frontiers in Research Metrics and Analytics

ORIGINAL RESEARCH
published: 29 January 2018
doi: 10.3389/frma.2017.00015

Application of Text Analytics to Extract and Analyze Material–Application Pairs from a Large Scientific Corpus

*Nikhil Kalathil[1], John J. Byrnes[2], Lucien Randazzese[1], Daragh P. Hartnett[2] and Christina A. Freyman[1]**

[1] Center for Innovation Strategy and Policy, SRI International, Arlington, VA, United States, [2] Artificial Intelligence Center, SRI International, San Diego, CA, United States

When assessing the importance of materials (or other components) to a given set of applications, machine analysis of a very large corpus of scientific abstracts can provide an analyst a base of insights to develop further. The use of text analytics reduces the time required to conduct an evaluation, while allowing analysts to experiment with a multitude of different hypotheses. Because the scope and quantity of metadata analyzed can, and should, be large, any divergence from what a human analyst determines and what the text analysis shows provides a prompt for the human analyst to reassess any preliminary findings. In this work, we have successfully extracted material–application pairs and ranked them on their importance. This method provides a novel way to map scientific advances in a particular material to the application for which it is used. Approximately 438,000 titles and abstracts of scientific papers published from 1992 to 2011 were used to examine 16 materials. This analysis used coclustering text analysis to associate individual materials with specific clean energy applications, evaluate the importance of materials to specific applications, and assess their importance to clean energy overall. Our analysis reproduced the judgments of experts in assigning material importance to applications. The validated methods were then used to map the replacement of one material with another material in a specific application (batteries).

Keywords: machine learning classification, science policy, coclustering, text analytics, critical materials, big data

OPEN ACCESS

Edited by:
Kevin Boyack,
SciTech Strategies, Inc.,
United States

Reviewed by:
Jos Winnink,
Leiden University, Netherlands
Karin Verspoor,
University of Melbourne, Australia

****Correspondence:***
Christina A. Freyman
christina.freyman@sri.com

***Received:** 31 August 2017*
***Accepted:** 26 December 2017*
***Published:** 29 January 2018*

Citation:
Kalathil N, Byrnes JJ, Randazzese L, Hartnett DP and Freyman CA (2018) Application of Text Analytics to Extract and Analyze Material–Application Pairs from a Large Scientific Corpus. Front. Res. Metr. Anal. 2:15. doi: 10.3389/frma.2017.00015

INTRODUCTION

Scientific research and technological development are inherently combinatorial practices (Arthur, 2009). Researchers draw from, and build on, existing work in advancing the state of the art. Increasing the ability of researchers to review and understand previous research can stimulate and accelerate scientific progress. However, the number of scientific publications grows exponentially every year both on the aggregate level and in an individual field (National Science Board, 2016). It is impossible for any single researcher or organization to keep up with the vastness of new scientific publications. The ability to use text analytics to map the current state of the art to detect progress would enable more efficient analyses of data.

The Intelligence Advanced Research Projects Activity recognized the scale problem in 2011, creating the research program Foresight and Understanding from Scientific Exposition. Under this program, SRI and other performers processed "the massive, multi-discipline, growing, noisy, and multilingual body of scientific and patent literature from around the world and automatically generated and prioritized technical terms within emerging technical areas, nominated those that exhibit technical emergence, and provided compelling evidence for the emergence" [Intelligence Advanced Research Projects Activity (IARPA), 2011]. The work presented here applies and extends that platform to efficiently identify and describe the past and present evolution of research on a given set of materials. This work applies text analytics to demonstrate how these computational tools can be used by analysts to analyze much larger sets of data and develop more iterative and adaptive material assessments to better inform and shape government and industry research strategy and resource allocation.

MATERIALS

Ground Truth

The Department of Energy (DOE) has a specific interest in critical materials related to the energy economy. The DOE identifies critical materials through analysis of their use (demand) and supply. The approach balances an analysis of market dynamics (the vulnerability of materials to economic, geopolitical, and natural supply shocks) with technological analysis (the reliance of certain technologies on various materials). The DOE's R&D agenda is directly informed by assessments of material criticality. The DOE, the National Research Council, and the European Economic and Social Committee have all articulated a need for better measurements of material criticality. However, criticality depends on a multitude of different factors, including socioeconomic factors (Poulton et al., 2013). Various organizations across the world define resource criticality according to their own independent metrics and methodologies, and designations of criticality tend to vary dramatically [National Research Council (US), 2008; National Research Council (US) and Committee on Assessing the Need for a Defense Stockpile, 2008; Erdmann and Graedel, 2011; European Economic and Social Committee, 2011; Poulton et al., 2013; European Commission, 2014; Graedel et al., 2015].

Experts tasked with assessing the role of materials must make decisions about what materials to focus on, what applications to review, what data sources to consult, and what analyses to pursue (Graedel et al., 2015). The amount of data available to assess is vast and far too large for any single analyst or organization to address comprehensively. In addition, to the best of our knowledge, previous assessments of material criticality have not involved a comprehensive review of scientific research on material use. [Graedel and colleagues have published extensively using raw data on supply and other indicators to measure criticality, see Graedel et al. (2012, 2015) and Panousi et al. (2016) and the references contained within.] Recent developments in text analytic computational approaches present a unique opportunity to develop new analytic approaches for assessing material criticality in a comprehensive, replicable, iterative manner.

The Department of Energy's 2011 Critical Materials Strategy (CMS) Report uses importance to clean energy as one dimension of the criticality matrix (see **Figure 1**) (US Department of Energy, 2011). In this regard, the DOE report serves as a form of ground truth for the validation of our technique, though the DOE report considered supply risk as the second dimension to criticality, which the analysis described in this paper does not address.

Scientific Publications

Data on scientific research articles was obtained from the Web of Science (WoS) database available from Thomson Reuters (now Clarivate Analytics). WoS contains metadata records. In principle, we could have analyzed this entire database; however, for budget-related reasons, the document set was limited by a topic search of keywords that appear in a document title, abstract, author-provided keywords, and WoS-added keywords for the following:

- The 16 materials listed in the 2011 CMS, or
- The 285 unique alloys/composites of the 16 critical materials.

The document set was also limited to articles published between 1992 and 2011, the 20-year period leading up the DOE's most recent critical material assessment.

The 16 materials listed in the 2011 CMS include europium, terbium, yttrium, dysprosium, neodymium, cerium, tellurium, lanthanum, indium, lithium, gallium, praseodymium, nickel, manganese, cobalt, and samarium. Excluded from our documents set was any publication appearing in 80 fields considered not likely to cover research in scope (e.g., fields in the social sciences, biological sciences, etc.). We used the 16 materials listed above because we were interested in validating a methodology against the 2011 CMI Strategy report, and these are the materials mentioned therein. The resulting data set consisted of approximately 438,000 abstracts of scientific papers published from 1992 to 2011.

METHODS

Text Analytics and Coclustering

The principle behind coclustering is the statistical analysis of the occurrences of terms in the text. This includes the processing of the relationships both between terms and neighboring (or nearby) terms, and between terms and the documents in which they occur. The approach presented here grouped papers by looking for sets of papers containing similar sets of terms. As detailed below, our analytics process meaning beyond simple counts of words, and thus, for example, put papers about earthquakes and papers about tremors in the same group, but would exclude papers in the medical space that discuss hand tremors.

Coclustering is based on an important technique in natural language processing which involves the embedding of terms into a real vector space; i.e., each word of the language is assigned a point in a high-dimensional space. Given a vector representation for a term, terms can be clustered using standard clustering techniques (also known as cluster analysis), such as hierarchical agglomeration, principle components analysis, K-means, and distribution mixture modeling (Hastie et al., 2004). This was first done in the 1980s under the name latent semantic analysis

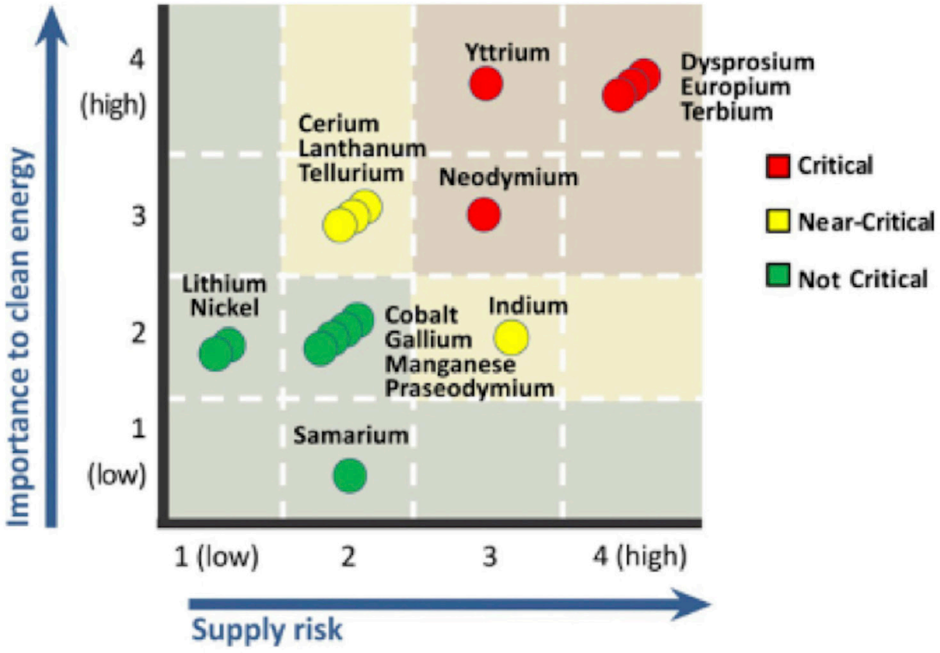


FIGURE 1 | 2011 Critical Materials Importance Analysis matrix, published by experts at the Department of Energy. This matrix served as ground truth for validation.

