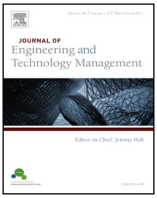
J. Eng. Technol. Manage. 32 (2014) 97–109

Contents lists available at ScienceDirect

Journal of Engineering and Technology Management

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



Comparing methods to extract technical content for technological intelligence

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

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

ARTICLE INFO

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

JEL classification:
C
O

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

ABSTRACT

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

© 2013 Elsevier B.V. All rights reserved.

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

* Corresponding author. Tel.: +1 678 296 2287.

E-mail addresses: newman@iisco.com (N.C. Newman), aporter@searchtech.com (A.L. Porter), newman@uci.edu (D. Newman), ctrumbac@uno.edu (C.C. Trumbach).

0923-4748/$ – see front matter © 2013 Elsevier B.V. All rights reserved.

http://dx.doi.org/10.1016/j.jengtecman.2013.09.001

Introduction

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

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

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

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

Background

Over the years many techniques have been used to model the research indexed in technology databases. This is done to analyze the structure of technical domains and enable analysts to solve technical problems, project the direction a technology is taking, and understand their own place within the technical domain. Latent Semantic Analysis (LSA), PCA, Support Vector Machines (SVM), and Topic Modeling are methods that have been utilized in the past (e.g., Fodor, 2002; Deerwester et al., 1990). The Georgia Tech process has used a variant of PCA to facilitate text analyses, usually focusing on technology topics, but also addressing MOT per se (e.g., Watts and Porter, 1999, 2005). However, the full process requires significant human interaction, iteration, and has scaling issues. Is there a more automated scalable approach? How do the outputs of varying approaches compare with our NLP/PCA-based results?

Since any type of text clustering is based on co-occurrence of words, whether some combination of keywords and/or words contained in a document or abstract, it would seem that the actual clustering algorithm chosen will not bring about large differences in the actual clusters developed. This hypothesis is supported by numerous projects (e.g., Courseault, 2004; Ding, 2003). The basis of these similarities is the fact that the objective of all clustering is to minimize associations between clusters and maximize the relationships within clusters. Different algorithms have different starting points and mechanisms of selection. This statement does not necessarily mean that the results are the same. The details of the chosen clustering algorithm are important to the end result and must be determined by the end goal. The difference, however, is primarily based on factors such as the following: whether the clusters are term clusters or document clusters; whether they are distinct groups or whether certain items can be excluded from any cluster; whether the location of certain words or documents has a distinct location in the space or can have multiple locations; whether the clusters remain the same each time the algorithm is run or, as in the case of probabilistic methods, may change each time the algorithm is run.

Cluster research contains a plethora of techniques and additions to well-known methods designed to improve the ability to find either documents or bits of information, as well as to provide a general landscape of the documents. These techniques fall into a number of categories. *Hierarchical methods* group items in a treelike structure. The methods can start with small groups and aggregate those clusters into larger clusters or start with one or more larger clusters and break those into smaller ones. In contrast, Leouski and Croft (1996) show that *non-hierarchical methods* simply break the corpus into subsets. Partitioning clustering divides the data into disjoint sets. Density-based clustering groups neighboring objects into a cluster, based on density criteria. A cluster is defined by a given density threshold. Statistical clustering methods, such as factor analysis, use similarity measures to partition documents (Halkidi and Vazirgiannis, 2001). While factor analysis is a more linear statistical approach, there are other statistical approaches, such as the probabilistic approach offered by Vinkourov and Girolami (2000). Bayesian Clustering is another probabilistic approach which uses Bayesian probability theory to calculate the probability that a certain object belongs in a certain group (Rauber et al., 2000). Kohonen Self-Organizing Maps is an artificial intelligence approach based on unsupervised neural networks. In general, each of these methods is based on frequency of term co-occurrence. One unique method is offered by Shah (2002). In this method, the semantic relationships among words in the document are captured. The Kohonen Self Organizing Map is used to cluster documents that have the most similar semantic maps.

In the context of text mining, clustering can be utilized in a number of different ways for a variety of purposes. Clustering may also serve as the basis for other types of analysis, such as those presented by Watts et al. (2000). In this paper, an algorithm based on combining various clustering techniques is used to find emerging technologies that accomplish a particular function in a corpus containing over 10,000 publication records. Clustering may be used to discover topic hierarchies giving structure to a corpus and allowing an individual to explore the corpus in a more organized fashion (e.g. Larsen and Chinatsu, 1999). Merkyl and Rauber (1999) use the Self Organizing Map as the basis for an approach designed to uncover associations among documents. Their approach is intended to make explicit the associations among clusters.

Much clustering addresses *document clustering* as a way to maneuver through documents, especially as clustering is being promoted as a visualization method for document retrieval (e.g. Lowden and Robinson, 2002). The increased number of Internet sites has sparked a great interest in this area (e.g. Zamir and Etzioni, 1998). Therefore, much research in this area is based on web pages. Broder et al. (1997) offer a method for determining the syntactic similarity of web documents for the purpose of filtering search results, updating web pages, and identifying copyright violations. Zamir and Etzioni (1998) evaluate clustering algorithms used on web documents and offer an algorithm called Suffix Tree Clustering, which analyzes phrases shared by multiple documents.

