Clustering scientific documents with topic modeling

Article *in* Scientometrics · September 2014

DOI: 10.1007/s11192-014-1321-8

CITATIONS
57

READS
1,348

4 authors, including:

Alan L. Porter
Georgia Institute of Technology
409 PUBLICATIONS **7,766** CITATIONS

SEE PROFILE

Nils C. Newman
Maastricht University
44 PUBLICATIONS **632** CITATIONS

SEE PROFILE

Arho Suominen
VTT Technical Research Centre of Finland
104 PUBLICATIONS **498** CITATIONS

SEE PROFILE

DRAFT - not final – for the published article please go to *Scientometrics*

Yau, C-K, Porter, A.L., Newman, N.C., and Suominen, A. (2014), Clustering scientific documents with topic modeling, Scientometrics, GTM special issue; 100 (3) 767-786; http://link.springer.com/content/pdf/10.1007%2Fs11192-014-1321-8.pdf.

Clustering scientific documents with topic modeling

Chyi-Kwei Yau

Alan Porter

Nils Newman

Arho Suominen

Received: date / Accepted: date

Abstract

Topic modeling is a type of statistical model for discovering the latent "topics" that occur in a collection of documents through machine learning. Currently, Latent Dirichlet Allocation (LDA) is the most popular and common modeling approach. In this paper, we investigate methods, including LDA and its extensions, for separating a set of documents into several clusters. To evaluate the results, we generate a collection of documents that contain academic papers from several different fields and see whether papers in the same field will be clustered together. We explore potential scientometric applications of such text analysis capabilities.

Key words: topic modeling, text analysis, ??

1 Introduction

With the increasing use of structured databases, our approaches toward measuring science are relying more and more on quantitative methods. Often referred to as bibliometrics (Borgman and Furner, 2002), studies have used different methodological tools to extract information from databases, striving to uncover underlying structures within the dataset (Daim et al, 2006). Moving from and adding to more elementary measures (Suominen, forthcoming), studies have used different methods to quantify not just the counts of countries, authors or technologies mentioned in a given year, but to augment those variables with textual content variables as well. The possibilities of text mining in scientometrics were shown in the pilot study by Glenisson (2005), which pointed out that text mining methods show promise as a valuable tool in mapping fields of science.

The text mining approach seeks to identify words or phrases that could explain possible underlying content and structures (relationships) in the data. Indentifying commonalities within the text, the analytical options with text mining have focused on analyzing co-occurrence data by distribution analysis, association rules, or different clustering approaches (Feldman and Sanger, 2006). Methods that could create practical categories from the text, rather than using preordained categories, enable an abundance of new research options. The work by Hofmann (1999) suggested that topic modeling would be a useful tool for extracting information from textual data. A year later, Wei and Croft (2006) showed that in information retrieval, topic models outperform more traditional, cluster-based approaches. Motivated by the possibilities of topic modeling, we approach scientific publication data with topic modeling and show a practical example that evaluates topic modeling ability to distinguish among scientific documents. Although a body of literature already exists on topic modeling, this literature mainly focuses on the methodology and its development. In analyzing and mapping science, we are more focused on the method's ability to distinguish the underlying structures within the text, thus forming a practical tool for the scientometrics community.

In this paper, we consider the problem of separating a scientific document collection into several clusters based on their content by topic modeling. To this end, we have created sample data with an underlying structure built in by the researcher. The objective is to use topic modeling as a relatively automated method that can generate document clusters containing similar documents and then compare the results to the actual structure of the data. By doing so, we evaluate the ability of the topic modeling method to distinguish structures from the text and measure its success by precision and recall.

To elaborate further, when referring to topic modeling, we actually refer to several different algorithms. In the field of machine learning, the straightforward approach would have been to separate a data collection by using classification algorithms. However, since labeling a set of documents would be time-consuming and inefficient, we prefer not to solve this problem with classification methods or other supervised learning algorithms. Instead, topic models are used to address this problem. Information retrieval researchers proposed the first probabilistic model, "probabilistic Latent Semantic Indexing" (pLSI) (Hofmann, 1999). Probabilistic Latent Semantic Indexing models each word in a document as a sample from a mixture model, where the mixture components are multinomial random variables that can be viewed as representations of topics. The problem in the pLSI algorithm is that the number of parameters in a pLSI model will grow linearly with the topic number, which means it tends toward over-fitting. To solve the problem, an improved model, "Latent Dirichlet Allocation" (LDA) was proposed (Blei et. al, 2003). LDA is a three-layer Bayesian model that overcomes the disadvantages of pLSI and is now widely used in different applications, such as text mining, bioinformatics and image processing. In addition, to further improve upon some constraints in LDA, various extensions, such as Correlated Topic Models (CTM), Hierarchical Latent Dirichlet Allocation (Hierarchical LDA) and the Hierarchical Dirichlet Process (HDP) have been proposed in recent years.