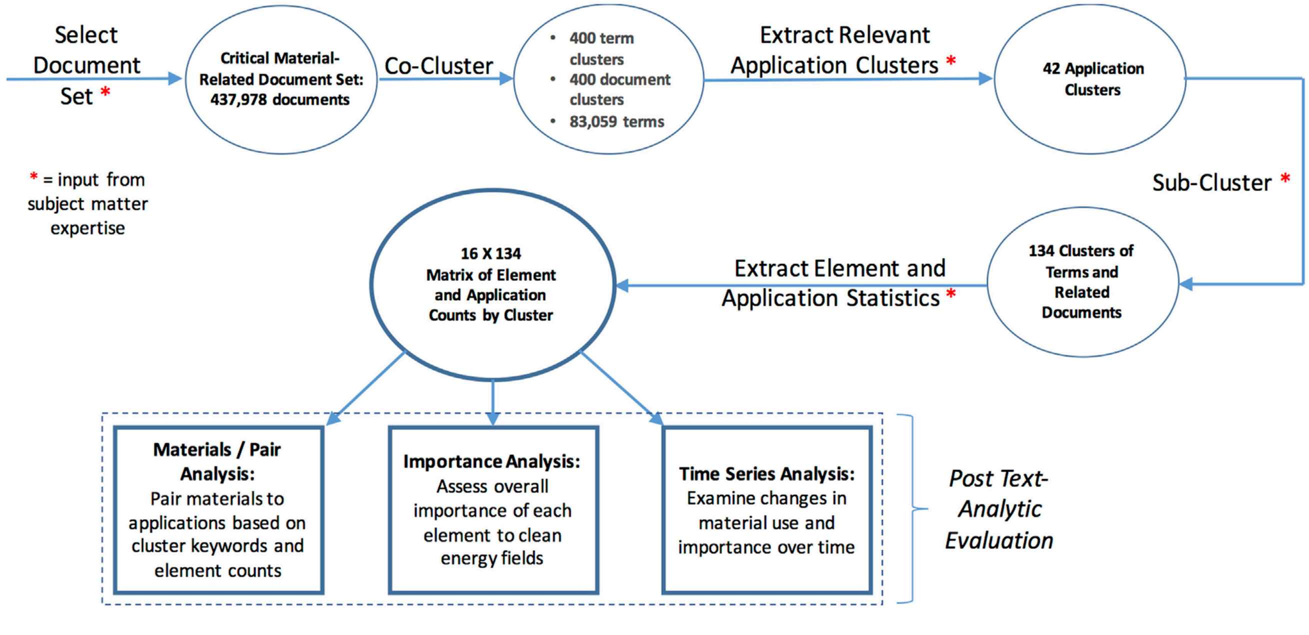


FIGURE 2 | Diagram of the text analytics workflow that this project developed. This work flow was utilized to identify material–application pairs.

(LSA) (Furnas et al., 1987). In the 1990s, neural networks were applied to find embeddings for terms using a technique called context vectors (Gallant et al., 1992; Caid et al., 1995). A Bayesian analysis of context vectors in the late 1990s provided probabilistic interpretation and enabled applying information-theoretic techniques (Gallant et al., 1992; Zhu and Rohwer, 1995). We refer to this technique as Association Grounded Semantics (AGS). A similar Bayesian analysis of LSA resulted in technique referred to as probabilistic-LSA (Hofmann, 1999), which was later extended to a technique known as Latent Dirichlet Allocation (LDA) (Blei et al., 2003). LDA is commonly referred to as “topic modeling” and is probably the most widely applied technique for discovering groups of similar terms and similar documents. Much more recently, Google directly extended the context vector approach of the early 1990s to derive word embeddings using much greater computing power and much larger datasets than had been used in the past, resulting in the word2vec product, which is now widely known (Mikolov et al., 2013; Bellemare et al., 2015).

The LDA model assumes that documents are collections of topics and those topics generate terms, making it difficult to apply LDA to terms other than in the contexts of documents. Because coclustering treats the document as a context for a term, any other context of a term can be substituted in the coclustering model. For example, contexts may be neighboring terms or capital letters or punctuation. This allows us to apply coclustering to a much wider variety of feature types than is accommodated by LDA. In particular, "distributional clustering" (clustering of terms based on their distributions over nearby terms), which has been proven to be useful in information extraction (Freitag, 2004a,b) is captured by coclustering. In future work, we anticipate recognizing material names and application references using these techniques.

Word embeddings are primarily used to solve the "vocabulary problem" in natural language, which is that many ways exist to describe the same thing, so that a query for "earthquakes" will not necessarily pick up a report on a "tremor" unless some generalization can be provided to produce soft matching. The embeddings produce exactly such a mapping. Applying the information-theoretic approach called AGS led to the development of coclustering (Byrnes and Rohwer, 2005), one of the key text analytic tools used in this research.

Information-theoretic coclustering is the simultaneous estimation of two partitions (a mutually exclusive, collectively exhaustive collection of sets) of values of two categorical variable (such as "term" and "document"). Each member of the partition is referred to as a cluster. Formally, if X ranges over terms x_0, x_1, …, Y ranges over documents y_0, y_1, …, and $\Pr(X = x, Y = y)$ is the probability of selecting an occurrence of term x in document y given that an arbitrary term occurrence is selected from a document corpus, then the mutual information $I(X;Y)$ between X and Y is given by

$$I(X;Y) = \sum_{x,y} \Pr(X = x, Y = y) \log \frac{\Pr(X = x, Y = y)}{\Pr(X = x)\Pr(Y = y)}$$

We seek a partition $A = \{a_0, a_1, \ldots\}$ over X and a partition $B = \{b_0, b_1, \ldots\}$ over Y such that $I(A;B)$ is as high as possible. Since the information in the A, B co-occurrence matrix is derived from the information in the X, Y co-occurrence matrix, maximizing $I(A;B)$ is the same as minimizing $I(X;Y) - I(A;B)$, we are compressing the original data (by replacing terms with term clusters and documents with document clusters) and minimizing the information lost due to compression.

Compound terms were discovered from the data through a common technique (Manning and Schutze, 1999) in which sequences are considered to be compound terms if the frequency of the sequence is significantly greater than that predicted from the frequency of the individual terms under the assumption that their occurrences are independent. As an example, when reading an American newspaper, the term "York" occurs considerably more frequently after the term "New" than it occurs in the newspaper overall, leading to the conclusion that "New York" should be treated as a single compound term. We formalize this as follows. Let $\Pr(X_i = x_0)$ be the probability that the term occurring at an arbitrarily selected position X_i in the corpus is the term x_0. Then, $\Pr(X_i = x_0, X_{i+1} = x_1)$ is the probability of the event that x_0 is seen immediately followed by x_1. If x_0 and x_1 occur independently of each other, then we would predict $\Pr(X_i = x_0, X_{i+1} = x_1) = \Pr(X_i = x_0)\Pr(X_{i+1} = x_1)$. To measure the amount that the occurrences do seem to depend on each other, we measure the ratio

$$\frac{\Pr(X_i = x_0, X_{i+1} = x_1)}{\Pr(X_i = x_0)\Pr(X_{i+1} = x_1)}.$$

As this ratio becomes significantly higher than 1, we become more confident that the sequence of terms should be treated as a single unit. This technique was iterated to provide for the construction of longer compound terms, such as "superconducting quantum interference device magnetometer." The compound terms that either start or end with a word from a fixed list of prepositions and determiners (commonly referred to as "stop-words") were deleted, in order to avoid having sequences such as "of platinum" become terms. This technique removes any reliance on domain-specific dictionaries or word lists, other than the list of prepositions and determiners.

Coclustering algorithms (detailed above) produce clusters of similar terms based on the titles and abstracts in which they appear, while grouping similar documents based on the terms they contain: thus, the name cocluster. In this project, a *document* is defined as a combined title and abstract. The process can be thought of as analogous to solving two equations with two unknowns. The process partitions the data into a set of collectively exhaustive document and term clusters resulting in term clusters that are maximally predictive of document clusters and document clusters that are maximally predictive of term clusters.

Coclustering results can be portrayed as an $M \times N$ matrix of term clusters (defined by the terms they contain) and document clusters (defined by the documents they contain). Terms that appear frequently in similar sets of documents are grouped together, while documents that mention similar terms are grouped together. Each term will appear in one, and only one, term cluster, while each document will appear in one, and only one, document cluster. When deciding on how many clusters to start with (term or document, i.e., M and N) there is a tradeoff between breadth and depth. The goal is to differentiate between sub-topics while including a reasonable range of technical discussion within a single topic. The balance between breadth and depth is reflected in the number of clusters that are created. Partitioning into a larger number of clusters would result in more narrowly defined initial clusters, but might mischaracterize topics that span multiple clusters. On the other hand, partitioning into fewer clusters would capture broader topics. Researchers interested in a more fine-grained understanding of materials use, e.g., interested in isolating a set of documents focused on a narrower scientific or technological topic, would need to sub-cluster these initial broad clusters in later analyses.

