Chapter 3

Text Clumping for Technical Intelligence

Alan L. Porter and Yi Zhang

Additional information is available at the end of the chapter

http://dx.doi.org/10.5772/50973

1. Introduction: Concepts, Purposes, and Approaches

This development responds to a challenge. Text mining software can conveniently generate very large sets of terms or phrases. Our examples draw from use of VantagePoint (or equivalently, Thomson Data Analyzer – TDA) software [1] to analyze abstract record sets. A typical search on an ST&I topic of interest might yield, say, 5,000 records. One approach is to apply VantagePoint's Natural Language Processing (NLP) to the titles, and also to the abstracts and/or claims. We also take advantage of available topic-rich fields such as keywords and index terms. Merging these fields could well offer on the order of 100,000 terms and phrases in one field (list). That list, unfortunately, will surely contain much noise and redundancy. The text clumping aim is to clean and consolidate such a list to provide rich, usable content information.

As described, the text field of interest can contain terms (i.e., single words or unigrams) and/or phrases (i.e., multi-word noun + modifiers term sets). Herein, we focus on such NLP phrases, typically including many single words also. Some of the algorithms pertain especially to multi-word phrases, but, in general, many steps can usefully be applied to single-word term sets. Here we focus on analyzing NLP English noun-phrases – to be called simply „phrases."

Our larger mission is to generate effective Competitive Technical Intelligence (CTI). We want to answer basic questions of „Who is doing What, Where and When?" In turn, that information can be used to build „innovation indicators" that address users' CTI needs [2]. Typically, those users might be:

- Information professionals (compiling most relevant information resources)
- Researchers (seeking to learn about the nearby „research landscape")
- R&D managers (wanting to invest in the most promising opportunities)

- Science, Technology and Innovation (ST&I) policy-makers (striving to advance their country's competitiveness)

We focus on ST&I information sets, typically in the form of field-structured abstract records retrieved from topical database searches [e.g., Web of Science (WoS), Derwent World Patent Index, Factiva]. These records usually contain a mix of free text portions (e.g., abstracts) and structured text fields (e.g., keywords, publication years). The software uses an import filter to recognize fields (i.e., to know where and how to find the authors and parse their names properly) for particular source sets, such as WoS. VantagePoint can merge multiple datasets from a given source database or from different sources (with guidance on field matching and care in interpreting).

Figure 1 presents our framework for „term clumping." We combine established and relatively novel bibliometric and text mining techniqueswithin this framework. Itincludesa number of steps to process alarge phrase list. The top portion of the figure indicates choices to be made concerning which data resources to mine and selection criteria for the records to be analyzed. The next tier notes additional choices regarding which content-laden fields to process. The following two blocks contain the major foci of this chapter. "Text Cleanup" includes stopword and common term handling, through several steps to consolidate related terms. "Consolidateion of terms into informative topical factors" follows. Here we treat basic "inductive methods." The elements of the Figure flagged with an asterisk (*) are addressed in depth herein.

Figure 1 also points towardinterests for future work. These include"purposive methods," wherein our attention focuses on particular terms based on external criteria – e.g., semantic TRIZ (Theory of Inventive Problem Solving) suggests vital functions and actions indicative of technological innovative potential [3, 4]. The idea is to search the target text fields for occurrences of theory-guided terms and adjacent content.

We are also keenly interested in pursuing single word analyses via Topic Modeling (TM) methods to get at themes of the record set under study. These hold appeal in providing tools that will work well in multiple languages and character sets (e.g., Chinese). The main language dependency that we confront is the use of NLP to extract noun phrases (e.g., VantagePoint's NLP is developed for English text).

The bottom portion of Figure 1 indicates interest in how best to engage experts in such topic identification processes. We distinguish three roles:

- Analyst: Professionals in data retrieval and analysis, who have analytical skills in handling text, but usually don't have domain knowledge
- Expert: Professional researchers in the specific domain, knowledgeable over the domain, and able to describe the current status of the domain at both macro and micro levels;
- Information & Computer Scientist: Covering a range of skills from in-depth programming, through preparation of macros, to operating software to accomplish particular text manipulations.

http://dx.doi.org/10.5772/50973

So defined, engagement of experts presents challenges in terms of motivation, time required, and communication of issues so that the domain experts can readily understand and respond to the analyst's needs. Simple, intermediate stage outputs could have value in this regard.

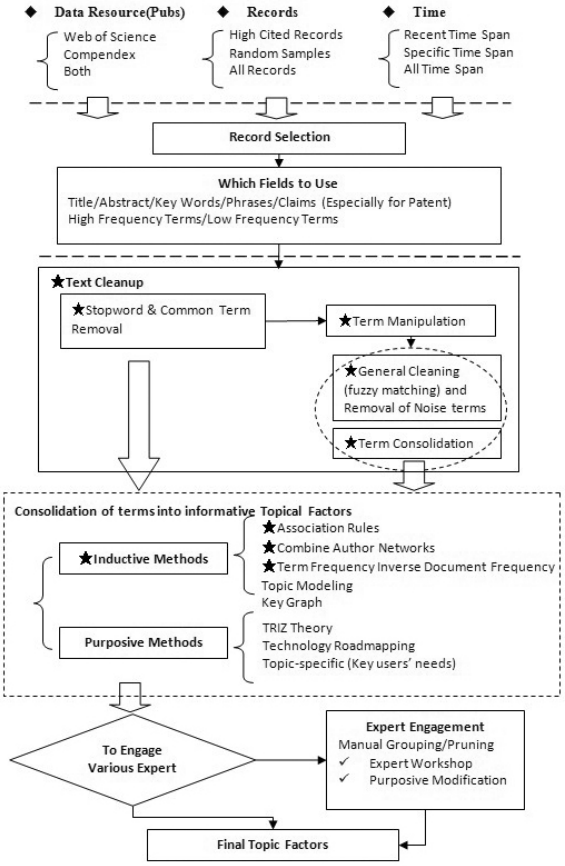

Figure 1. Term Clumping for Technical Intelligence

In summary, this chapter addresses how best to clean and consolidate ST&I phrase lists from abstract record sets. The target is to semi-automate this „inductive" process (i.e., letting the data speak without predetermined identification of target terms). We aim toward semi-automation because the process should be tailorable to study needs. We are exploring a series of text manipulations to consolidate phrase lists. We are undertaking a series of experiments that vary how technical the content is, which steps are performed, in what sequence, and what statistical