Here, the focus is on *term clustering*. We "clump" (combine) terms, first, based on term commonalities, mainly using a combination of thesauri and fuzzy matching routines. We then draw on co-occurrence-based approaches to group terms that tend to appear together in records using "latent" similarities, as introduced in this section. We focus on PCA because we have a version in our software that generates effective Multi-Dimensional Scaling (MDS) factor maps for further analyses that have proven effective in our Tech Mining work. We compare with Topic Modeling because it offers enticing prospects of more automated processing and cross-language capabilities. In particular, our term clumping-to-PCA approach relies on NLP tailored for technical English; Topic Modeling can use single words ("tokens") far less dependent on language structuring.

Data

The topic analyzed here is Dye-Sensitized Solar Cells (DSSCs) – a nanotechnology enhanced, third-generation photovoltaic technology, just entering commercialization. We select this topic because the team is actively researching it (e.g. Guo et al., 2009, 2011, 2012; Ma et al., 2013; Porter et al., 2011; Zhang et al., 2012a,b)]. The data source is Web of Science (expressly, Science Citation Index). The range for the data is 1991 (the year of the first DSSC paper by O'Regan and Gratzel) to 2011. The search strategy is presented in the Appendix. It has been refined over time through a series of analyses and review by persons knowledgeable about solar cells. Compared to routine researcher searches, it tends to be encompassing – i.e., favoring recall over precision to a degree. The intent is to analyze a relatively full R&D landscape.

The 4104 Web of Science records were downloaded and imported into *VantagePoint* text mining software [www.thevantagepoint.com]. The software was used to extract abstract and title phrases using a Natural Language Processing (NLP) module oriented toward S&T text. [For instance, it strives to retain chemical formulas and names as singular entities.] The resulting phrase list contains 64,480 noun phrases. These phrases were cleaned using the general clean-up module within *VantagePoint* to reduce the list to 56,800 phrases. This set of 56,800 DSSC phrases formed the basis of the comparison between the PCA-based approach and the Topic Modeling approach.

Analyses: term clumping and PCA

The PCA approach applied here followed a number of discrete steps. These are of interest for their potential inclusion in a semi-automated methodology of "term clumping" (e.g. Zhang and Porter, 2012). The aim of such an approach is to expedite the reduction of large compilations of term phrases from a document set, such as these 64,480 phrases, to a more manageable, highly informative subset that could be analyzed to gain insight into topical patterns. This is a work in progress. Key steps included here are as follows:

a. *Field selection*: These Web of Science (WOS) records offer four promising topical sources: titles, abstracts, and two types of keywords (a set deriving from the authors that are not always available – covering only 52% of these records; and a set constructed by WOS based on cited reference title terms – for 94% of the present set). In addition, we explored "borrowing" keywords from another source (EI Compendex) and extracting that controlled vocabulary from the WOS records. Use of such meta-data, however, accentuates certain terms to the disadvantage of the raw records. So, for the present case, we favored letting the "records speak" with less intrusion, so we report on the combination of title and abstract phrases extracted using *VantagePoint's* NLP routine.
b. *Basic cleaning*: The **64,480** Title + Abstract phrases were reduced to **56,800** by use of *VantagePoint's* general.fuz "fuzzy matching" routine. This consolidates terms with shared stems ("stemming") and other phrase variations expected to be highly related concepts (e.g., combine singular and plural versions). The basic cleaned set forms the input to Topic Modeling.

c. *Further cleaning*: In VantagePoint, we applied several thesauri (".the" files) to further consolidate term variations and cull "noise." This process reduced the set to 51,960. These include both very general and quite topic-specific collections:
   - stopwords.the – a standard thesaurus provided with the software that uses Regular Expression (RegEx) to batch some 280+ stemmed terms as "stopwords" [e.g. – the, you, and, are, single letters, and numbers]
   - common.the – over 48,000 general scientific terms
   - trash term remover.the – compiled from scanning such WOS phrase collections, including the DSSC records, to remove some 500 noise terms [e.g. – copyright 2011, references, United States Abstract, 650 nm]
   - topic variations consolidator.the – combines variations on a few prominent DSSC terms
   - DSSC data fuzzy matcher results.the – 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 TiO_2 films" with "nanostructured TiO_2 thin films"]

d. *Additional cleaning*: We ran *VantagePoint's* list cleanup routine, again, using a variation of a routine provided for general use – "general-85cutoff-95fuzzywordmatch-1exact.fuz." As the title hints, this was derived by varying parameters offered by the software to adjust fuzzy matching routines. This reduced the set to 47,842 phrases.

e. *Consolidation*: We ran a macro devised by Cherie Courseault Trumbach and Douglas Porter of IISC to consolidate noun phrases of differing numbers of terms. As described by Courseault Trumbach and Payne (2007), this concept-clumping algorithm first identifies a list of relevant noun phrases and then applies a rule-based algorithm for identifying synonymous terms based on shared words. It is intended for use with technical periodical abstract sets. Phrases with, first, 4 or more terms in common; then 3; then 2; are combined and named after the shortest phrase, or most prevalent phrase. In case of conflict, it associates a term with the most similar phrase. This reduced the set to 43,074 terms.