Terms in term clusters, and titles and abstracts (i.e., document data) from document clusters reveal the content of each cluster. This information provides the basis for identifying what each term cluster "is about" and for selecting term clusters for further scrutiny. Term clusters of interest were filtered using a glossary, in this case, terms pertaining to common applications of the 16 critical materials. This list of terms was manually created through a brief literature review of common applications associated with the materials in question.

Each term cluster was correlated with each of the 437,978 scientific abstracts in our corpus, and the degree of similarity was determined between each term cluster and each abstract through an assessment of their mutual information. As discussed, the term clusters and document clusters described above were selected so as to maximize the mutual information between terms. In order to find document abstracts most strongly associated with a given term cluster, we want to choose those abstracts which were most predictive of the term cluster. These are the abstracts that contain the words in the cluster, but especially the words in the cluster that are rare in the corpus in general. We formalize this by defining the association of a term cluster *t* and document abstract *d* as the value of the (*t*,*d*)-term of the mutual information formula.

To formalize this, we consider uniformly randomly selecting an arbitrary term occurrence from the entire document set, and we write $\Pr(T = t, D = d)$ for the probability that the term was a member of cluster *t* and that the occurrence was in document *d*. We adopt the maximum likelihood estimate for this probability:

$$\Pr\left(T = t, D = d\right) = \frac{n\left(t,d\right)}{N}$$

where $n(t,d)$ is the number of occurrences of terms from term cluster *t* in document abstract *d* and *N* is the total of number of term occurrences in all documents: $N = \sum_{t,d} n(t,d)$.

We define the association between *t* and *d* by the score:

$$\text{Assoc}\left(t,d\right) = \Pr\left(T = t, D = d\right)\log\frac{\Pr\left(T = t, D = d\right)}{\Pr\left(T = t\right)\Pr\left(D = d\right)}$$

Given this score, document abstracts can be arranged from most associated to least associated with a given term cluster. This was done for all 437,978 abstracts in our corpus and all term clusters. This methodology allowed the identification of those abstracts most closely associated with each term cluster.

Not all disperse term clusters were easy to interpret. As with the construction of an appropriate filtering glossary, deciding on the appropriate number of sub-clusters can be subjective. In general, the more diffuse a term cluster appears to be in its subject matter, the more it should be sub-clustered as a means to separate its diverse topic matter. Even imprecise sub-clustering is effective in narrowing the focus of these clusters. Once we determined the number of sub-clusters to generate, sub-clustering of terms was done in the same general manner as the original coclustering: the same coclustering algorithms were applied only to the terms in the initial cluster, but it is instructed not to discover any new abstract clusters. Rather, the set of terms in a cluster are grouped into the pre-specified number of bins according to similarity in the already existing groups of documents in which the terms appear.

RESULTS

Workflow

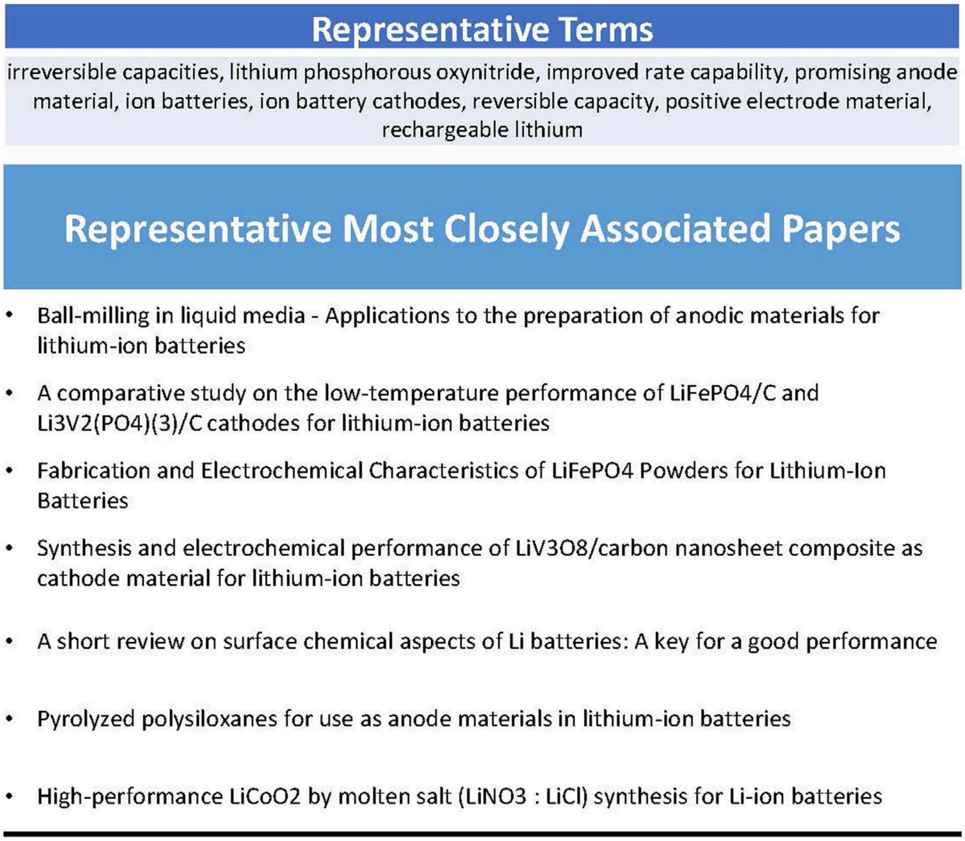
This research resulted in the creation of the following workflow, detailed in (**Figure 2**). The workflow was developed to assess the material–application pairing matrices, and illustrates how text analytics can be used to aid in assessment of material importance to specific application areas, identify pairings of materials and applications, and augment a human expert's ability to monitor the use and importance of materials.

The final data set acquired and ingested into the Copernicus platform contained 437,978 documents. We extracted 83,059 terms from titles and abstracts, including compound (multi-part) terms such as "chloroplatinic acid" and "alluvial platinum." Initial clustering of terms showed some term clusters were very precise, others were less focused. Multiple clustering sizes were experimented with to find an optimal size, which was 400 document clusters and 400 term clusters to start, creating a 400 × 400 matrix of document and term clusters. By manually analyzing the terms in each term cluster, we identified which clusters focused narrowly on areas most relevant to our study, namely those that use materials in clean energy applications. The term cluster in **Figure 3** below, for example, focused around lithium-ion batteries, as indicated by the representative term list and an analysis of the most closely associated papers.

Analysis relies on our ability to extract meaningful statistics about the dataset from term clusters, which can be poorly defined, as seen in the example term cluster in **Figure 4**. When term clusters are poorly defined, we have limited ability to interpret the statistics we extract. There are multiple ways to differentiate term clusters. Principal among these was the division between term clusters that cover basic scientific research versus those that focused on specific technological applications. In addition, because this project used the 2011 DOE CMS as a validation source, a differentiation between clean energy and non-clean energy relevance was also necessary, especially when the same material discoveries were used in both a clean energy and non-clean energy context. More broadly, if a researcher is interested in a specific application area, a division between that application area and others can be used as the means for differentiation.

As discussed in the Section "Methods" for this project, we manually determined the number of necessary sub-clusters. For example, one of the glossary terms we used was "bulbs." A term cluster was identified as the bulb cluster. Once it was examined closely, however, it was clear it contained material about both plant bulbs as well as light bulbs, as shown in **Figure 5**. Accordingly, it was clear that this cluster should be split (sub-clustered) into two clusters. After sub-clustering, the term cluster split into two clearly defined clusters.

The process of sub-clustering expanded the list of initial 42 application term clusters into a total of 134 relevant term clusters for analysis. Clusters were considered "relevant" based on the weights that were assigned to them. For this project, we developed a methodology for weighting clusters to replicate the ground truth of the 2011 DOE CMS.



Representative Terms

irreversible capacities, lithium phosphorous oxynitride, improved rate capability, promising anode material, ion batteries, ion battery cathodes, reversible capacity, positive electrode material, rechargeable lithium

Representative Most Closely Associated Papers

- Ball-milling in liquid media - Applications to the preparation of anodic materials for lithium-ion batteries
- A comparative study on the low-temperature performance of LiFePO4/C and Li3V2(PO4)(3)/C cathodes for lithium-ion batteries
- Fabrication and Electrochemical Characteristics of LiFePO4 Powders for Lithium-Ion Batteries
- Synthesis and electrochemical performance of LiV3O8/carbon nanosheet composite as cathode material for lithium-ion batteries
- A short review on surface chemical aspects of Li batteries: A key for a good performance
- Pyrolyzed polysiloxanes for use as anode materials in lithium-ion batteries
- High-performance LiCoO2 by molten salt (LiNO3 : LiCl) synthesis for Li-ion batteries

FIGURE 3 | The coclustering algorithm produced term clusters. This is an example of a narrowly focused cluster.

