soc. Change, 69 (2002)495–506.

[3] A. L. Porter, S. W. Cunningham, Tech Mining: Exploiting New Technologies for Competitive Advantage, Wiley, New York, NY, 2005.

[4] Y. Zhang, Y. Guo, X. Wang, D. Zhu, A. L. Porter, A Hybrid Visualization Model for Technology Roadmapping: Bibliometrics, Qualitative Methodology, and Empirical Study, 2011 Global Tech Mining Conference, Atlanta, USA, 2011.

[5] 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.

[6] M. Verbitsky, Semantic TRIZ, The TRIZ Journal, Feb (2004),

http://www.triz-journal.com/archives/2004/ (accessed 20 May 2012).

[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] D. S. Price, Little science, big science and beyond, Columbia University Press, New York, NY, 1986.

[9] 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.

[10] A. F. J. van Raan, Advanced Bibliometric Methods to Assess Research Performance and Scientific Development: Basic Principles and Recent Practical Applications, Research Evaluation, 3(3) (1992) 151-166.

[11] N. De Bellis, Bibliometrics and Citation Analysis, The Scarecrow Press, Lanham, MD, 2009.

[12] VantagePoint,www.theVantagePoint.com (accessed 17 August 2012).

[13] R. J. Watts, A. L. Porter, S. W. Cunningham, D. Zhu, TOAS intelligence mining, an analysis of NLP and computational linguistics, Lecture Notes in Computer Science 1997, 1263 (1997) 323-334.

[14] P. Losiewicz, D. W. Oard, R. N. Kostoff, Textual data mining to support science and technology management, Journal of Intelligent Information Systems, 15(2) (2002)99-119.

[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 Management, 42(2) (2012) 133–149.

[17] A. Bookstein, T. Klein, T. Raita, Clumping properties of content-bearing words, Journal of the American Society for Information Science, 49(2) (1998) 102–114.

[18] A. Bookstein, T. Raita, Discovering term occurrence structure in text, Journal of the American Society for Information Science and Technology, 52(6) (2000) 476–486.

[19] A. Bookstein, K. Vladimir, T. Raita, N. John, Adapting measures of clumping strength to assess term-term similarity, Journal of the American Society for Information Science and Technology, 54(7) (2003) 611–620.

[20] S. W. Cunningham, The Content Evaluation of British Scientific Research, D.Phil. Thesis, Science Policy Research Unit, University of Sussex, Brighton, United Kingdom,1996.

[21] Haywood, S. Academic Vocabulary, Nottingham University. http://www.nottingham.ac.uk/~alzsh3/acvocab/wordlists.htm (accessed 26 May, 2012)

[22] 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).

[23] S. Deerwester, S. T. Dumais, G. W. Furnas, T. K. Landauer, R. A. Harshman, Indexing by latent semantic analysis, Journal of the Society for Information Science, 41(6) (1990) 391–407.

[24] T. K. Landauer, D. S. McNamara, S. Denis, W. Kintsch (Eds.), Handbook of Latent Semantic Analysis, Erlbaum Associates, Mahwah, NJ, 2007.

[25] D. Blei, A. Ng, M. Jordan, Latent Dirichlet allocation, Journal of Machine Learning Research,3 (2003) 993–1022.

[26] M. W. Berry, M. Castellanos, Survey of text mining II: clustering, classification, and retrieval, Springer, New York, NY, 2008.

[27] 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.

[28] H. Thomas, Probabilistic latent semantic indexing, Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval, (1999) 50–57.

[29] 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) 216–223.

[30] 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) 483–490.

[31] 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.

[32] 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) 46–53.

[33] J. I. Maletic, A. Marcus, Supporting program comprehension using semantic and structural information, Proceedings of the 23rd International Conference on Software Engineering, (2001) 103–112.

[34] Q. Mei, L. Xu, M. Wondra, H. Su, C. Zhai, Topic sentiment mixture: modeling facets and opinions in weblogs, Proceedings of the 16th international conference on World Wide Web, 2007.

[35] 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.

[36] A. L. Porter, Y. Zhang, Text Clumping for Technical Intelligence, In: S. Sakurai (Ed.), Text Mining, InTech Publishing, Beijing, China, to appear.

[37] 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.

[38] 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.

[39] 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.