approaches are then used to further cluster the phrases or terms. In particular, we also vary and assess the degree of human intervention in the term clumping. That ranges from almost none, to analyst tuning, to active domain expert participation [5-7].

2. Review of Related Literatures

Given the scope of Figure 1, several research areas contribute. This chapter does not address the purposive analyses, so we won't treat literatures on importing index terms, or on TRIZ and Technology RoadMapping (TRM) -- of great interest in suggesting high value terms for CTI analyses.

Several of the steps to be elaborated are basic. Removal of „stopwords" needs little theoretical framing. It does pose some interesting analytical possibilities, however. For instance, Cunningham found that the most common modifiers provided analytical value in classifying British science [8]. 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 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 [9]. We are interested in whether removal of these enhances or, possibly, degrades further analytical steps' performance (e.g., Topic Modeling).

To state the obvious -- not all texts behave the same. Language and the venue for the discourse, with its norms, affect usage and text mining. In particular, we focus on ST&I literature and patent abstracts, with outreach to business and attendant popular press coverage of topics (e.g., the Factiva database). English ST&I writing differs somewhat from „normal" English in structure and content. For instance, scientific discourse tends to include many technical phrases that should be retained, not parsed into separate terms or part-phrases by NLP. VantagePoint's NLP routine [1] strives to do that. It also seeks to retain chemical formulas.

A research community has built around bibliometric analyses of ST&I records over the past 60 or so years, see for instance [10-12]. DeBellis nicely summarizes many facets of the data and their analyses [13]. 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 [1, 2, 14]. We have explored ways to expedite such text analyses, c.f. [15, 16], as have others [17]. We increasingly turn toward extending such „research profiling" to aid in Forecasting Innovation Pathways (FIP), see for example [18].

Over the years many techniques have been used to model content retrieved from ST&I text databases. Latent Semantic Indexing (LSI) [19], Principal Components Analysis (PCA), Sup-

http://dx.doi.org/10.5772/50973

port Vector Machines (SVM), and Topic Modeling (TM) are among the key methods that have come forth [20].

PCA is closely related to LSI. Both use 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, 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.]

VantagePoint uses a special variant of PCA developed to facilitate ST&I text analyses (used in the analyses reported here). 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)[21] He has also empirically compared PCD (inductive) results with a deductive approach based on use of class codes [22].

We apply PCA to term sets to generate co-occurrence based principal components. Because of the familiar use of "clusters," we also use that terminology, although other clustering approaches can yield different forms (e.g., K-means, hierarchical clustering). This PCA approach allows terms to appear in multiple factors

We use the concept, „term clumping," as quite general – entailing various means of text consolidation (e.g., application of thesauri, fuzzy matching, stemming) with noise removal. Bookstein, Raita, and collegues offer a somewhat more specialized, but related, interpretation pointing toward the aim of condensing terminology to better identify content-bearing words [23-25]. Term clumping addresses text (not document) „clustering." Any type of text clustering is based on co-occurrence of words in records (documents). Clustering, in turn, includes many variations plus additional statistical analyses with considerable commonality -- in particular, factor analysis. PCA can be considered as a basic factoring approach; indeed, we call its output principal components „factors. "Similarity among these term grouping approaches arises in that they generally aim to maximize association within clusters and minimize association among clusters. Features to keep in mind include whether terms or phrases being clustered are allowed to be included in multiple clusters or not; whether algorithms yield the same results on rerun or may change (probabilistic methods); and whether useful visualization are generated. Many further variations are available – e.g., hierarchical or non-hierarchical; building up or partitioning down; neural network based approaches (e.g., Kohonen Self-Organizing Maps), and so forth [26]. Research is actively pursuing many refinements, for many objectives, for instance [27]. Our focus is on grouping terms, but we note much complementary activity on grouping documents (based on co-occurrence with particular terms) [26], with special interest in grouping web sites, for instance [28].

Latent Semantic Indexing (LSI) or Latent Semantic Analysis, is a classical indexing method based on a Vector Space Model that introduces Singular-Value Decomposition (SVD) to uncover the underlying semantic structure in the text set. The key feature of LSI is to map those terms that occur in similar contexts into a smaller "semantic space" and to help determine the relationships among terms (synonymy and polysemy) [17, 29, 30]. When applied on co-occurrence information for large text sources, there is no need for LSI to import domain literatures or thesauri (what we call „purposive" or aided text clumping). There are also various extended LSI methods [31]. Researchers are combining LSI with term clumping variations in order to relate synonymous terms from massive content.For example, Maletic and Marcus combine semantic and structural information [32] and Xu et al. seek to associate genes based on text mining of abstracts [30].

Topic modeling is a suite of algorithms that automatically conforms topical themes from a collection of documents [33, 34]. This stream of research begins with Latent Dirichlet Allocation (LDA), which remains the basic algorithm. Topic modelling is an extended LSI method, that treats association probabilistically. Various topic modeling algorithms extend the basic approach, for example [35-44]. Topic modeling is being applied in many contexts – e.g., NLP extension, sentiment analysis, and topic detection.