Weighting

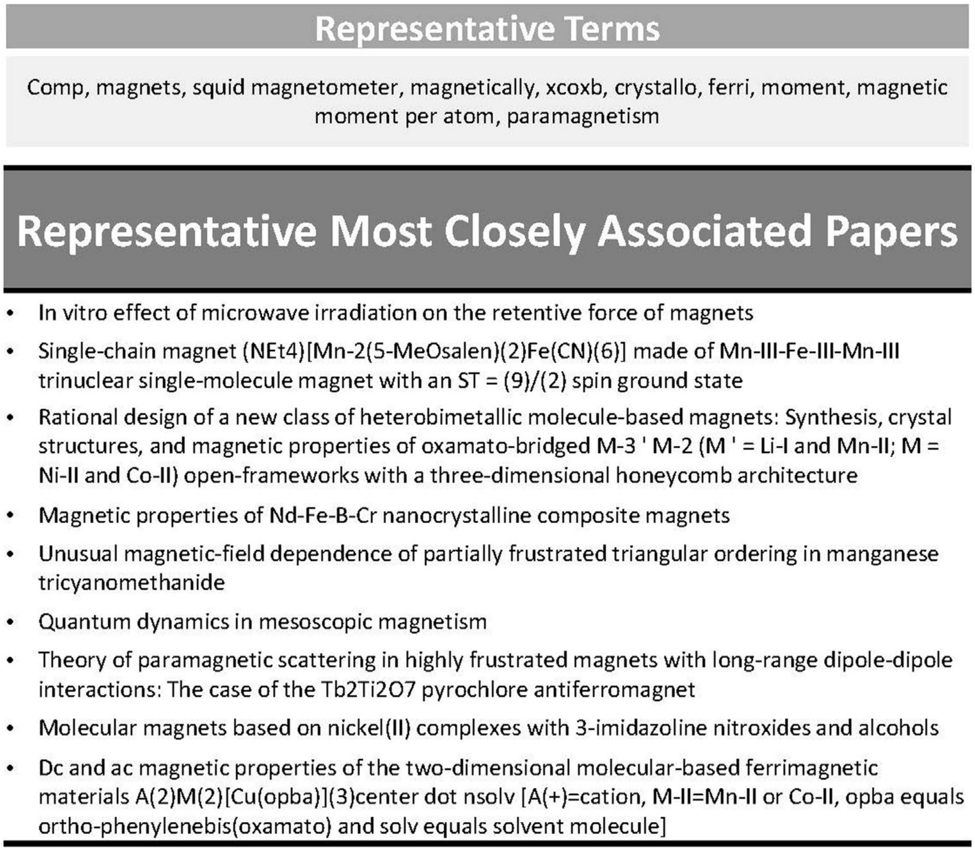
One of the principal questions addressed by the 2011 DOE CMS was the importance of specific materials to clean energy applications. The DOE ranked the 16 materials based on their importance to clean energy, and their supply risk, as detailed in **Figure 1**. We developed a methodology to replicate the *y*-axis of this figure (material importance to clean energy) by combining the data on material distribution over clusters with an assessment of the clean energy importance of each cluster. To assess the clean energy importance of each of these clusters, we developed a clean energy importance weighting. A set of key words was constructed manually and searched the top 500 associated abstracts of each term cluster for these key words. Constructing this glossary is a crucial step that allows researchers to define their key words of interest. Document abstracts were analyzed for mentions of any clean energy field per cluster, and the number of different clean energy fields mentioned across the cluster. The clean energy weighting considers the extent and depth of impact within a cluster, and is equal to:

$$\frac{\text{Total \# of Abstracts Mentioning Any Clean Energy Field}}{\text{\# of Different Clean Energy Fields Mentioned Across the Cluster}}.$$

This weighting essentially captures the number of abstracts per clean energy field. The goal of the weighting is to discount clusters that mention a number of different clean energy fields, but do not discuss these clean energy fields in a substantial or significant manner. Materials that were mentioned frequently in clusters with high clean energy weights can be thought of as "important" to clean energy. The importance of each material to clean energy was determined by counting the number of document mentions in clean energy important clusters (i.e., clusters above some cutoff of minimum clean energy weight).

Material-to-clean-energy field application pairings were derived from term clusters. Recall that for each term cluster, the number of abstracts mentioning any of the 16 materials were determined as well as the number of abstracts mentioning any of 33 clean energy field keywords. In the importance analysis previously discussed, mentions of 33 clean energy fields were aggregated together to establish the clean energy weighting system. For each clean energy field, all term clusters were ranked based on a normalized count of document mentions of that field.

From this ranking, the "top" three term clusters were identified for a specific field for review. Then, the terms and keywords were manually reviewed to ensure that term clusters were actually relevant and the occasional false positives (term clusters that score high based on keyword counts but are not in fact relevant based on human content analysis) were discarded. This manual review, while requiring human intervention, is done



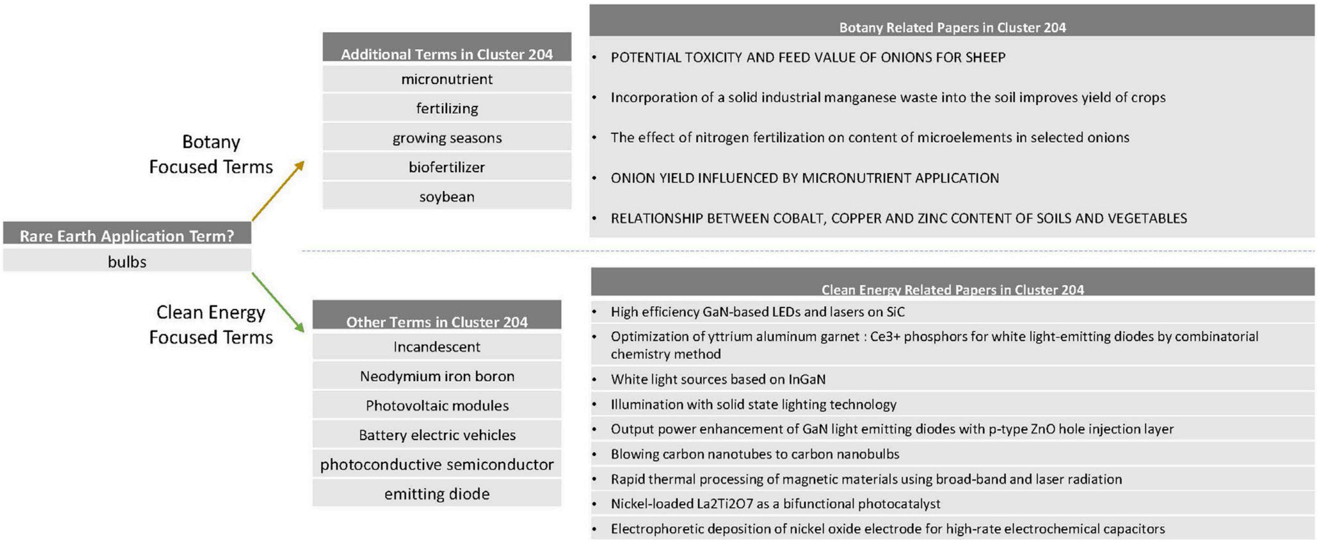
Representative Terms

Comp, magnets, squid magnetometer, magnetically, xcoxb, crystallo, ferri, moment, magnetic moment per atom, paramagnetism

Representative Most Closely Associated Papers

- In vitro effect of microwave irradiation on the retentive force of magnets
- Single-chain magnet (NEt4)[Mn-2(5-MeOsalen)(2)Fe(CN)(6)] made of Mn-III-Fe-III-Mn-III trinuclear single-molecule magnet with an ST = (9)/(2) spin ground state
- Rational design of a new class of heterobimetallic molecule-based magnets: Synthesis, crystal structures, and magnetic properties of oxamato-bridged M-3 ' M-2 (M ' = Li-I and Mn-II; M = Ni-II and Co-II) open-frameworks with a three-dimensional honeycomb architecture
- Magnetic properties of Nd-Fe-B-Cr nanocrystalline composite magnets
- Unusual magnetic-field dependence of partially frustrated triangular ordering in manganese tricyanomethanide
- Quantum dynamics in mesoscopic magnetism
- Theory of paramagnetic scattering in highly frustrated magnets with long-range dipole-dipole interactions: The case of the Tb2Ti2O7 pyrochlore antiferromagnet
- Molecular magnets based on nickel(II) complexes with 3-imidazoline nitroxides and alcohols
- Dc and ac magnetic properties of the two-dimensional molecular-based ferrimagnetic materials A(2)M(2)[Cu(opba)](3)center dot nsolv [A(+)=cation, M-II=Mn-II or Co-II, opba equals ortho-phenylenebis(oxamato) and solv equals solvent molecule]

FIGURE 4 | The coclustering algorithm produced term clusters. This is an example of an unfocused cluster that spans basic science research and non-clean energy research.