f. *Pruning*: We scanned the term list and added a few terms to trash remover.the, and removed a few more general DSSC terms – reducing just a little to 43,060. We then removed phrases appearing in only a *single record* – reducing the phrase set to 10,350 (after removing 2 cumulative trash terms). So, this is clearly the critical reduction, albeit an extremely simple one to execute.

g. *Parent–child consolidation*: We ran a macro devised by Webb Myers of IISC originally to consolidate junior authors under a more senior collaborator – Combineauthornetworks.vpm. This reduces the set to 7179 terms. To give a sense of term prevalence, here are some frequency benchmarks: 4941 terms associated with 3 or more records; 1086 with 10 or more; 348 with 25 or more; 194 with 40 or more; 49 with 102 or more.

h. *PCA*: *VantagePoint's* "factor mapping" [and/or factor matrix] routine that applies Principal Components Analysis (PCA) was then run on groups of the 7179 terms. These followed a strategic approach devised by Dave Schoeneck of Search Technology to take three tiers of top terms based roughly on the percentage of records that contain them. Several runs, as follows – PCA on the top 194 terms; cleaned terms a bit more to 7164; reran PCA on the top 204 terms (occurring in 37 or more records) – got 15 factors, but the human analyst thought those could be consolidated better, so reran PCA requesting 10 factors; got 10 that look pretty coherent. The terms listed below in Table 1 are from this PCA analysis. To check robustness, another PCA on terms appearing in $\geq$25 records, removing several uninteresting terms (human judgment), was based on 319 terms. Initial result of 18 factors included several that seemed to warrant consolidation; preferred a result with 11 factors that were pretty similar to those presented below (based on the 10-factor solution based on 194 terms).

To give the flavor of term clumping, a "find" on the 56,800 phrases yields 3874 containing the strings "dye-sensiti" or "dye sensiti." These range in frequency from 1 to 2125 records. So, depending on analytical intent, one aim is to consolidate many of these variants before pruning (removing the very low frequency terms). Step d (more aggressive fuzzy matching routine – i.e., changing the adjustable parameters to match more such variants) seeks to do this. Presently we are experimenting

Table 1
PCA factors. Each row shows information about one "factor" (Principal Component), including: a short machine-assigned label for the factor; the percent of variance explained by the factor; and the phrases that load highly on the factor (i.e., that relate most closely to that factor).

PCA 10 factors		
Factors	Percent coverage	Terms
Voc	1.54%	mA cm; fill factor; Voc; open circuit voltage Voc; Jsc; photocurrent density Jsc; eta; open circuit photovoltage Voc; current density Jsc; ISC
Density functional theory DFT	1.37%	Density functional theory DFT; electronic structures
Conduction band	1.22%	TiO; sensitizer; photocurrent; electron injection; conduction band
Electron lifetime	1.08%	Electron transport; electron lifetime; electron diffusion coefficient
Transmission electron microscopy	1.00%	Electron microscopy; transmission electron microscopy; electron microscopy SEM; X ray diffraction; X ray diffraction XRD; X ray photoelectron spectroscopy XPS
Electron donor	0.98%	MW cm irradiance; electron acceptor; photophysical; electron donor
XRD	0.94%	XRD; SEM; TEM
Counter electrode	0.88%	Counter electrode; Pt
Ionic conductivity	0.86%	Electrolyte liquid; ionic conductivity; polymer electrolytes; polymer gel electrolyte
Open circuit voltage	0.83%	mA cm; fill factor; open circuit voltage; overall conversion efficiency

with an iterative approach, that runs fuzzy matching repeatedly, until results stabilize. Step e works to combine varying length phrases with substantial term commonality – e.g., "dye sensitized nanocrystalline TiO_2 anatase" (1 occurrence) could be combined with "dye sensitized nanocrystalline TiO_2 solar cell" (66 records). Another paper illustrates the stepwise changes in addressing a mixed DSSC search set combining results from Web of Science and EI Compendex database searches (Zhang et al., 2013). In other words, Term Clumping is very much an "engineered" solution. The approach focuses on refining discrete steps, automating them to the extent possible, and then combining them into a packaged, semi-automated process.

Analyses: topic modeling

Topic Modeling is a statistical process that discerns topical structure in a collection of text documents (e.g. Blei et al., 2003; Griffiths and Steyvers, 2004). Topic Modeling assumes that each document in the collection incorporates a small number of topics. It simultaneously learns a set of topics to describe the entire collection, and the topics most associated with each document. Formally, each topic is a probability distribution over terms, and is typically displayed by listing the ten to twenty most likely terms.

Topic Models are learned in a fully automated fashion. Like other unsupervised methods, such as PCA, there is no need for an ontology, thesaurus, or dictionary. Instead Topic Modeling works directly from the text data by observing patterns of terms that tend to co-appear in documents, such as *dye* and *sensitized*. The topic model works at a very granular level, assigning a topic label to every word in every document. Topic tags on a per-document level are obtained by aggregating these word-level topic assignments.