We are pursuing topic modeling in conjunction with our text clumping development in several ways. We are experimenting to assess whether and which term clumping steps can refine term or phrase sets as input into topic modeling to enhance generation of meaningful topics. We also compare topic modeling outputs to alternative processes, especially PCA performed on clumped phrases. We additionally want to assess whether some form of text clumping can be applied after topic modeling to enhance topic interpretability.

We have also tried, but are not actively pursuing, Key Graph, a kind of visualization technique that treats the documents as a building constructed by a series of ideas and then retrieves these ideas and posts as a summary of original points on the segmentation of a graph [45-47]. Usually, Key Graph has 3 major components: (1) Foundations, which are the subgraphs of highly associated and frequent terms; (2) Roofs, which are terms highly related to the foundations; and (3) Columns, which are keywords representing the relationships between foundations and roofs.

We are especially interested in term grouping algorithms to refine large phrase sets through a sequence of steps. These typically begin with noise removal and basic cleaning, and end with some form of clustering of the resulting phrases (e.g., PCA). „In-between" we are applying several intermediate stage term consolidation tools. Kongthon has pursued an object oriented association rule mining approach [48], with a „concept grouping" routine [49] and a tree-structured network algorithm that associates text parent-child and sibling relationships [50].

Courseault-Trumbach devised a routine to consolidate related phrases, particularly of different term lengths based on term commonality [51]. Webb Myers developed another routine to combine authors. The notion was that, say, we have three papers authored by X. Perhaps two of those are co-authored with Y, and one with Z; and Y and Z never appear as authors on another paper without X. In that case, the operation surmises that Y and Z are likely junior authors,

and eliminates them so that further author analyses can focus on the senior authors or author team. The macro [available at www.theVantagePoint.com] adds major co-authors into the term name. We incorporate these two routines in the present exercises.

Lastly, we consider various quality assessment approaches. Given that one generates clustered text in various forms, which are best? We look toward three approaches. First, we want to ask the target users. While appealing, this also confronts issues – e.g., our PCA output „names" the resulting factors, whereas topic modeling does not. How can we compare these even-handedly? Second are statistical approaches that measure some form of the degree of coherence within clusters vs. among clusters [52]. Third are record assignment tests – to what extent do alternative text clumping and clustering sequences correctly distinguish mixed dataset components? Here we seek both high recall and precision.

3. Empirical Investigation:Two Case Analyses

Figure 1 arrays a wide range of possible term clumping actions. As introduced in the previous sections, we are interested in many of those, but within the scope of this chapter we focus on many of the following steps and comparisons:

Term Clumping STEPS:

a. Fuzzy matching routines

b. Thesauri to reduce common terms

c. Human-aided and topic tailored cleaning

d. Phrase consolidation macro (different lengths)

e. Pruning of extremely high and low frequency terms

f. Combine term networks (parent-child) macro

g. g.TFIDF (Term Frequency Inverse Document Frequency)

h. Term normalization vs. parent database samples

i. PCA variations to generate high, medium, and low frequency factors

j. Topic Modeling

k. Quality assessment of the resulting factors – comparing expert and statistical means

We are running multiple empirical comparisons. Here we compare results on two topical datasets:

"MOT" (for Management of Technology) – 5169 records covering abstract records of the PICMET (Portland International Conference on Management of Engineering and Technology) from 1997 through 2012.

"DSSCs" (for Dye-Sensitized Solar Cells) – 5784 abstract records compiled from searches for 2001-2010 in WoS and in EI Compendex, merged in VantagePoint

Elsewhere, we elaborate on these analyses in various ways. Substantive interpretations of the topical MOT thrusts based on the human-selected MOT terms are examined over time and regions [55]. Comparisons of three MOT analyses -- 1) 3-tier, semi-automatic PCA extraction, 2) PCA based on human-selected MOT terms, and 3) Topic Modeling of unigrams – found notably different factors extracted. Human quality assessment did not yield a clear favorite, but the Topic Modeling results edged ahead of the different PCA's [7]. Additional explorations of the WoS DSSC data appear in [6], comparing Topic Modeling and term clumping-to-PCA – finding quite different emphases in the extracted factors. Zhang et al. [54] track through a similar sequence of term clumping steps on the combined WoS-Compendex DSSC dataset.

Here, we focus on stepping through most of the term clumping operations for these two cases. To avoid undue complexity, we set aside data variations (e.g., stepping through for the WoS DSSC set alone), Topic Modeling comparisons, and quality assessment. As noted, we have done one version of human assessment for the MOT data [7]. We are pursuing additional quality assessments via statistical measures [52] and by comparing how well the alternative analytics are able to separate out record sets from a combination of 7 searches. We also intend to pursue Step h – term normalization based on external (e.g., entire database) frequencies. So, here we treat Steps a-g and i, not Steps h, j, or k.

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, coresponding to the list of steps just above.The second column shows the results of those actions applied in sequence on the MOT data. Blank cells indicate that particular action was not performed on the MOT (or DSSC) dataset. The last row notes additional human-informed analyses done on the MOT data, but not treated here (to recognize that this is a selective presentation). The third column relates the results of application of the steps to the DSSC data, but here we indicate sequence within the column, also showing the resulting term reduction. [So, the Table shows the Term Clumping Steps in the order performed on MOT; this was arbitrary. It could as well have been ordered by the list (above) or in the order done for DSSC data.]