[40] A. L. Porter, W. Read (Eds.), The Information Revolution: Current and Future Consequences,

JAI/Ablex, Westport, CT, 1998.

Appendix

Table 2. Term Clumping Stepwise Results

DSSCs5784 records (WoS + Compendex), 2001-2010	
Field selection	**Title & Abstract (NLP phrases + keywords)**
Phrases with which we begin	90980
Step a.Applying Thesauri for Common Term Removal	
01- Stopword.the	89360
02- GeneralTerm.the	87769
03- AcademicTerms.the	87589
04- Common.the	85887
05- BasicEnglish.the	85887
06- XMLencoding.the	77872
07- GeneralScientificTermsConsolidator.the	75156
08- DSSCDataFuzzyMatcherResults.the	72527
09- Remove.the	72527
10- TrashTermRemover.the	72112
11- Combo general term removal2.the	72091
12- NumPunctToSpace.the	63812

Step b. Fuzzy Matching	
13- General.fuz	58577
14- General-85cutoff-95fuzzywordmatch-1 exact.fuz	53718
Step c. Combining	
15- Combine_Terms_Network.vpm (Optional)	Not Applied Here
16- Term_Clustering.vpm	52161 to 37928*
Step d. Pruning	
17- Remove Single terms	15299
18- General-85cutoff-95fuzzywordmatch-1 exact.fuz	14840
Stepe. Screening	
19- Term Frequency Inverse Document Frequency (TFIDF)	14840 (with the Sequence of TFIDF) to 14740**
20- Combine_Terms_Network.vpm (Optional)	8038
Stepf. Clustering	
21- Principal Components Analysis (PCA)	11 Topical Clusters

**We ran an unfinished "Term_Clustering.vpm" in VantagePoint, and reduced the terms from 53718 to 52161, then, we selected the 37928 terms which contain 2, 3 or 4 words.*

***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 14740 terms for the next steps.*

Table 2. Trash Terms List

	# Records	# Instances	Abe + Ti Phrases + keys
1	5966	67333	**Remove**
2	3759	8411	trash
3	3266	7103	ACAD COMMON
4	2835	13844	
5	296	419	BASIC ENGLISH
6	259	289	NUMBERS

Table 3. Multiply Word Terms Classification

Multiple Word Terms	Number
1 Word Terms	2680
2 Word Terms	18795
3 Word Terms	13962
4 Word Terms	5171
5 Word Terms	2430
6 Word Terms	1187
7 Word Terms	399
8 or more Word Terms	201

Table 4.Clusters and Related Factors of DSSCs

Clusters	Coverage/1924	Coverage/5784	Factors
1	8.42%	2.80%	**Photoelectric property**, Hydrothermal method, Higher conversion efficiency
2	7.80%	2.59%	**Polyethylene oxide**, Polyethylene glycol, Ethylene glycol
3	10.34%	3.44%	**Sol gel**, Sol gel process
4	7.33%	2.44%	**Electron donor**, Electron acceptor, Molecular design
5	22.40%	7.45%	**Ruthenium sensitizers**, Ruthenium complex, Efficient sensitizers, Absorption spectrum, Charge transfer sensitizer, Red shift, Density functional theory, High molar extinction coefficient
6	5.87%	1.95%	**Electric resistance**, Sheet resistance, Internal resistance
7	13.15%	4.37%	**Modulated photocurrent spectroscopy**, Electron diffusion coefficient, Electron traps, Recombination kinetics, Photo-injected electron, 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, Conductive film

**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

Table 6.Comparisons with Expert Judgments and PCA Records 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 with 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	13509	1	4672	13509

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

#R = Number of Records containing that term; #I =Number of Instances -- How many times the terms appear, counting multiple occurrences in a record; Rank is based on the #R.