The basic form of Topic Modeling – Latent Dirichlet Allocation – evolved as a Bayesian approach for LSA/PCA (e.g. Blei et al., 2003; Griffiths and Steyvers, 2004). Topic Modeling is possibly more suited to text data since it is a model of discrete counts rather than real-valued data. Furthermore, topics from the Topic Model can be easier to understand since topics can be interpreted as probabilities, whereas PCA factors can have positive and negative values (required for orthonormality). But the bigger difference is that the Topic Model uses "T" topics to explain the entire corpus, whereas PCA computes the top-T factors that account for the most variance in the data (that is, there exists no different set of T factors that accounts for more variance). So learning a Topic Model with twice as many topics will result in finer-grained topics, whereas computing twice as many factors in PCA will produce the same top-T factors.

Topic Modeling has been used in a variety of application areas, ranging from information retrieval to research portfolio analysis. One application of relevance to science and technology is the characterization of National Institutes of Health (NIH) funded research, available online as the NIH Map Viewer (Talley et al., 2011). Another Topic Modeling endeavor addressed National Science Foundation awards, generating a thousand topic characterization of that research (Nichols, 2012). Topic Modeling is also highly scalable. One can efficiently and quickly generate Topic Models on millions of documents, making it possibly preferential over PCA for large-scale analyses (e.g. Newman and Smyth, 2009).

Topic Modeling uses a bag-of-words representation of a corpus, where word counts in each document are preserved, but word order is discarded. In this work PCA is applied to term phrases (multi- or single-word), not to unigrams (i.e., single words). One difference is that the Topic Model uses integer term counts, and there is no term frequency/inverse-document-frequency (TF-IDF) weighting of term frequencies that can be used in LSA/PCA.

The basic version of the Topic Model is parameterized by a single input parameter, T, the setting of the number of topics to learn. One can use heuristics to set T, for example based on corpus size or experience. Note that this differs from LSA/PCA where one can request the top-T factors to be computed (and the selection of larger T does not change factors already computed). There are also nonparametric versions of the Topic Model that use the data to learn an appropriate number of topics to explain the data.

Our preprocessing of the 4104 DSSC abstracts followed a slightly different procedure to that used for the PCA analysis. We used all the text from title and abstract, did some simple normalization (lower-casing and removal of punctuation), and limited stemming. Here our focus was on unigrams, so we did no chunking or noun-phrase extraction, except for some limited replacement of frequent bigrams (such as *thin film*). We removed a short list of standard stopwords (e.g. *the*, *and*), as well as some frequently occurring terms (e.g. *journal*, *elsevier*, *all*, *rights*, *reserved*). After tokenization, there were on average 100 terms per document. Given this relatively small collection, Topic Models were learned with T = 10, 15 and 20 topics, running for 400 iterations. The setting of T = 15 topics seemed to have a reasonable resolution, and we present that model here for illustration purposes (for more detailed analyses, one might be interested in more fine-grained topics learned using a higher setting for number of topics, T).

Results for DSSC analyses

Table 2 shows the T = 15 topics learned on the collection of 4104 DSSC abstracts. Each row shows information for one topic. For each topic there is a human assigned label (in the first column) and percentage (second column). The percentage indicates the overall prevalence of a topic, i.e., the percentage of all ~400,000 tokens in the corpus that are assigned to that topic.

We see from the table that the Topic Model learns a range of topics or facets about this collection of DSSC abstracts. It divides various aspects of this technology from topics related to PERFORMANCE (terms like *performance*, *efficiency*, *effect*) to topics related to CIRCUIT INFO (terms like *current*, *voltage*). It also captures the divide of TiO_2-based technology from ZnO-based technology (affirmed as vital by interview). While not shown here, we can examine any one of the 4104 abstracts and show what topics are discussed. Likewise, we can find relevant abstracts for a particular topic, for example, we could rank all abstracts on their relevance to ELECTROLYTE TYPES. Since topics need to represent every word in every document, we see a range of topic types, covering different technologies, aspects of the technology (PERFORMANCE), and topics accounting for various terms that appear less related to technical aspects (e.g. the two PUBLICATION INFO topics).

Discussion

Comparing term clumping-to-PCA with topic modeling

The results of the two analyses are intriguingly different, particularly from the perspective of an analyst viewing the results. The PCA factors appear focused to specific sub-technologies within the

Table 2
Topics learned by the topic model. Each row shows information about one topic, including: a short human-assigned label; the percent of words in the corpus assigned to that topic; the most likely terms in the topic, in order of likelihood.