Additional Terms in Cluster 204

- micronutrient
- fertilizing
- growing seasons
- biofertilizer
- soybean

Botany Related Papers in Cluster 204

- POTENTIAL TOXICITY AND FEED VALUE OF ONIONS FOR SHEEP
- Incorporation of a solid industrial manganese waste into the soil improves yield of crops
- The effect of nitrogen fertilization on content of microelements in selected onions
- ONION YIELD INFLUENCED BY MICRONUTRIENT APPLICATION
- RELATIONSHIP BETWEEN COBALT, COPPER AND ZINC CONTENT OF SOILS AND VEGETABLES

Other Terms in Cluster 204

- Incandescent
- Neodymium iron boron
- Photovoltaic modules
- Battery electric vehicles
- photoconductive semiconductor
- emitting diode

Clean Energy Related Papers in Cluster 204

- High efficiency GaN-based LEDs and lasers on SiC
- Optimization of yttrium aluminum garnet : Ce3+ phosphors for white light-emitting diodes by combinatorial chemistry method
- White light sources based on InGaN
- Illumination with solid state lighting technology
- Output power enhancement of GaN light emitting diodes with p-type ZnO hole injection layer
- Blowing carbon nanotubes to carbon nanobulbs
- Rapid thermal processing of magnetic materials using broad-band and laser radiation
- Nickel-loaded La2Ti2O7 as a bifunctional photocatalyst
- Electrophoretic deposition of nickel oxide electrode for high-rate electrochemical capacitors

FIGURE 5 | Unfocused clusters were then sub-clustered to produce more useable and focused clusters, as shown in this figure. A blub term cluster was split. The sub-clustering algorithms produced two new clusters, as shown above. The new smaller clusters were more focused than the larger cluster.

on a significantly narrowed set of documents, and thus does not represent a major bottleneck in the process. Having linked clusters to a specific clean energy field, material importance to each field was evaluated by examining the distribution of material mentions across associated documents. The number of mentions divided by the average mentions across all 134 relevant term clusters was used to account for keywords or phrases that were mentioned at a high frequency. In the photovoltaics case, these normalized number of mentions dropped significantly after the first three term clusters. The first three term clusters were manually reviewed to ensure that they were relevant.

Extracting Statistics

Statistics related to material importance and material–application pairings were extracted from this final set of 134 relevant term clusters. For the purpose of this project, the top 500 abstracts from each of the 134 final clusters were analyzed, for a resulting 49,573 abstracts. Each of these 500 abstracts were automatically searched for mentions of the 16 materials[1] identified in the DOE strategy report, providing a count and distribution of material counts across the final set of granular term clusters that will serve as the basis for a subsequent manual analysis. An alternative to coclustering is hierarchical clustering in which terms are joined with their nearest matches only, and then each cluster is joined with its nearest match only, etc. Such an approach makes subclustering trivial by reducing it to undoing the merges that generated a given cluster. We opted against this structure because in past experience hierarchical clusters have generated qualitatively inferior clusterings of terms.

Figure 6 displays the count of how many times each material was mentioned in the 49,573 abstracts that had the highest mutual information with the final 134 relevant term clusters (the 500 abstracts with the most mutual information with each of the 134 relevant term clusters). Material mentions were counted in the papers most closely associated with each term. This revealed what materials were most strongly associated with which terms. Thus, the methodology provides a way to analyze the distribution of materials over different topics—setting the foundation for material importance and material–application pair analysis.

Validation of Method

To compare the results with the CMS, we need to weight for clean energy, as described above. Multiple clean energy cutoffs were reviewed with and ultimately a cutoff of 26 was employed. The cutoff of 26 corresponds to the beginning of the step up in the distribution presented in **Figure 7**. Thus, term clusters with a clean energy weight of less than 26 were excluded.

Material mentions over the top 500 abstracts associated with the 19 "clean energy" term clusters were aggregated together, yielding a measure of overall importance to clean energy, as seen in **Figure 8**. This measure of overall importance counted the number of times each material was mentioned in the top 500 abstracts associated with the 19 term clusters that are highly important to clean energy, treating those term clusters as one set. The top and the bottom of our determination of importance matched the top and bottom of the DOE's list. However, there was some variation in the middle. Such variation, in the case of real-world analysis, would provide analyst prompts regarding where to consider more study, and whether some aspects of a materials clean energy use and importance.

Material–Application Pairings

SRI utilized the term and document clusters to develop a matrix of material–application pairs to compare the ground truth. This was compared to the DOE 2011 CMS matrix that mapped materials to specific clean energy technologies (see **Figure 9**). The prevalence of different materials to a given clean energy field was considered, as defined and described above.

Figure 10 displays three term clusters, as illustration of the results and filtering step. The term cluster labeled 388 discussed the economic implications of photovoltaic technologies, and so was discarded after manual review. The other two clusters can

[1] These counts included any compounds or alloys associated with those minerals.

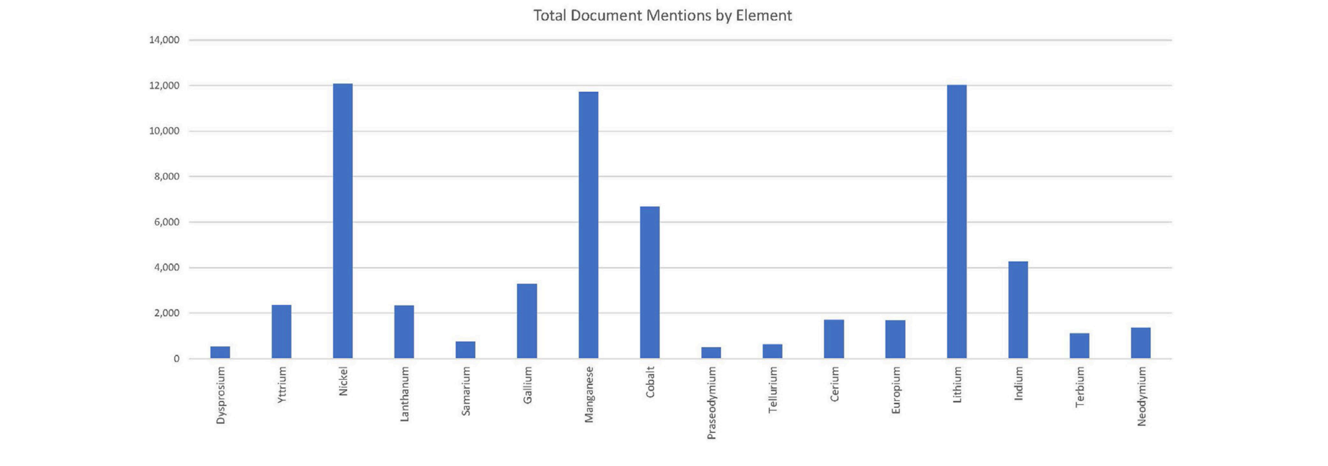


FIGURE 6 | The number of documents in the 134 relevant clusters that contained a specific element varied from under 1,000 to over 12,000, as shown above.

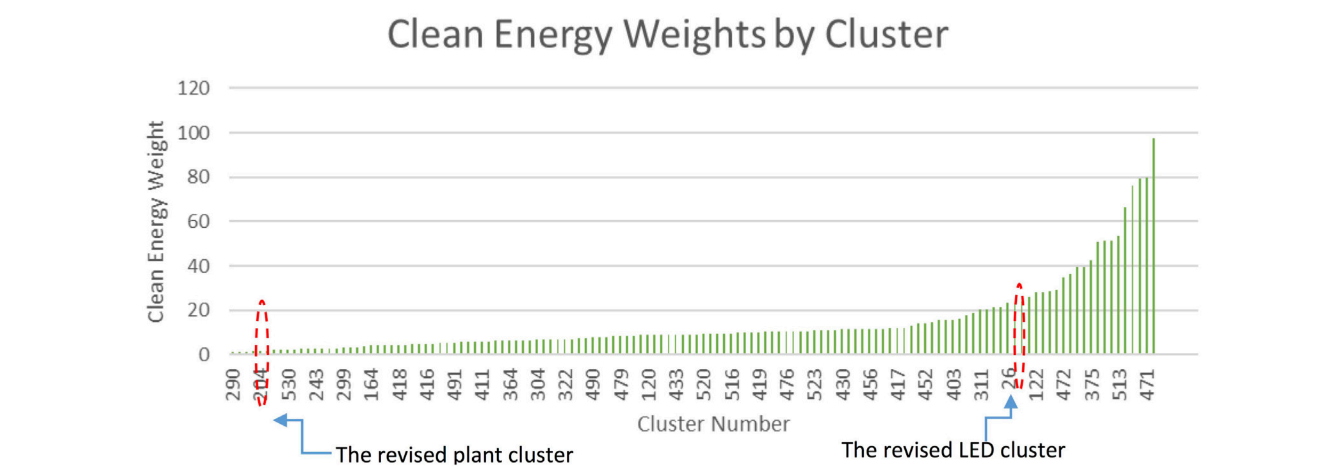
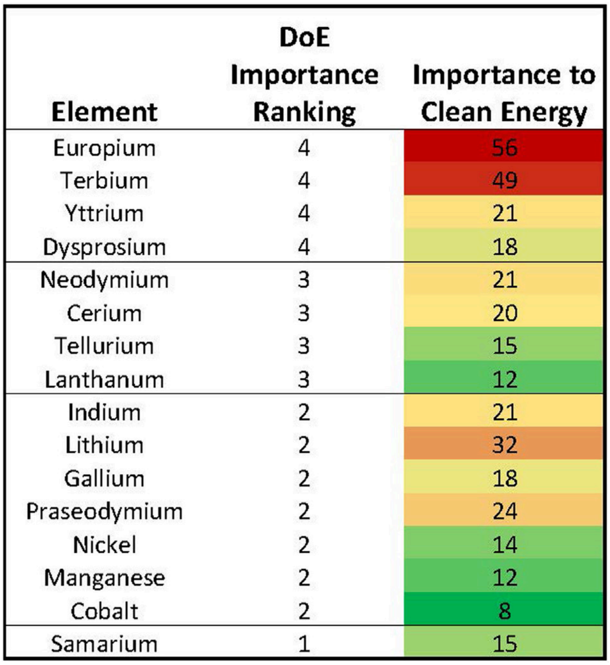


FIGURE 7 | To determine the importance rankings for clean energy, clusters were weighted using a weight reflecting their relevance to clean energy, as shown here. Weights were determined using a "clean energy" glossary, manually created by the analysts. When applied, the revised plant "bulb" cluster has a low clean energy weight, while the revised LED "bulb" has a high clean energy weight.