Term Clumping Steps	MOT 5169 PICMET records	DSSCs 5784 records (WoS+Compendex), 2001-2010
Field selection	Title&Abstract NLP phrases	Title&Abstract NLP phrases + keywords
Phrases with which we begin	86014	90980
a-1) Apply general.fuz routine	76398	Applied 10th, reducing 82701 to 74263
b-1) Apply stopwords thesaurus	76105	Applied 1st, reducing 90980 to 89576) and applied 7th, reducing 85960 to 84511

Term Clumping Steps	MOT 5169 PICMET records	DSSCs 5784 records (WoS+Compendex), 2001-2010
b-2) Apply common academic/scientific terms thesaurus	73232	Applied 2d, reducing 89576 to 89403; and applied8th, reducing 84511 to 82739
b-3) multiple tailored cleaning routines -- trash term remover.the; topic variations consolidator.the; DSSC data fuzzy matcher results.the*		Applied such actions as 3d-6th steps, reducing 89403 to 85960; applied 9th, reducing 82739 to 82701;
a-2) Apply general-85cutoff-95fuzzywordmatch-1exact.fuz	69677	Applied 11th, reducing 74263 to 65379
d) Apply phrase consolidation macro (different lengths)	68062	Applied 4th, reducing 89355 to 86410
e) Prune (remove phrases appearing in only 1 record)	13089	Applied 12th, reducing 65379 to 23311
c-1) Apply human-aided and general.fuz routine		Applied 13th, reducing 23311 to 21645
c-2) Manual noise screens (e.g., copyrights, stand-alone numbers)		Applied 14th, reducing 21645 to 20172
f) Apply combine term networks (parent-child) macro	10513	Applied 15th, reducing 20172 to 8181
g) Apply TFIDF	1999	Applied 16th, reducing 8181 to 2008
i) Auto-PCA: highest frequency; 2d highest; 3d highest	201, 256, 299	203;214;230
PCA factors	9 factors (only top tier)	12 (only top tier)
c-3) Tuned phrases to 7164; reviewed 15 factors from 204 top phrases; reran to get final PCA		

Table 1. Term Clumping Stepwise Results

*a compilation of phrase variations that VantagePoint's "List Cleanup" routine suggested combining [e.g. – various singular and plural variations; hyphenation variations; and similar phrases such as "nanostructured TiO2 films" with "nanostructured TiO2 thin films"]

Some steps are broken out in more detail – e.g., Step a -- Fuzzy matching routines – is split into use of VantagePoint's general matching routine (a-1) and application of a variant tuned-for this term clumping (a-2). Note also that some steps appear more than once, especially for the DSSC clumping.

For Step b – application of thesauri to remove common terms – we distinguish the use of a modest size stopwords thesaurus (fewer than 300 words) as Step b-1 and the application of the 48,000 term thesaurus of common academic/scientific terms as Step b-2.

Step c -- Human-aided and topic tailored cleaning (Steps c-1, c-2 & c-3) groups a variety of „obvious" cleaning routines. Our dilemma is whether to eliminate these, to facilitate development of semi-automated routines, or to include them, for easy improvement of the term consolidation? In the MOT term clumping reported in Table 1, we essentially avoid such cleaning. In the DSSC step-through, we include limited iterations of human-aided cleaning to see whether this makes a qualitative difference by the time the progression of steps is completed. [It does not seem to do so.]

Step d -- Phrase consolidation macro – consolidates only a modest percentage of the phrases (as applied here, reducing the phrase count by 2.3% for MOT and by 3.3% for DSSCs), but the improvements appear worthwhile. For instance, combining "Dye-Sensitized Solar Cells" with "Sensitized Solar Cells" can provide important conceptual concentration.

Step e – Pruning – is simply discarding the phrases that appear in only one record. Those would not add to co-occurrence based analyses. The challenge is to sequence pruning after consolidation so that potentially useful topical information is not discarded. Pruning is the overwhelmingly potent step in reducing the term or phrase counts. For MOT, it effects a reduction of 81%; for DSSCs, 64%.

Step f -- Combine term networks (parent-child) – appears a powerful reducer. As discussed, Webb Myers devised this macro to consolidate author sets.We apply the macro to the phrases field, showing sizable reductions for MOT (19.7%) and DSSCs (59.4%). The macro will combine major co-occurring terms in the new phrase name with a "&" between them. It also results in terms that appear in a single record being combined into a single phrase [hence, we perform the Pruning step prior to applying this macro].

Step g – TFIDF – strives to distinguish terms that provide specificity within the sample set.For example, if some form of „DSSC" appears in nearly every DSSC record, this would not be a high-value term in distinguishing patterns within the dataset. VantagePoint offers three TFIDF routines – A) un-normalized, B) log, and C) square root. We compared these and proceed with the square root term set for DSSCs, whose 2008 terms are all included in sets A or B. Of the 2008 phrases, 1915 are in both A and B (so differences in this regard are small), with 42 in set A and 51 in set B. For the MOT data, B and C yield the same 1999 terms, whereas A yields 2052. Inspection of the distinct terms find the 78 only in sets B & C to appear more substantive than the 131 terms only in set A, so we opt for the 1999 term result.

Step h is included as a place-holder.On the one hand, Step b aims to remove generally common terms.On another, Step g favors more specific terms within the document set being analyzed. With access to full databases or general samples from sources such as WoS, one could sort toward terms or phrases that are relatively unique to the search set.We have not done that here.