Topics		
Short label	%	Topic terms
DSSC GENERIC	9%	Dye sensitizer DSSC ruthenium group acid complexes efficient nanocrystalline_TiO_2 organic_dye
PERFORMANCE & EFFICIENCY	9%	TiO_2 layer dye recombination performance surface efficiency effect increase electrode
DEVICE DESCRIPTION	9%	Device material cell photovoltaic dsc organic application efficiency low high semiconductor cost
ELECTROLYTE TYPES	8%	Electrolyte polymer solid_state ionic_liquid iodide polymer_-electrolyte poly gel_electrolyte
ELECTRON TRANSPORT	8%	Electron recombination charge transport diffusion spectroscopy electron_transport kinetic
TiO_2 FILM	8%	TiO_2 film TiO_2_film temperature electrode particle layer prepared thin_film deposition
MOLECULAR CHARACTERISTICS	7%	State dye TiO_2 surface molecular band absorption electronic level excited density functional
TiO_2 NANOSTRUCTURES	6%	TiO_2 nanotube arraytitania nanoparticle mesoporous anatase light structure scattering
CIRCUIT INFO	6%	Cell current voltage short_circuit open_circuit dye photocurrent factor density TiO_2 solid_state
PHOTOELECTROCHEMICAL	6%	Dye light film absorption sensitization electrode photoelectrochemical sensitized photon visible
SPECTROSCOPY	5%	Surface ray spectroscopy electron microscopy properties characterized temperature scanning film
ELECTRODE	5%	Counter_electrode carbon electrode dsc substrate layer resistance performance glass fto
PUBLICATION INFO	5%	Doi chem phy_chem chem_soc mater sol phy mat commun lett sci adv energ nature
ZnO NANOSTRUCTURES	5%	ZnO film nanowire nanorod oxide growth zinc deposition array thin_film nanoparticle nanostructure
PUBLICATION INFO	4%	Chemical physics society applied letter American electrode conversion Chinese material energy

DCCS domain. The percentage of variance within the data explained by these factors is relatively low, but not unexpected, since the pre-processing tended to move the analysis away from the most common terms and toward the middle-high occurring ones. This is intentional. An analysis of high frequency terms from our DSSC work (Ma et al., 2013) finds a good degree of structure within the field and can explain much of the variance. This would indicate that the technology is organized and relatively well developed (key terms are agreed upon, the lexicon is established, definitional battles are minimal). Present interest therefore rests with the mid-to-high tier frequency terms. The PCA analysis shows some variability at the middle tier indicating that DSSC technology is still undergoing transformation. For example, variability in microscopy terms (XRD, SEM, TEM) indicates that best practices are still under development. However, these same terms show some of the weakness of non-expert trained systems – PCA does not understand that TEM and Transmission Electron Microscopy are the same. We are working on an acronym identifier macro to improve this.

Topic Modeling results present a somewhat clearer view of the state for the technology – with a much shorter route to obtain these results. Like the term clumping-to-PCA process, the Topic Modeling process shows some variability of the technology at the second tier. However, it is interesting to note, that key to effective representation of the topics is the labeling – a non-automated (human) process. In this experiment, we only compute topics using unigrams and some frequent bigrams, so the absence of n-grams (multi-word phrases) can make human interpretation of a topic challenging. However, it is possible to post-process the topics, and print out significant (high likelihood) phrases associated with each topic. These can be used to help label the topics.

Do note that in the present analysis we are focusing on the analyst perspective – considering how the analysis interacts with analyst. In other works, we consider statistical validation of topics and factors through standard measures such as precision and recall. For example, we have composed artificial search sets consisting of some seven distinct searches in Web of Science and seven intersecting topic searches in EI Compendex. We then compare the efficacy of term clumping-to-PCA with Topic Modeling in enabling record clustering to match the actual search sets. Even this is not an unequivocal test in that the distance among the search sets is not absolute (i.e., a given record could appear in more than one search set; moreover, the topical distance surely varies among the sets).

Processing topical content and MOT

Validation of topical content manipulation is important. Throughout our DSSC studies, we have enlisted persons knowledgeable in solar cells to check our search strategy, topic content, and other facets. In particular, one Georgia Tech PhD student, Chen Xu, has collaborated in several of the papers (e.g., Guo et al., 2012; Ma et al., 2013) and a faculty member, Jud Ready, in one (Guo et al., 2011).

In July, 2013, three of members of Georgia Tech team had the opportunity to present selected results to Dr. M. Nazeeruddin, a long-time leader in DSSC research at EPFL (Ecole Polytechnique Federale Lausanne – where DSSC research was initiated in 1991 and continuing as the single organizational leader in the field). He found the major topics useful in characterizing DSSC research emphases. He could peruse a list of the higher frequency terms to spotlight pivotal R&D thrusts pertinent to the future of the field. In addition, results were presented to a DSSC conference, receiving generally supportive feedback on the topical characterizations (Guo et al., 2011).

We engaged seven persons knowledgeable about these solar cells to assess results of our DSSC topical treatment of the combined Web of Science and EI Compendex search sets (Zhang et al., 2013). We sought their judgment on which topics and/or PCA factors would be useful in analyzing this emerging technology. For the terms, 183 of 322 elicited endorsement by at least one of six selective raters (249 of 322 got endorsement by at least one of seven). For 11 PCA factors, 10 received endorsement by at least 2 of 7 (8 were selected by 3 or more) as useful in characterizing the field. So the clumped terms and factors do reflect strong value in further study of the field.