Element	DoE Importance Ranking	Importance to Clean Energy
Europium	4	56
Terbium	4	49
Yttrium	4	21
Dysprosium	4	18
Neodymium	3	21
Cerium	3	20
Tellurium	3	15
Lanthanum	3	12
Indium	2	21
Lithium	2	32
Gallium	2	18
Praseodymium	2	24
Nickel	2	14
Manganese	2	12
Cobalt	2	8
Samarium	1	15

FIGURE 8 | The importance rankings as determined by SRI methodology. Red denotes more important, while green denotes less important. Results from the clustering algorithms matched the expert-produce high and low importance rankings.

be used to measure the distribution of material mentions across the associated abstracts of the top term clusters, an indicator of material importance to that specific application.

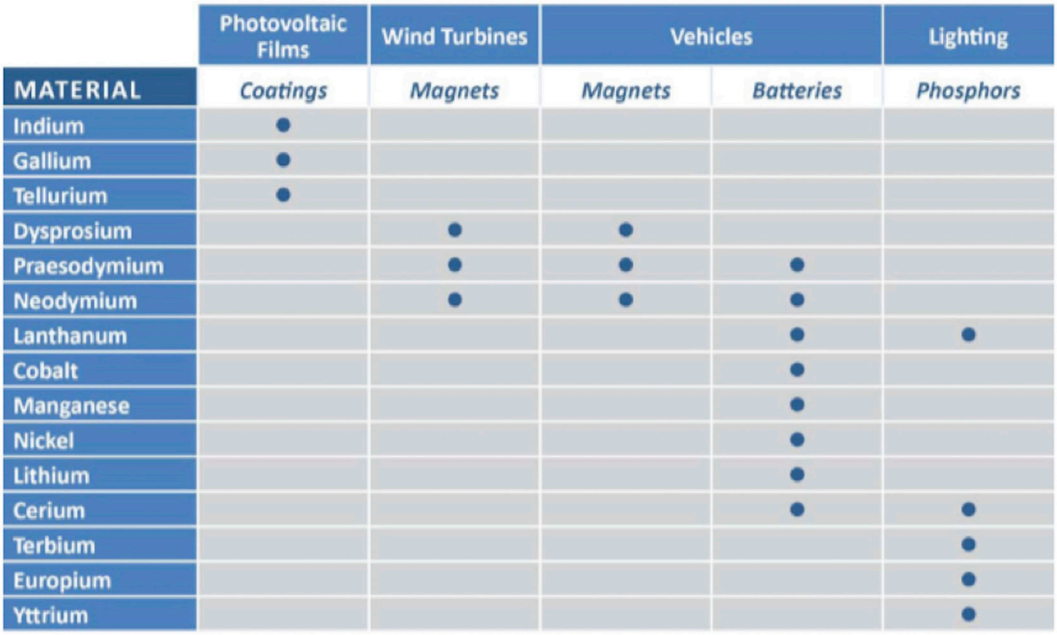
Each material was scored based on a normalized number of mentions across abstracts associated with each of the term clusters of analysis. In this case, the results from our methodology mirror the results from the DOE's report exactly: indium, gallium, and tellurium are considered the most important materials to photovoltaics (see **Table 1**). Similarly, the results for magnets mirrored the DOE's results (see **Table 2**).

Topic Replacement

One potential application of text analytics is the ability to examine trends in material use both within and between term and document clusters to measure how technology may be changing or "trending" within specific application areas. Work on detection of technology emergence has focused on keyword occurrence, sometimes within in the context of an existing taxonomy (Eusebi and Silberglitt, 2014). SRI's method produces both the terms and applications from the corpus. Manual evaluation determined that the term cluster labeled 361 mentioned nickel–metal hydride and other battery technologies, specifically in the context of electric vehicles. In 2008, document counts for lithium surpassed document counts for nickel in term cluster 361 (see **Figure 11**). The graph shown that around 2008, nickel–metal hydride battery technology for electric vehicles had reached a point of relative maturity. Lithium-ion batteries for use in electric vehicles, however, were a less mature technology, and research shifted to focus on advancing this immature technology. While we cannot draw any conclusions about the actual usage of these technologies from these data, the case study demonstrates the potential of text analytics to be used to analyze trends over time and identify how materials and technologies may replace one another. We have also seen this in dye-sensitized solar cells in a previous project (Randazzese, 2016).

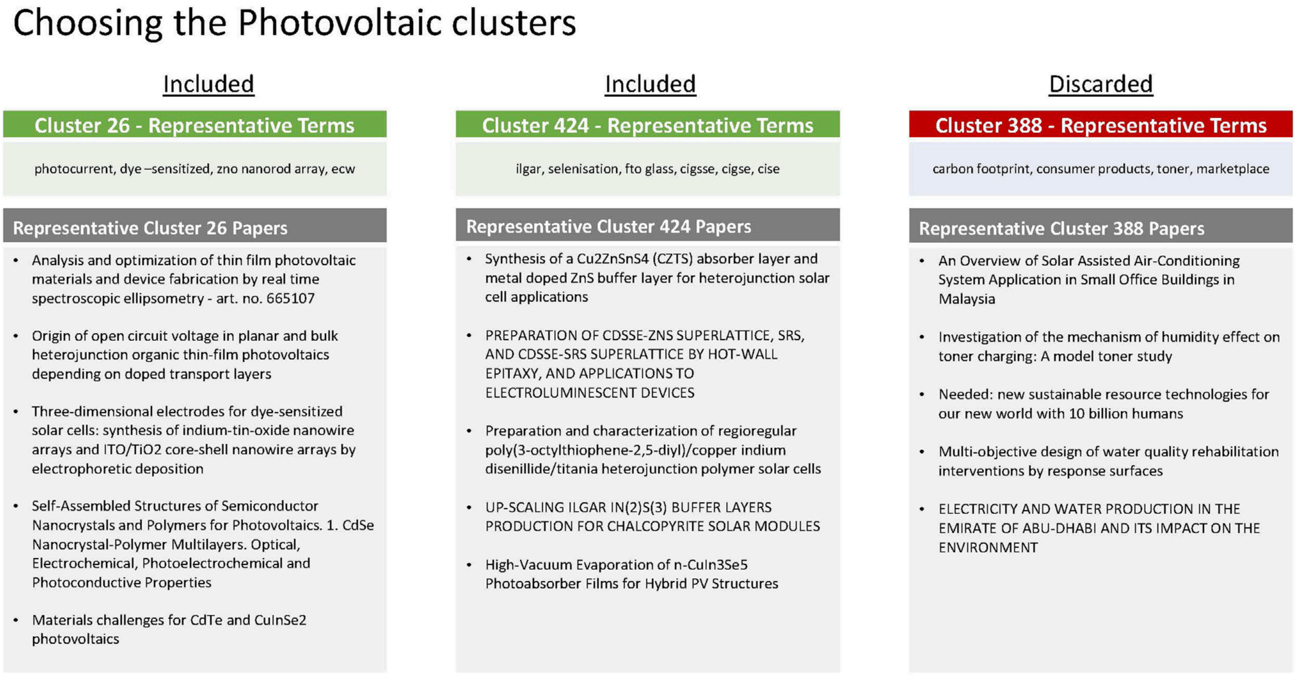
DISCUSSION

The results presented show the application of a text analytics method to extract meaningful information for use to evaluate the progress of research and development. This tool automates away a large amount of manual labor; however, human intervention is still necessary. Human experts are required to define the parameters of evaluation. Fundamentally, this methodology is not designed to replace human analysis or input nor is it intended



	Photovoltaic Films	Wind Turbines	Vehicles		Lighting
MATERIAL	Coatings	Magnets	Magnets	Batteries	Phosphors
Indium	●				
Gallium	●				
Tellurium	●				
Dysprosium		●	●		
Praesodymium		●	●	●	
Neodymium		●	●	●	
Lanthanum				●	●
Cobalt				●	
Manganese				●	
Nickel				●	
Lithium				●	
Cerium				●	●
Terbium					●
Europium					●
Yttrium					●

FIGURE 9 | Material–application pairing matrix from the U.S. Department of Energy Critical Materials Strategy Report, December 2011. These classifications were used to validate the results from the clustering algorithms.