At this stage, we have very large, but clumped, phrase sets. In our two cases, these consist of about 2000 phrases. Consider the illustrative „post-TFIDF" tabulations in Table 2. We be-

http://dx.doi.org/10.5772/50973

lieve these offer rich analytical possibilities. For instance, we could scan their introduction and frequency over time to identify „hot" topics in the field. Or, we could compare organizational emphases across these phrases to advance CTI interests. We might ask technical and/or business experts in the field to scan those 2000 phrases to identify particularly important or novel ones for in-depth analyses.

Steps i and j represent a major „last step" for these sorts of term analyses. Here we explore select PCA steps; elsewhere, as noted, we pursue Topic Modeling [6, 7]. This factoring (~clustering) step reduces thousands of phrases to tens of phrases. If done accurately, this can be game-changing in terms of opening conceptual insights into topical emphases in the field under study.

VantagePoint's PCA routine is now applied as Step i. In these cases we have tried to minimize human-aiding, but we explore that elsewhere [6, 7]. We select three top tiers of terms to be subjected to separate Principal Components Analysis. Such selection can be handled by various coverage rules – e.g., terms appearing in at least 1% of the records. In the present exercises, we set thresholds to provide approximately 200 phrases as input to each of three PCA analyses. We run the default requested number of factors to extract – this is the square root of the number of terms submitted. We review the resulting three sets of factors in terms of recall (record coverage) and determine to focus on just the top tier PCA results here. For DSSCs, the top-tier PCA yields 12 factors that cover 98% of the records, whereas the 2d tier factors cover 47% and the 3d tier only 18%.For the MOT analyses, results are comparable – the 9 top-tier factors cover 90% of the records; 2d tier, 36%; 3d tier, 17%. [We have performed additional analyses of these data, exploring various PCA factor sets, including ones in which we perform post-PCA term cleaning based on inspection of initial results., then re-run PCAFor instance, a very high frequency term might be removed, or simple relations handled by refining a thesaurus (e.g., in one result „Data Envelopment Analysis" and its acronym, DEA, constituted a factor).

Step j is of high interest, and we are exploring several alternative approaches, as mentioned. Here, we just present the high tier set of PCA factors for face validity checks.

4. Term Clumping Case Results

Having stepped through multiple term clumping steps, what do we get? One has many reasonable choices as to which term clumping steps to apply, in what sequence. To get a feel for the gains, let's compare sample results at four Stages:

1. Initial phrase set
2. After the term clumping steps up to TFIDF
3. After TFIDF
4. After PCA

Referring to Figure 1, the Text Cleaning stage, in general, would be carried out in preparation for nearly all further analyses. We would not anticipate aborting that processing partway, except in special cases (e.g., as mentioned in Cunningham's analysis of British science titles). The next stage of consolidating the cleaned and, therefore, partly consolidated phrases, is where interesting choices arrive. Based on the analyses of the MOT and DSSC data, we note the significant effect of selecting the high TFIDF terms. We thus compare the phrase sets at Stage 1 (before cleaning and clumping), Stage 2 (before filtering to the top TFIDF terms), Stage 3 (after TFIDF), and Stage 4 (after applying one of the clustering family of techniques – PCA).

Stage 1 - Initial			Stage 2 - Clumped		
Top 10	# Records	# Instances	Top 10	# Records	#Instances
study	1177	1874	technology	475	1113
results	894	1177	case study	472	931
research	792	1050	applicable	444	998
development	603	829	knowledge	414	1022
analysis	518	690	relationship	356	801
One	494	574	competition	303	699
innovation	465	800	governance	248	517
knowledge	412	750	technology manager	241	526
process	400	506	literature	227	344
industry	399	637	implication	221	327

Table 2. Stages 1 & 2 – Top 10MOT Phrases

Considering the MOT data first, Table 2 compares the ten most frequent terms or phrases as of Stages 1 and 2.As per Table 1, the clumping and, especially single-record term pruning, has reduced from 86014 to 10513 phrases – an 88% reduction. Table 2 lists the highest frequency terms and phrases based on record coverage. For instance, study appears in 1177 of the 5169 records (23%). The Table also shows instances, and we see that study appears more than once in some records to give a total of 1874 instances. MOT is Management of Technology. That said, the terms and phrases after clumping are somewhat more substantive. As one scans down the Stage 2 set of 10513 phrases, this is even more the case. Our sense is that a topical expert reviewing these to tag a set of themes to be analyzed (e.g., to track trends, or institutional emphases) would definitely prefer the clumped to the raw phrases.

In Tables 2-5, we show in how many of the full sample of MOT and DSSC records the particular terms appear. We also show instances (i.e., some terms appear more than once in a record). These just convey the changes in coverage resulting from the various clumping operations applied.

Table 3 shows the „Top 10“ post-TFIDF terms and phrases, based on TFIDF scores. Recall that the 1999 terms and phrases at this Stage 3 are based on an arbitrary threshold – we sought about 2000. Note that term counts are unchanged for terms present in both Stages 2 & 3. TFIDF is not clumping, but rather screening based on occurrence patterns across the 5169 records.

Stage 3 - post-TFIDF			
Top 10	# Records	# Instances	SqRt TFIDF value
Knowledge	414	1022	35.05
technology	475	1113	34.59
applicable	444	998	33.68
relationship	356	801	32.89
competition	303	699	32.57
innovation technology	200	527	32.42
case study	472	931	31.72
technology manager	241	526	30.54
R&D	191	446	30.25
Governance	248	517	29.99
developed country	179	406	29.43

Table 3. Stage 3 – Top 10 MOT Phrases based on TFIDF

Table 4 presents another 10-term sample pair for Stages 1 and 2. Here, we alphabetically sort the phrase lists and arbitrarily take the ten phrases beginning with „knowledge“ or „knowledg“ --i.e., a stem version of the term. Notice that the big consolidation is for the stemmed version of „knowledg,“ for which the record count has gone up a tiny amount (2), whereas the instance count has increased by 272. In general, the term clumping increases term frequencies and consolidates related terms pretty well (but by no means completely).