Addressing the research question

Can we reduce the time an analyst spends text mining and improve output? Both techniques appear promising. The engineered approach of Term Clumping, despite its complex interior, can be packaged to reduce cycle time and it can be implemented within existing software. We have been conducting a series of DSSC analyses that do use such topical content. Of particular note, term clumping helps consolidate terms and phrases from which the analyst can key on:

- Higher frequency terms – to help distinguish sub-systems, and to identify major topics for further analyses that address:
  - o who is doing what (e.g., Guo et al., 2012, profiled top DSSC research organizations to contrast their research emphases)
  - o trends for major topics (e.g., Guo et al., 2011, comparing major topic trends), and distinguishing "hot topics" (e.g., Guo et al., 2010, contrasting temporal concentrations for TiO_2 and ZnO research)
  - o social network analyses (considering similarities in topical interests; e.g., Guo et al., 2009, presented a research network map based on contrasting national emphases)
- Lower frequency terms – to help identify new R&D emphases based on the year of first use (a standard *VantagePoint* macro provides this in minutes)
- PCA factors – to pursue topical thrust differences – e.g., which organizations are recently most active on Voc issues (as one investigates Merger & Acquisition candidates, likely through patent analyses)?

Or, does EPFL or CAS most emphasize counter-electrode aspects (as one seeks expertise on same)? Zhang et al. (2012b) compare PCA factors with high frequency clumped terms for mapping DSSC research emphases. A combination approach, taking the high-loading terms on the high frequency PCA together with the top 50 clumped phrases, proved helpful in profiling the R&D.

The Topic Modeling approach for addressing "what" issues looks attractive. The fewer steps needed to complete the analysis and little analyst input required suggests significant time savings is possible. The cycle times for current Term Clumping runs can be measured in terms of hours while the Topic Modeling runs take minutes. The issue of labeling topics, which at first pass, appears to be a risk, might ultimately be the step that allows the analyst sufficient time to get grounding in the data. Both approaches look promising enough to pursue deployment within tech mining software. Once implemented, we can conduct further trials to provide a side by side comparison of the approaches.

Conclusions

The key issue with the PCA process is that much of the understanding of the state of the technology comes from the analyst's journey through the process. Table 1 in isolation is not particularly illuminating when disconnected from the process used to create it. A full analysis encompasses addressing not just "what," but also who, when and where. Furthermore, the "what" question addressed by the PCA-based process is, at its core, a process to separate signal from noise – or more precisely, noise from signal. The analyst might not possess the technical expertise to fully understand the nuances of the topic at hand, but an experienced text mining analyst uses the PCA process to ferret out and remove noise from the data, leaving the final PCA step to order the remaining signal.

This noise removing interplay between the analyst and data has been the core of the process at Georgia Tech for over 20 years. It is also the most time consuming step in the analytical process and is more of an art than a science. Over the years we have endeavored to systematize the process (e.g. Porter and Zhang, 2012). However, we remain at the mercy of the data we have available (bibliographic abstracts) and core techniques we rely on to order that data. In the 1990s, our research pushed us toward PCA as a workable approach. We have looked at other techniques. Several techniques, particularly those based on supervised learning, would probably perform better in a single analysis. However, supervised learning would greatly restrict the eclectic range of topics we typically address, forcing us to stick to topics where we feel comfortable conducting supervised training.

The Topic Modeling approach looks promising. This DSSC comparison, although not comprehensive, is still quite helpful. Topic Modeling appears to have utility to reduce the cycle time, the complexity, and analyst input required for a technology analysis. Its ability to separate signal from noise is superior to a blunt implementation of PCA. The fact that it is unsupervised fits our analytical needs. The approach also presents some intriguing possibilities for layering different techniques together. The scalability is very attractive, suggesting the possibility to move beyond abstracts into full text analysis. Use of unigrams reduces language dependency, and that holds great appeal. However, Topic Modeling is not a silver bullet that can be used in isolation. Human analysts, working in concert with other techniques, will still be required to produce an effective assessment of a technology.

Acknowledgments

This work was supported by the Intelligence Advanced Research Projects Activity (IARPA) via Department of Interior National Business Center contract number D11PC20155. Earlier work was supported by the Defense Advanced Research Projects Agency (DARPA), the U.S. Army Tank-automotive and Armaments Command, and the U.S. Army Aviation and Missile Command under Contract DAAH01-96-C-R169. The U.S. government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright annotation thereon.

Disclaimer: The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of IARPA, DoI/NBC, DARPA, or the U.S. Government.

Appendix

The search algorithms and component results for the Web of Science search are as follows:

#1: 4000 records
TS= (((dye-sensiti*) or (dye* same sensiti*) or (pigment-sensiti*) or (pigment same sensiti*) or (dye* same sense)) same (((solar or Photovoltaic or photoelectr* or (photo-electr*)) same (cell or cells or batter* or pool*)) or photocell* or (solar-cell*)))
Annotation: #1 search term is various expression of Dye sensitized solar cell (pigment sensitized solar cell is a kind of DSSC)

#2: 1204 records (#2 not #1: 32 records)
TS=((DSSC or DSSCs) not ((diffuse cutaneous Systemic sclerosis) or (diffuse cutaneous SSc) or (diffuse SSc) or (distributed switch and stay combining) or (Distributed Static Series Compensator*) or (decoupled solid state controller*) or (Active Diffuse Scleroderma*) or (systemic sclerosis) or (diffuse scleroderma) or (Deep Space Station Controller) or (Data Storage Systems Center) or(decompressive stress strain curve) or (double-sideband-suppressed carrier) or (Flexible AC Transmission Systems) or (DSS induced chronic colitis) or (Dynamic Slow-start) or (dextran sulfate sodium) or (disease or patient* or QSRR)))
Annotation: It is the papers which include 1) DSSC but not includes #1 and relate to Dye sensitized solar cell and exculde 2) noisy data.