Table 8.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.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."*

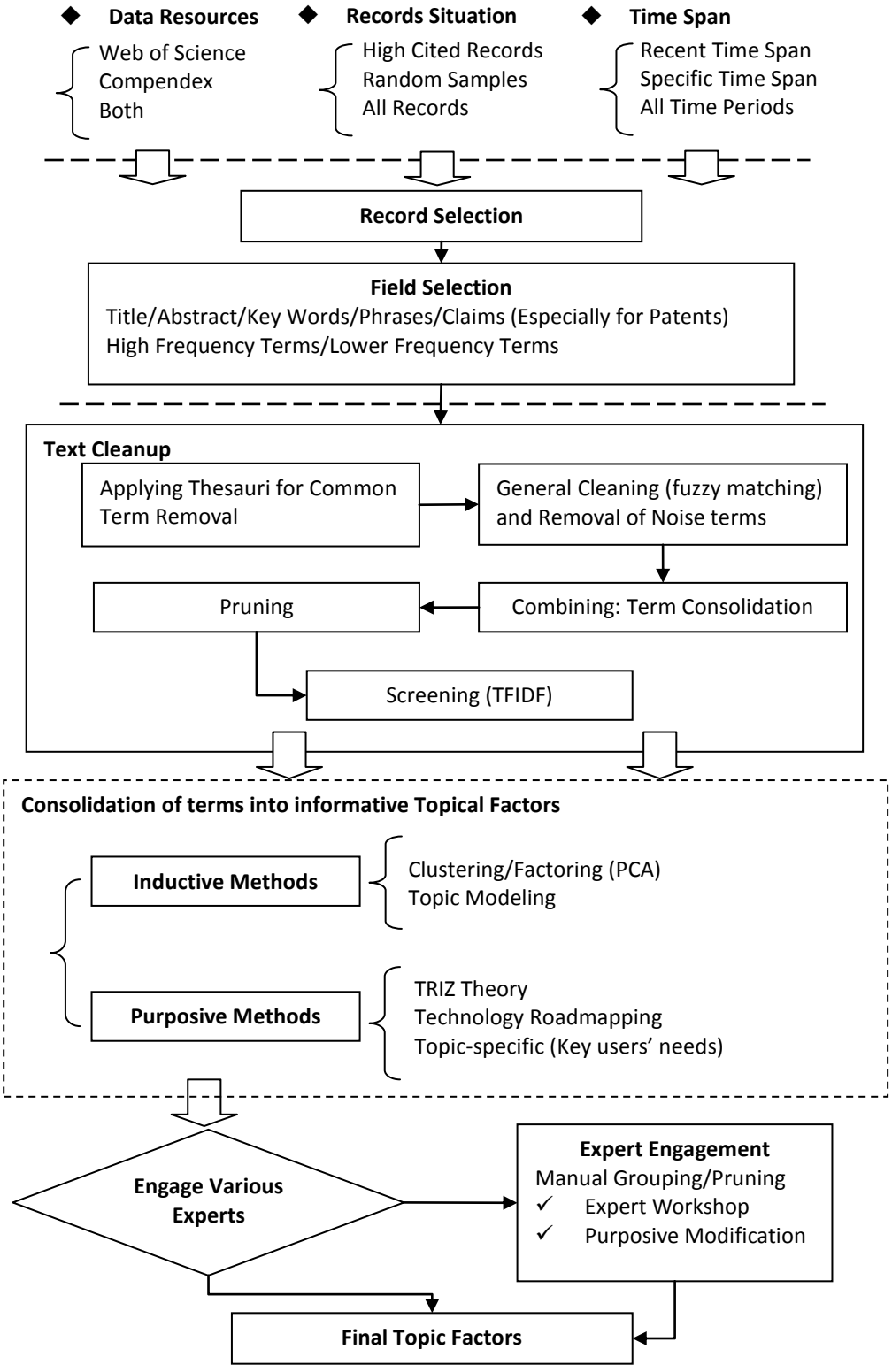

Figure 1 Framework for Term Clumping