Choosing the Photovoltaic clusters

Included

Cluster 26 - Representative Terms

photocurrent, dye –sensitized, zno nanorod array, ecw

Representative Cluster 26 Papers

- Analysis and optimization of thin film photovoltaic materials and device fabrication by real time spectroscopic ellipsometry - art. no. 665107
- Origin of open circuit voltage in planar and bulk heterojunction organic thin-film photovoltaics depending on doped transport layers
- Three-dimensional electrodes for dye-sensitized solar cells: synthesis of indium-tin-oxide nanowire arrays and ITO/TiO2 core-shell nanowire arrays by electrophoretic deposition
- Self-Assembled Structures of Semiconductor Nanocrystals and Polymers for Photovoltaics. 1. CdSe Nanocrystal-Polymer Multilayers. Optical, Electrochemical, Photoelectrochemical and Photoconductive Properties
- Materials challenges for CdTe and CuInSe2 photovoltaics

Included

Cluster 424 - Representative Terms

ilgar, selenisation, fto glass, cigsse, cigse, cise

Representative Cluster 424 Papers

- Synthesis of a Cu2ZnSnS4 (CZTS) absorber layer and metal doped ZnS buffer layer for heterojunction solar cell applications
- PREPARATION OF CDSSE-ZNS SUPERLATTICE, SRS, AND CDSSE-SRS SUPERLATTICE BY HOT-WALL EPITAXY, AND APPLICATIONS TO ELECTROLUMINESCENT DEVICES
- Preparation and characterization of regioregular poly(3-octylthiophene-2,5-diyl)/copper indium disenillide/titania heterojunction polymer solar cells
- UP-SCALING ILGAR IN(2)S(3) BUFFER LAYERS PRODUCTION FOR CHALCOPYRITE SOLAR MODULES
- High-Vacuum Evaporation of n-CuIn3Se5 Photoabsorber Films for Hybrid PV Structures

Discarded

Cluster 388 - Representative Terms

carbon footprint, consumer products, toner, marketplace

Representative Cluster 388 Papers

- An Overview of Solar Assisted Air-Conditioning System Application in Small Office Buildings in Malaysia
- Investigation of the mechanism of humidity effect on toner charging: A model toner study
- Needed: new sustainable resource technologies for our new world with 10 billion humans
- Multi-objective design of water quality rehabilitation interventions by response surfaces
- ELECTRICITY AND WATER PRODUCTION IN THE EMIRATE OF ABU-DHABI AND ITS IMPACT ON THE ENVIRONMENT

FIGURE 10 | Three term clusters as an illustration of results and filtering step. Analysis of the terms and associated abstracts from the “top” three photovoltaic term clusters reveal that two are relevant while one is not.

to act independently. Instead, it is a tool that researchers can use to more effectively and efficiently analyze their entire domain of research, while also reaching into tangential domains that contain relevant concepts, components, and ideas. Researchers can utilize this methodology to perform objective, replicable, and adaptable reviews of the relative importance of individual components to work toward an understanding of how different pieces fit together. This methodology can be applied to help researchers

TABLE 1 | Materials/application pairing matrix comparing the results from our methodology with the Department of Energy (DOE)'s CM strategy report for the photovoltaic technology and coatings component.

	SRI clusters: 26, 424, 476 **Terms: photovoltaics, photocurent, tandem solar cells, cdte solar cells; CISe, CIGSe, CuInS; charge compensation, doping** **Cluster-field importance: 0.993**	**2011 DOE CM report** **Photovoltaic films**
Material		**Coatings**
Indium	4.58	●
Gallium	2.51	●
Tellurium	1.98	●
Dysprosium	0.33	
Praseodymium	0.80	
Neodymium	0.59	
Lanthanum	0.78	
Cobalt	0.59	
Manganese	0.69	
Nickel	0.51	
Lithium	0.55	
Cerium	0.79	
Terbium	0.44	
Europium	0.68	
Yttrium	0.72	

The numerical scores for each material represent normalized number of document mentions. SRI results matched expert results (right column).

TABLE 2 | Materials/application pairing matrix comparing the results from our methodology with the Department of Energy (DOE)'s CM strategy report for wind turbines and vehicle technologies, and magnet component.

	SRI clusters: 501 **Terms: effective magnetic moment, magnet, paramagnetism** **Cluster-field importance: 0.992**	**2011 DOE CM report** **Wind turbines|vehicles**
Material		**Magnets**
Indium	0.31	
Gallium	0.28	
Tellurium	0.00	
Dysprosium	5.89	
Praseodymium	2.13	●
Neodymium	4.54	
Lanthanum	0.80	●
Cobalt	1.85	
Manganese	0.91	●
Nickel	0.68	
Lithium	0.45	
Cerium	1.10	
Terbium	1.57	
Europium	0.95	
Yttrium	0.90	

The numerical scores for each material represent normalized number of document mentions. SRI results matched expert results (right column).

and inventors better understand how specific components or materials are involved in a given technology or research stream, thereby increasing their potential to create new inventions or discover new scientific findings.

The specific manual construction of the glossary is a crucial step in this methodology. The choice of screening terms will have an obvious impact on what clusters are chosen for analysis. The glossary must be assembled with the end goal in mind. Alternatively, one could use a more limited glossary if the domain of interest was narrow and known. Instead of screening in and out various term clusters, one could keep all clusters and use weightings to rate relevance of clusters. In this case, we created our glossary by selecting keywords from a broad list of material applications (both clean energy relevant and not) to ensure that we did not artificially restrict our results early in the process. This selection was done purely manually based on our analysts' background knowledge of applications of critical materials. This screening identified 42 application term clusters of interest, about 10% of all term clusters.

Future Work

Obvious extensions to this work include text analysis of additional document corpora, more analysis of trends over time, and more sophisticated use of text analytics; for example, to include natural language processing approaches. We did preliminary clustering of patent data that suggested a path forward similar to what we did for papers. New types of insight, for example on the influence of external conditions on materials research, might be obtained from looking at non-technical corpora such as news articles.

The workflow developed for this project allows a subject matter analyst to leverage state-of-the-art text analytic tools without requiring those tools to produce perfect output (which is significant, because the state of the art in text analysis produces enormous amounts of noise when applied in any practical setting). We are able to reduce the overall manual effort required to understand the content of large volumes of scientific reporting, at the cost of shifting some of that effort to tasks dealing with the text analytic tools. For a targeted investigation such as ours, analysts will always select the target and shape the investigation.

We used term clustering primarily as a way for analysts to collect "concepts" from text, but of course there are many ways to carve up the concept space. In our workflow, analysts were able to specify sets of terms that should be further specified, but an improved interface to these tools would allow a user to suggest a split and see resulting changes immediately as opposed to the semi-manual process we currently have in which the data analyst invokes a standalone process to subdivide individual clusters. This was applied, for example, to distinguish the two different senses of the term "bulb," for lighting and as a type of plant. Polysemy of this type is common in natural language and experimental techniques existing for automatically identifying and resolving this polysemy (Freitag, 2004a,b; Freitag et al., 2005). Incorporating these techniques into our workflow and extending them where possible would reduce the effort by the analyst by providing superior semantic distinctions with each round of clustering. It is possible that the alternative clustering technique LDA, referred