#3: 330 recordes (#3 not (#1 or #2): 54 records)
TS=((((dye- Photosensiti*) or (dye same Photosensiti*) or (pigment- Photosensiti*) or (pigment same Photosensiti*)) same ((solar or Photovoltaic or photoelectr* or (photo-electr*)) same (cell or cells or batter* or pool*))) not (melanocyte* or cancer))
Annotation: #3 search term is 1) various expression of Dye photo-sensitized solar cell, and use 2) (melanocyte or cancer) to exclude noisy data.*

#4: 188 records (#4 not (#1 or #2 or #3): 18 records)
TS = (((((dye adj (sensiti* or photosensiti*)) and (conduct* or semiconduct*)) same electrode*) and electrolyte*) not (wastewater or waste-water or degradation))
Annotation: #4 search term searches DSSC papers according to1) the component of DSSC and use 2) (wastewater or waste-water or degradation) to exclude noisy data.

Total: 4104 records
#1 or #2 or #3 or #4

References

Blei, D., Ng, A., Jordan, M., 2003. Latent dirichlet allocation. Journal of Machine Learning Research 3, 993–1022.

Broder, A., Glassman, S.C., Manasse, M.S., Zweig, G., 1997. Syntactic clustering of the web. In: Selected Papers from the Sixth International Conference on the World Wide Web. Elsevier Science Publishers Ltd., Santa Clara, CA, pp. 1157–1166.

Courseault, C.R., 2004. A Text Mining Framework Linking Technical Intelligence from Publication Databases to Strategic Technology Decisions. Georgia Tech (dissertation).

Courseault Trumbach, C., Payne, D., 2007. Identifying synonymous concepts in preparation for technology mining. Journal of Information Science 33 (6) 660–677.

Deerwester, S., Dumals, S., Furnas, G., Landauer, T., Harshman, R., 1990. Indexing by latent semantic analysis. Journal of the American Society for Information Science 41, 391–407.

Ding, C., 2003. H.Q. document retrieval and clustering from principal component analysis to self-aggregation. In: Proceedings from the 9th International Workshop on Artificial Intelligence and Statistics, Key West, FL.

Fodor, I.K., 2002. A Survey of Dimension Reduction Techniques. U. S. Department of Energy, Lawrence Livermore National Lab.

Griffiths, T., Steyvers, M., 2004. Finding scientific topics. Proceedings of the National Academy of Science 101 (Suppl. 1) 5228–5235.

Guo, Y., Huang, L., Porter, A.L., 2009. Profiling research patterns for a new and emerging science and technology: dye-sensitized solar cells. In: The Atlanta Conference on Science and Innovation Policy, Atlanta, GA.

Guo, Y., Huang, L., Porter, A.L., 2010. The research profiling method applied to nano-enhanced, thin-film solar cells. R&D Management 40 (2) 195–208.

Guo, Y., Ma, T., Porter, A.L., Rafols, I., 2011. A comparative analysis of Asia-Pacific research thrusts vs. Euro-North American for DSSCs by employing tech mining approach.In: The 6th Aceanian Conference on Dye-sensitized and Organic Solar Cells, Beppu, Japan.

Guo, Y., Xu, C., Huang, L., Porter, A.L., 2012. Empirically informing a technology delivery system model for an emerging technology: illustrated for dye-sensitized solar cells. R&D Management 42 (2) 133–149.

Halkidi, M., Vazirgiannis, M., 2001. Clustering Validity Assessment: Finding the Optimal Partitioning of a Data Set. Athens University of Economics & Business, Department of Informatics.

Larsen, B., Chinatsu, A., 1999. Fast and effective text mining using linear-time document clustering. In: Proceedings of the International Conference on Knowledge Discovery and Data Mining (KDD), San Diego, CA.

Leouski, A.V., Croft, W.B., 1996. An evaluation of techniques for clustering search results. University of Massachusetts at Amherst, Computer Science Department, Amherst.

Lowden, B.G., Robinson, J., 2002. An analysis of file space properties using clustering. In: Proceedings of the World Multi-conference on Systemics, Cybernetics and Informatics: International Conference on Information Systems, Analysis and Synthesis: Computer Science I. v.(5).

Ma, T., Porter, A.L., Ready, J., Xu, C., Gao, L., Wang, W., Guo, Y., 2013. A technology opportunities analysis model: applied to dye-sensitized solar cells for China. Technology Analysis and Strategic Management (in press).

Merkyl, D., Rauber, A., 1999. Uncovering Associations Between Documents. Institut fur Softwaretechnik, ARGE Information Retrieval, Technische Universitat Wiem, Wiem.

Newman, A., Smyth, W., 2009. Distributed algorithms for topic models. Journal of Machine Learning Research 10, 1801–1828.

Nichols, L.G., 2012. Measuring interdisciplinarity at the National Science Foundation. Global Tech Mining Conference, Montreal.