Table 5 presents the top-tier PCA analysis results. The phrases appearing here tend to be more topically specific than those seen as most frequent at Stages 2 and 3. Only two terms -- „competition“ and „knowledge“ -- happen to be approximately in common. These nine factors pass a face validity check – they seem quite coherent and potentially meaningful to study of the MOT research arena. Naming of the factors is done by VantagePoint, using an algorithm that takes into account relative term loading on the factor and term commonalities among phrases.

Stage 1 Sample	#R	#I	Stage 2 Sample	#R	#I
knowledge	412	750	knowledge	414	1022

knowledge absorption ability KAA	1	1	knowledge acquisition	6	11
knowledge access	1	1	knowledge age	4	10
knowledge accumulated	1	1	knowledge asset	4	5
knowledge accumulation	4	8	knowledge base	14	17
knowledge accumulation model	1	2	knowledge based competencies	2	3
knowledge acquisition	6	11	knowledge based economy	21	28
knowledge acquisition KA	1	1	knowledge based organizational strategy	2	4
knowledge acquisition strategies	1	4	knowledge based perspective	3	4
knowledge across different sectors	1	1	Knowledge Based Product Models	2	2

Table 4. Stages 1 & 2 – 10 Sample MOT Phrases

Note: #R = # of Records; #I = # of Instances

As mentioned, we have done additional analyses of these data. In another PCA, starting with the 10513 terms (pre-TFIDF), we extracted a high frequency term set (112 terms or phrases appearing in 50-475 records). In addition we extracted a second-tier PCA based on 185 terms appearing in 25-49 records, and a third-tier PCA from 763 terms in 10-24 records. Each set was run using VantagePoint default settings for number of factors, yielding, respectively, 7, 9, and 16 factors. Of the present 9 top-tier factors, 3 show clear correspondence to either top or second-tier factors in the 10513-term based PCA; one shows partial correspondence; 5 are quite distinct. Which factor sets are better? Impressionalistically, the 9 post-TFIDF factors seem reasonable and somewhat superior, but lacking some of the specificity of (7 + 9 + 16 = 32) factors. As noted, we don't pursue the corresponding post-TFIDF PCA second and third tier factors because their record coverage is low.

Examination of DSSC phrase sets shows generally similar progressions as term clumping proceeds.In some respects, results are even more satisfactory with that more technical terminology.In the interest of space, we don't present tables like Tables 2-4 here.But here's a synopsis of one fruitful topical concentration within the DSSC phrase list:

Principle Component (Factor)	High Loading Phrases
Managing Supply Chain	Managing Supply Chain
	Supply Chain
Nanotechnology	Nanotechnology
	Commercial
Competing Technologies	Competition
	Capability

http://dx.doi.org/10.5772/50973

	Technology Capability
	Global Competition
	Competing Technologies
Technology Roadmap	Roadmap
	Technology Roadmap
Innovation Process	Innovation Process
	Innovation Activity
	Open Innovation
Knowledge	Knowledge
	Knowledge Manager
	Individual
	Knowledge Creation
	New Knowledge
	Share Knowledge
Project Success	Project Manager
	Project Success
Make Decisions	Make Decisions
	Decision Making Process
Communication Technology	ICT
	Communication Technology

Table 5. Stage 4 – Top Tier MOT PCA Factors and Constituent Phrases

- In the initial 90980 term list, there are 807 terms on "electron/electrons/electronic"
- In the 8181 term list, there are 119 terms on this
- In the 2008 term list, there are 40 terms remaining, such as "electron acceptor," "electron diffusion," "electron injection," "electron transfer," "electronic structure," etc.

Table 6 shows the twelve top-tier DSSC PCA factors and the phrases that load highly on those factors. These results pace a face validity test in that the grouping of terms seems generally sensible. These factors appear to be reasonable candidates for thematic analysis of this solar cell research & development activity.

Principle Component (Factor)	High Loading Phrases
Sol Gel Process	Sol Gel
	Sol Gel Process

Principle Component (Factor)	High Loading Phrases
	Gel-Sol Method
Polymer Electrolyte	Electrolyte
	Polym
	Ionic Liquid
	Polymer Electrolyte
	Gel Electrolyte
	Electrolyte Liquid
	Ionic Conduction
	Gel Polymer Electrolyte
	Electrolysis
	Gelator
	Poly electrolyte
	Temperature Molten-Salt
Conduction Band	Electron Injection
	Conduction Band
	Mobile Electrons
	Density Functional Theory
Coumarin Dye	Organic Dye
	Coumarin Dye
Solar Equipment	Photo Electrochemical cell
	Efficient Conversion
	Solar Energy
	Solar Equipment
Material Nanostructure	Material Nanostructure
	Redox Reaction
Electron Transport	Electron Transpot
	Back Reaction
ZnO	ZnO
	Nanowire
	Nanorod
	Semiconducting zinc compounds
Scanning Electron Microscopy	Scanning Electron Microscopy
	X-ray Diffraction
	Transmission Electron Microscopy
	Electron Microscopy
	X-ray Diffraction Analysis
	X-ray Photoelectron spectroscopy

Principle Component (Factor)	High Loading Phrases
Open Circuit Voltage	Open Circuit Voltage
	Fill Factor
Electrochemical Impedance Spectroscopy	Electrochemical Impedance Spectroscopy
	Electrochemical Corrosion
	Ion Exchange
Nanotube	Nanotube
	Anode
	Nanotube TiO2

Table 6. Stage 4 – Top Tier DSSC PCA Factors and Constituent Phrases

5. 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 [53]. 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

Phrase consolidation advantages stand out in one DSSC example. Starting with some 2000 terms relating to variations of titanium dioxide (e.g., TiO2, TiO_2, TiO_2 film), we reduce to 4 such terms, with the "combine term networks" (Step f) particularly helpful.