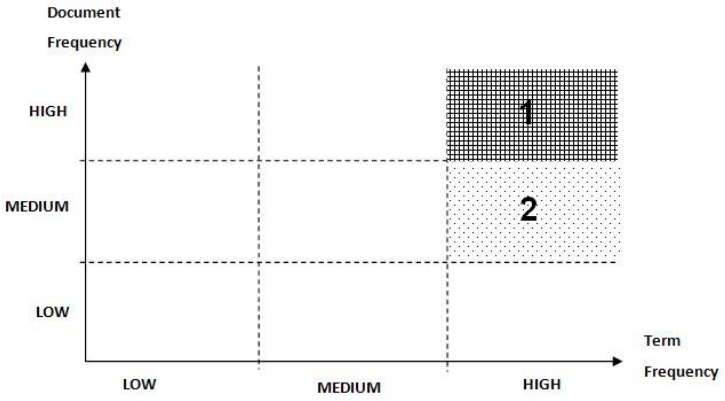

Figure 2.TFIDF Analysis Results Evaluation

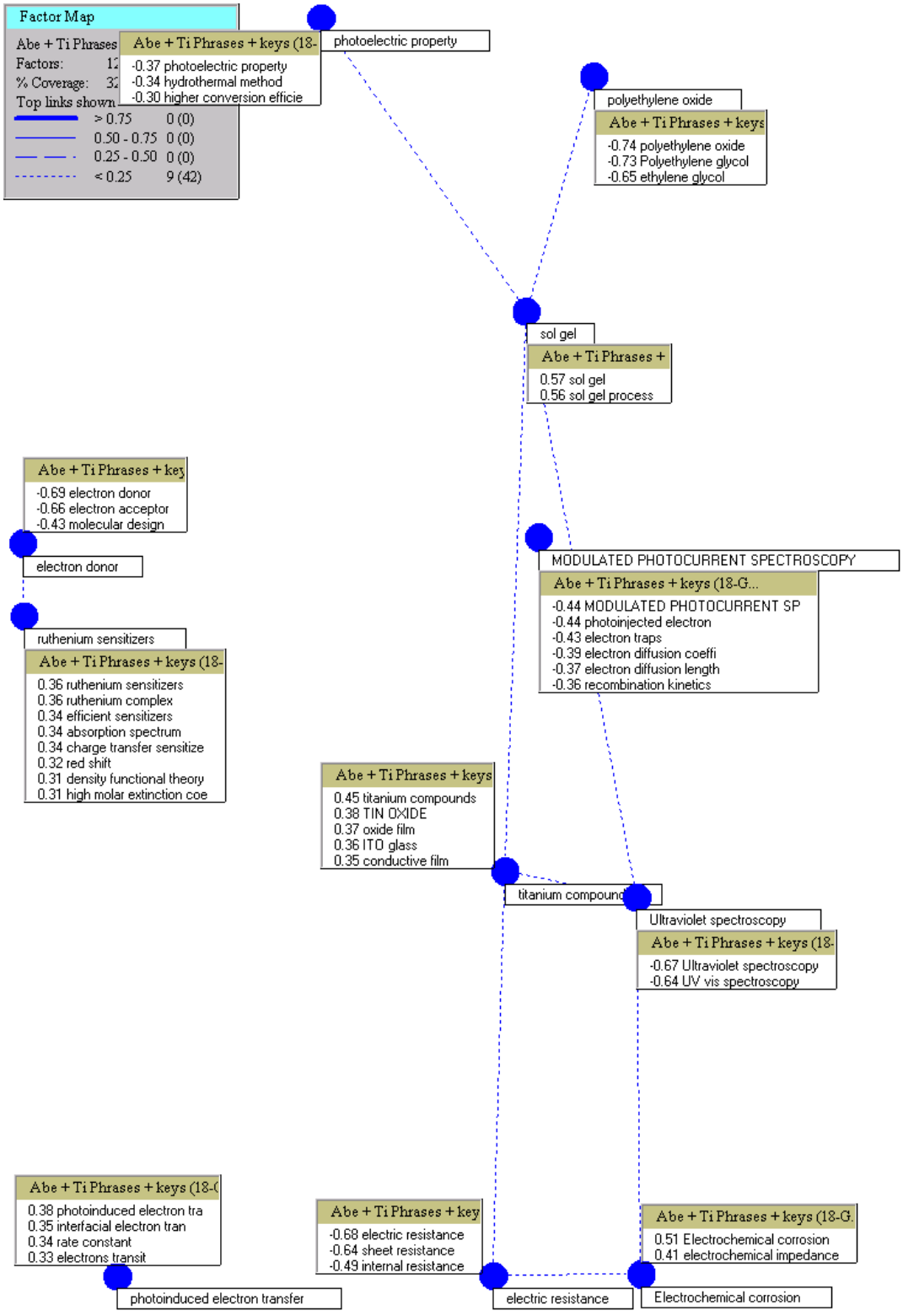

Figure 3.Factor Map of DSSCs (based on the Top 200 Terms)