O'Regan, B., Grätzel, M., 1991. A low-cost high efficiency solar-cell based on dye-sensitized colloidal TiO_2 films. Nature 353 (6346) 737–740.

Porter, A.L., Cunningham, S.W., 2005. Tech Mining: Exploiting New Technologies for Competitive Advantage..

Porter, A.L., Zhang, Y., 2012. Text clumping for technical intelligence. In: Sakurai, S. (Ed.), Theory and Applications for Advanced Text Mining. InTech Publishing, ISBN: 978-953-51-0852-8, http://www.intechopen.com/articles/show/title/text-clumping-for-technical-intelligence/.

Porter, A.L., Ma, T., Guo, Y., 2011. Patents in newly emerging science & technology: tracking emergence of dye-sensitized solar cells. In: Patent Statistics for Decision Makers Conference, Alexandria, VA, November 16–17.

Rauber, A., Paralic, J., Pampalk, E., 2000. Empirical Evaluation of Clustering Algorithms. Vienna University of Technology, Department of Software Technology; Technical University of Kosice, Department of Cybernetics and Artificial Intelligence.

Shah, C., 2002. Automatic organization of text documents in categories using self-organizing map (SOM). IEEE's Regional Student Paper Contest.

Talley, E., Newman, D., Mimno, D., Herr, B., Wallach, H., Burns, G., Leenders, M., McCallum, A., 2011. Database of NIH grants using machine-learned categories and graphical clustering. Nature Methods.

Tang, L., Walsh, J.P., 2010. Bibliometric fingerprints: name disambiguation based on approximate structure equivalence of cognitive maps. Scientometrics 84 (3) 763–784.

Vinkourov, A., Girolami, M., 2000. A probabillistic hierarchical clustering method of organizing collections of text documents. University of Paisley, Department of Computing and Information Systems, Computational Intelligence Research Unit, Paisley.

Watts, R.J., Porter, A.L., 1999. Mining foreign language information resources. In: Proceedings, Portland International Conference on Management of Engineering and Technology (PICMET), Portland, OR, USA.

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

Watts, R.J., Porter, A.L., 2005. Mining conference proceedings for corporate technology knowledge management. In: Portland International Conference on Management of Engineering and Technology (PICMET), Portland.

Watts, R.J., Porter, A.L., 2007. Mining conference proceedings for corporate technology knowledge management. International Journal of Technology Management 4 (2) 103–119.

Watts, R.J., Porter, A.L., Cunningham, S.W., Zhu, D., 1997. TOAS intelligence mining, an analysis of NLP and computational linguistics. In: Proceedings of First European Symposium on Principles of Data Mining and Knowledge Discovery, Bergen, Norway. Springer-Verlag, New York, pp. 323–334.

Watts, R.J., Porter, A.L., Newman, N.C., 1998. Innovation forecasting using bibliometrics. Competitive Intelligence Review 9 (4) 11–19.

Watts, R.J., Porter, A.L., Courseault, C., 1999. Functional analysis: deriving systems knowledge from bibliographic information resources. Information, Knowledge, Systems Management 1 (1) 45–61.

Watts, R., Courseault, C., Kapplin, S., 2000. In: Khalil, Lefebvre, L.A., Mason, R.M. (Eds.), Identifying Unique Information Using Principal Component Decomposition. Management of Technology: The Key to Prosperity in the Third Millennium. Edited by Tarek. Elsevier Science.

Watts, R.J., Porter, A.L., Minsk, B., 2004. Automated text mining comparison of Japanese and USA multi-robot research, data mining 2004. In: Fifth International Conference on Data Mining, Text Mining and their Business Applications, Malaga, Spain.

Zamir, O., Etzioni, O., 1998. Web document clustering: a feasibility demonstration. In: Annual ACM Conference on Research and Development in Information Retrieval Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. ACM Press, New York, pp. 46–54.

Zhang, Y., Porter, A.L., 2012. Tech mining management of technology research abstracts to track development of the field. In: International Conference on Innovative Methods for Innovation Management and Policy (IM2012).

Zhang, Y., Porter, A.L., Gomila, J.M.V., 2012a. Text mining methods for consolidating topical factors: topical analyses, triz, and case study on dye-sensitized solar cells. In: Proceedings of The 8th International Conference on Webometrics, Informetircs and Scientometrics and 13th COLLENT, Seoul, Korea.

Zhang, Y., Porter, A.L., Hu, Z., 2012b. An inductive method for term clumping: a case study on dye-sensitized solar cells. In: The International Conference on Innovative Methods for Innovation Management and Policy (IM2012), Beijing.

Zhang, Y., Porter, A.L., Hu, Z., Guo, Y., Newman, N.C., 2013. Term clumping for technical intelligence: a case study on dye-sensitized solar cells. Technology Forecasting and Social Change (in press).

Zhu, D., Porter, A.L., 2002. Automated extraction and visualization of information for technology intelligence and forecasting. Technological Forecasting and Social Change 69, 495–506.

Zhu, D., Porter, A.L., Cunningham, S., Carlisle, J., Nayak, A., 1999. A process for mining science & technology documents databases, illustrated for the case of 'knowledge discovery and data mining. Ciencia da Informacao 28 (1) 1–8.