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 in conjunction with the list of steps, of which we are actively working on Steps h, j, and k. We are particularly interested in processing unigrams, 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.

Key Acronyms:

CTI - Competitive Technical Intelligence

DSSCs - Dye-Sensitized Solar Cells [one of two topical test sets]

LSI - Latent Semantic Indexing

MOT - Management of Technology [the second of two topical test sets]

NLP - Natural Language Processing

PCA - Principal Components Analysis

ST&I - Science, Technology & Innovation

TM - Topic Modeling

WoS - Web of Science (including Science Citation Index)

6. 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 David J. Schoeneck for devising groundrules for a semi-automated, 3-tier PCA and Webb Myers for the macro to combine term networks.Nils Newman has contributed pivotal ideas as we build our term clumping capabilities and determine how to deploy them.

Author details

Alan L. Porter[1*] and Yi Zhang[2]

*Address all correspondence to: alan.porter@isye.gatech.edu

1 Search Technology, Inc., Norcross, Georgia, USA, and Technology Policy & Assessment Center, Georgia Tech, Atlanta, Georgia, USA

2 School of Management and Economics, Beijing Institute of Technology, Beijing, China

References

[1] VantagePoint. www.theVantagePoint.com, (accessed 20 May 2012).

[2] Porter, A.L., & Cunningham, S.W. (2005). *Tech Mining: Exploiting New Technologies for Competitive Advantage*, New York, Wiley.

[3] Kim, Y., Tian, Y., Jeong, Y., Ryu, J., & Myaeng, S. (2009). Automatic Discovery of Technology, Trends from Patent Text.In. *Proceedings of the 2009 ACM symposium on Applied Computing, ACMSAC2009, 9-12 March 2009,Hawaii, USA*.

[4] Verbitsky, M. Semantic TRIZ. *The TRIZ Journal2004; Feb.*, http://www.triz-journal.com/archives/2004/, (accessed 20 May 2012).

[5] Porter, A. L., Zhang, Y., & Newman, N. C. (2012). Tech Mining to Identify Topical Emergence in Management of Technology. *The International Conference on Innovative Methods for Innovation Management and Policy, IM2012, 23-26 May 2012. Beijing, China*.

[6] Newman, N. C., Porter, A. L., Newman, D., Courseault-Trumbach, C., & Bolan, S. D. (2012). Comparing Methods to Extract Technical Content for Technological Intelligence. *Portland International Conference on Management of Engineering and Technology, PICMET2012, 29 July-2 August, Vancouver, Canada*.

[7] Porter, A. L., Newman, D., & Newman, N. C. (2012). Text Mining to identify topical emergence: Case study on'Management of Technology. *The 17th International Conference on Science and Technology Indicators, STI2012, 5-8 September, Montreal, Canada*.

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

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

[10] Price, D.S. (1986). *Little science, big science and beyond*, New York, Columbia University Press.

[11] Garfield, E., Malin, M., & Small, H. (1978). Citation Data as Science Indicators. *Y. Elkana,et al, (Eds.), The Metric of Science: The Advent of Science Indicators,*, New York, Wiley.

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

[13] De Bellis, N. (2009). *Bibliometrics and Citation Analysis*, Lanham, MD, The Scarecrow Press.

[14] Porter, A.L., & Detampel, M.J. (1995). Technology opportunity analysis. *Technol. Forecast. Soc. Change*, 49, 237-255.

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

[16] Zhu, D., & Porter, A.L. (2002). Automated extraction and visualization of information for technological intelligence and forecasting. *Technol. Forecast. Soc. Change*, 69495-506.

[17] Losiewicz, P., Oard, D.W., & Kostoff, R.N. (2000). Textual data mining to support science and technology management. *Journal of Intelligent Information Systems*, 15(2), 99-119.

[18] Robinson, D.K.R., HuangL., , Guo, Y., & Porter, A.L. Forecasting Innovation Pathways for New and Emerging Science & Technologies. *Technological Forecasting & Social Change*.

[19] Deerwester, S., Dumals, S., Furnas, G., Landauer, T., & Harshman, R. (1990). Indexing by latent semantic analysis. *Journal of the American Society for Information Science*, 41391-407.

[20] FodorI.K., . A survey of dimension reduction techniques. *U.S. Department of Energy, Lawrence Livermore National Lab.9 May 2002.https://e-reports-ext.llnl.gov/pdf/240921.pdf*, (accessed 22 May 2012).

[21] Watts, R. J., & Porter, A. L. (1999). Mining Foreign language Information Resources, Proceedings., *Portland International Conference on Management of Engineering and Technology, PICMET1999, July 1999, Portland, OR, USA;*.

[22] Watts, R. J., Porter, A. L., & Minsk, B. (2004). Automated text mining comparison of Japanese and USA multi-robot research, data mining 2004. *Fifth International Conference on Data Mining, Text Mining and their Business Applications,15-17Sep. 2004, Malaga, Spain;*.

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

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

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

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

[27] Beil, F., Ester, M., & Xu, X. Frequent term-based text clustering. *Proceedings of the 8th International Conference on Knowledge Discovery and Data Mining, KDD2002*, http://dl.acm.org/citation.cfm?id=775110, (accessed 21 May 2012).

[28] Scime, A. (2005). *Web mining: applications and techniques.*, Hershey, PA, Idea Group Pub.