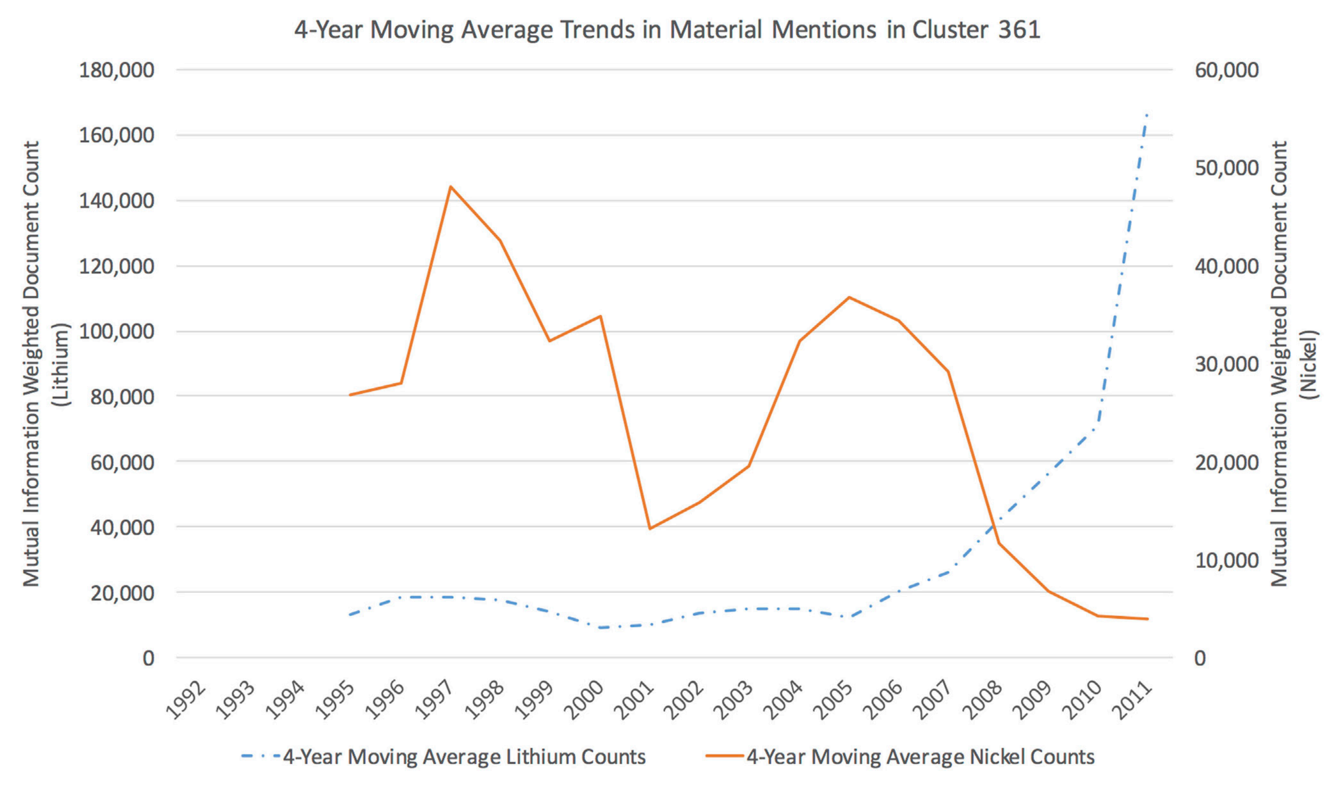


FIGURE 11 | 4-year moving averages for lithium and nickel document counts in term cluster 361. These results from SRI algorithms show decreasing activity in battery research using nickel hydrides by 2008 while lithium research increased for battery applications.

to in the background section, would contribute to this problem in complementary ways to the coclustering approach, as it allows for terms to be members of multiple topics rather than requiring each term to be in a single cluster. In addition, metrics such as entropy or information gain can be used to attempt to automatically recognize clusters most likely to need to be split, although we expect that such metrics will not be sufficient to replace human judgment.

Set expansion techniques (Xu and Croft, 1996; Wang and Cohen, 2009) have been developed for finding a set of terms which are related to a given set of terms in the same way that those terms are related to each other. For example, the word "Achilles" is related to many words in Greek mythology, the word "Alabama" is related to US state names and famous Alabamans, and the word "Queen Mary" is related to historical political figures. When taken together, however, the main thing that these terms have in common is that they are all ship names, and set expansion techniques find such hidden connections and then find additional terms sharing the relationship. Applying such techniques to keyword lists for target technologies would reduce the burden on the analyst to come up with comprehensive and specific keyword lists for targeting concepts.

AUTHOR CONTRIBUTIONS

NK contributed to the design and performed analysis. JB designed the text analytics workflow and extended the Copernicus tools to apply to this project. LR led the development of the overall analytical approach. DH managed the data, applied Copernicus, and implemented extensions designed by JB. CF contributed to the design of the study and methodology. All authors contributed to the drafting of the manuscript.

FUNDING

This work was supported by the National Renewable Energy Laboratory through contract number AEU-6-62527-01. Any opinions, findings, and conclusions expressed in this material are those of the authors and do not necessarily reflect the views of the National Renewable Energy Laboratory or its staff. The authors acknowledge John Chase for his contributions to the initial proposal and project plan. The authors wish to thank Alberta Carpenter, Joe Cresko, Fletcher Fields, and Rod Eggert for their helpful comments during the project. We thank the reviewers for their helpful comments.

REFERENCES

Arthur, W. B. (2009). *The Nature of Technology: What It Is and How It Evolves*. Free Press.

Bellemare, M. G., Naddaf, Y., Veness, J., and Bowling, M. (2015). "The arcade learning environment: an evaluation platform for general agents," in *Proceedings of the 24th International Conference on Artificial Intelligence* (Buenos Aires, Argentina: AAAI Press), 4148–4152.

Blei, D. M., Ng, A. Y., and Jordan, M. I. (2003). Latent Dirichlet allocation. *J. Mach. Learn. Res.* 3, 993–1022. doi:10.1162/jmlr.2003.3.4-5.993

Byrnes, J., and Rohwer, R. (2005). "Text modeling for real-time document categorization," in *2005 IEEE Aerospace Conference*, Big Sky, MT

Caid, W. R., Dumais, S. T., and Gallant, S. I. (1995). Learned vector-space models for document retrieval. *Info. Process. Manag.* 31, 419–429. doi:10.1016/0306-4573(94)00056-9

Erdmann, L., and Graedel, T. E. (2011). Criticality of non-fuel minerals: a review of major approaches and analyses. *Environ. Sci. Technol.* 45, 7620–7630. doi:10.1021/es200563g

European Commission. (2014). *Report on Critical Raw Materials for the EU: Report of the Ad Hoc Working Group on Defining Critical Raw Materials*.

European Economic and Social Committee. (2011). *Commodity Markets and Raw Materials E. Commission*. Brussels: European Commission.

Eusebi, C. A., and Silberglitt, R. (2014). *Identification and Analysis of Technology Emergence Using Patent Classification*. RAND Corporation.

Freitag, D. (2004a). "Toward unsupervised whole-corpus tagging," in *Proceedings of the 20th International conference on Computational Linguistics* (Geneva, Switzerland: Association for Computational Linguistics), 357.

Freitag, D. (2004b). "Trained named entity recognition using distributional clusters," in *Proceedings of EMNLP 2004*, eds D. Lin and D. Wu (Barcelona, Spain: Association for Computational Linguistics), 262–269.

Freitag, D., Blume, M., Byrnes, J., Chow, E., Kapadia, S., Rohwer, R., et al. (2005). "New experiments in distributional representations of synonymy," in *Proceedings of the Ninth Conference on Computational Natural Language Learning* (Ann Arbor, MI: Association for Computational Linguistics), 25–32.

Furnas, G. W., Landauer, T. K., Gomez, L. M., and Dumais, S. T. (1987). The vocabulary problem in human system communication. *Commun. ACM* 30, 964–971. doi:10.1145/32206.32212

Gallant, S. I., Caid, W. R., Carleton, J., Hecht-Nielsen, R., Qing, K. P., and Sudbeck, D. (1992). HNC's MatchPlus system. *SIGIR Forum* 26, 34–38. doi:10.1145/146565.146569

Graedel, T. E., Barr, R., Chandler, C., Chase, T., Choi, J., Christoffersen, L., et al. (2012). Methodology of metal criticality determination. *Environ. Sci. Technol.* 46, 1063–1070. doi:10.1021/es203534z

Graedel, T. E., Harper, E. M., Nassar, N. T., Nuss, P., and Reck, B. K. (2015). Criticality of metals and metalloids. *Proc. Natl. Acad. Sci. U.S.A.* 112, 4257–4262. doi:10.1073/pnas.1500415112

Hastie, T., Tibshirani, R., and Friedman, J. H. (2004). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction: With 200 Full-Color Illustrations*. New York: Springer.

Hofmann, T. (1999). "Probabilistic latent semantic indexing," in *Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (Berkeley, CA: ACM), 50–57.

Intelligence Advanced Research Projects Activity (IARPA). (2011). *IARPA Launches New Program to Enable the Rapid Discovery of Emerging Technical Capabilities*. Available at: https://www.dni.gov/index.php/newsroom/press-releases/press-releases-2011/item/327-iarpa-launches-new-program-to-enable-the-rapid-discovery-of-emerging-technical-capabilities

Manning, C. D., and Schutze, H. (1999). *Foundations of Statistical Natural Language Processing*. MIT Press.

Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013). Efficient estimation of word representations in vector space. arXiv:1301.3781.

National Research Council (US). (2008). *Minerals, Critical Minerals, and the U.S. Economy*. Washington, DC: National Academies Press.

National Research Council (US) and Committee on Assessing the Need for a Defense Stockpile. (2008). *Managing Materials for a Twenty-First Century Military*. Washington, DC: National Academies Press.

National Science Board. (2016). *Science & Engineering Indicators 2016*. Washington, DC: National Science Board.

Panousi, S., Harper, E. M., Nuss, P., Eckelman, M. J., Hakimian, A., and Graedel, T. E. (2016). Criticality of seven specialty metals. *J. Ind. Ecol.* 20, 837–853. doi:10.1111/jiec.12295

Poulton, M. M., Jagers, S. C., Linde, S., Van Zyl, D., Danielson, L. J., and Matti, S. (2013). State of the world's nonfuel mineral resources: supply, demand, and socio-institutional fundamentals. *Annu. Rev. Environ. Res.* 38, 345–371. doi:10.1146/annurev-environ-022310-094734

Randazzese, L. (2016). *Helios: Understanding Solar Evolution through Text Analytics*. Menlo Park, CA: SRI International.

US Department of Energy. (2011). *Critical Materials Strategy*. Available at: https://energy.gov/sites/prod/files/DOE_CMS2011_FINAL_Full.pdf

Wang, R. C., and Cohen, W. W. (2009). "Automatic set instance extraction using the web," in *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 1* (Suntec, Singapore: Association for Computational Linguistics), 441–449.

Xu, J., and Croft, W. B. (1996). "Query expansion using local and global document analysis," in *Proceedings of the 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (Zurich, Switzerland: ACM), 4–11.

Zhu, H., and Rohwer, R. (1995). Bayesian invariant measurements of generalization. *Neural Process. Lett.* 2, 28–31. doi:10.1007/BF02309013

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.