[29] Homayouni, R., Heinrich, K., Wei, L., & Berry, M.W. (2005). Gene clustering by latentsemantic indexing of MEDLINE abstracts. *Bioinformatics*, 21104-115.

[30] Xu, L., Furlotte, N., Lin, Y., Heinrich, K., & Berry, M.W. *Functional Cohesion of Gene Sets Determined byLatent Semantic Indexing of PubMedAbstracts.PLoS ONE 2011*, 6(4), e18851.

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

[32] Maletic, J. I., & Marcus, A. Supporting program comprehension using semantic and structural information. *Proceedings of the 23rd International Conference on Software Engineering, ICSE2001*.

[33] Blei, D., Ng, A., & Jordan, M. (2003). Latent Dirichlet allocation. *Journal of Machine Learning Research*, 3993-1022.

[34] Griffiths, T., & Steyvers, M. (2004). Finding Scientific Topics. *Proceedings of the National Academy of Sciences*, 101 (suppl.1), 5228-5235.

[35] Thomas, H. Probabilistic latent semantic indexing. *Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval, SIGIR1999*.

[36] Ando, R.K. 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, SIGIR2000*.

[37] Li, W., & McCallum, A. Pachinko allocation: DAG-structured mixture models of topic correlations. *Proceedings of the 23rd international conference on Machine learning, ICML2006*.

[38] David, M., Li, W., & Mc Callum, A. Mixtures of hierarchical topics with Pachinko allocation. *Proceedings of the 24th international conference on Machine learning, ICML 2007*.

[39] Wang, X., & McCallum, A. Topics over time: A non-Markov continuous-time model of topical trends. *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining, KDD2006*.

[40] David, M. B., & John, D. H. Dynamic topic models. *Proceeding Proceedings of the 23rd international conference on Machine learning, ICML 2006*.

[41] Gruber, A., Rosen-Zvi, M., & Weiss, Y. *Hidden topic Markov models*, http://www.cs.huji.ac.il/~amitg/aistats07.pdf., Accessed March 20, 2012.

[42] Rosen-Zvi, M., Griffiths, T., Steyvers, M., & Smyth, P. The author-topic model for authors and documents. *Proceeding of the 20th conference on Uncertainty in artificial intelligence, UAI2004*.

[43] Mc Callum, A., Corrada-Emmanuel, A., & Wang, X. The Author-Recipient-Topic Model for Topic and Role Discovery. *Social Networks: Experiments with Enron and Academic Email.,* http://scholarworks.umass.edu/cgi/viewcontent.cgi?article=1024&context=cs_faculty_pubs., Accessed March 20, 2012.

[44] Mei, Q., Xu, L., Wondra, M., Su, H., & Zhai, C. Topic sentiment mixture: modeling facets and opinions in weblogs. *Proceedings of the 16th international conference on World Wide Web, WWW2007*.

[45] Ohsawa, Y., Benson, N.E., & Yachida, M. Keygraph: Automatic indexing by co-occurrence graph based on building construction metaphor. *Proceedings of the Advances in Digital Libraries Conference, ADL1998*.

[46] Tsuda, K., & Thawonmas, R. (2005). KeyGraph for Visualization of Discussions in Comments of a Blog Entry with Comment Scores. *World Scientific and Engineering Academy and Society (WSEAS) Trans. Computers*, 12(4), 1794-1801.

[47] Sayyadi, H., Hurst, M., & Maykov, A. Event Detection and Story Tracking in Social Streams. *Proceeding of 3rd Int'l AAAI Conference on Weblogs and Social Media, ICWSM09, May 17-20, 2009, San Jose, California, USA; 2009*.

[48] Kongthon, A. A. (2004). Text Mining Framework for Discovering Technological Intelligence to Support Science and Technology Management, Doctoral Dissertation,. *Georgia Institute of Technology*, http://smartech.gatech.edu/bitstream/handle/1853/5151/kongthon_alisa_200405_phd.pdf.txt?sequence=2, (accessed 20 May 2012).

[49] Kongthon, A., Haruechaiyasak, C., & Thaiprayoon, S. Constructing Term Thesaurus using Text Association Rule Mining. *Proceedings of the 2008 Electrical Engineering/Electronics, Computer, Telecommunications and Information Technology International Conference, ECTI2008.*

[50] Kongthon, A., & Angkawattanawit, N. Deriving Tree-Structured Network Relations in Bibliographic Databases. *Proceedings of the 10th International Conference on Asian Digital Libraries, ICADL 2007, December 10-13, Hanoi, Vietnam; 2007*.

[51] Courseault-Trumbach, C., & Payne, D. (2007). Identifying Synonymous Concepts in Preparation for Technology Mining. *Journal of Information Science*, 33(6).

[52] Watts, R.J., & Porter, A.L. (2003). R&D cluster quality measures and technology maturity. *Technological Forecasting and Social Change*, 70(8), 735-758.

[53] Porter, A.L., & Read, W. The Information Revolution: Current and Future Consequences. *Westport, CT: JAI/Ablex; 1998*.

[54] Zhang, Y., Porter, A. L., & Hu, Z. An Inductive Method for "Term Clumping": A Case Study on Dye-Sensitized Solar Cells. *The International Conference on Innovative Methods for Innovation Management and Policy, IM2012, 23-26May 2012, Beijing, China; 2012*.

[55] Porter, A.L., Schoeneck, D.J., & Anderson, T.R. (2012). PICMET Empirically: Tracking 14 Management of Technology Topics. *Portland International Conference on Management of Engineering and Technology, PICMET2012, 29 July-2 August, Vancouver, Canada*.

INTECH