Topic models have also been extended to take into account time series (Blei and Lafferty, 2006b) and the scientific communities and topics in textual data (Yan, 2012; Steyvers, 2004), which will enable practical methods for analyzing structures of scientific communities and topics studied in the future. This can be seen for example in Ni et. al (2012) for example, where an author-conference-topic (ATC) extension was

applied. However, all of these rely on the ability of the topic model algorithm to distinguish between different topics.

Recently we have seen that topic modeling is becoming more approachable as the availability of applications enables researchers to take advantage of these algorithms. MALLET (McCallum, 2002), a Java-based package for Topic Modeling, and topicmodels (Grün and Hornik, 2011), an R package for fitting topic models, both give researchers a practical approach to using different topic modeling algorithms. Accessible software such as these two packages facilitates the use of topic modeling in scientometrics research. Studies focusing on the applicability and limitations of the approach and different algorithms are valuable in better understanding what is achieved by topic modeling. In this paper, we will first describe the data in Section 2, give a short review of LDA and its extensions in Section 3, and show how we use those methods to separate documents in Section 4. Finally, we present the practical implications of the results in Section 5.

2 Research design, data and pre-processing

This research is designed to elaborate on the ability of different topic modeling algorithms—namely LDA, CTM, Hierarchical LDA, and HDP—to distinguish and automatically cluster documents. To this end, we design a two-phased research process: first we focus on testing and then on verification. In the first phase, flow-charted in Figure 1, we test different pre-processing methods and topic modeling algorithms with a test sample. The results are used in the verification phase with a more correlated dataset.

In the first phase, we evaluate the different algorithms , excluding Hierarchical LDA, where we show the path clustering results by precision, recall and subsequent F-score. Both precision and recall are measures of relevance. Precision is defined as the number of relevant documents retrieved by the algorithm divided by the total number of documents retrieved. Recall, on the other hand, is defined as the number of relevant documents retrieved divided by the total count of relevant documents. The F-score is also counted as a measure of accuracy -- the weighted average of precision and recall -- with a maximum value of one and minimum value of zero. The equation used to calculate the F-score is as follows:

$$F = 2 \times \frac{precision \times recall}{precision + recall} \tag{1}$$

We first calculate precision, recall and F-score for the test sample described in Subsection 2.1 with each of the three algorithms described in Section 3 and then use the highest-performing algorithm to validate the result with the sample described in Subsection 2.2. For the verification sample, we use the highest-performing algorithm to create a set of top topics, give them labels, and show the five top words.

2.1 Test sample

We select seven different research fields for which we download bibliographical data from the Web of Science and merge together to create one set. After a pre-processing phase, we use four different topic modeling algorithms to divide the articles back to their original sets and compare the effect of two different preprocessing approaches and the ability of each algorithm.

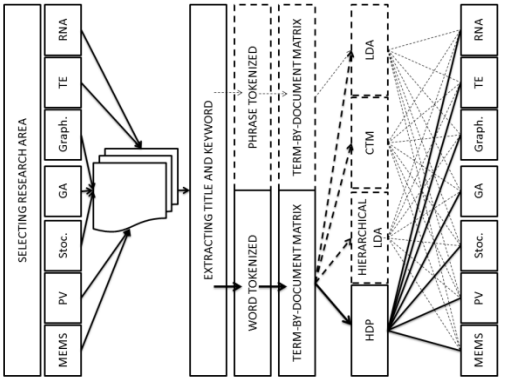

Figure 1: Overall illustration of the research design.

To elaborate in detail, we generate a document collection from seven different scientific areas, as listed in Table 1. This collection serves as our test sample. We choose these seven scientific areas because some of the areas are related with others and some are not. For example, Genetic Algorithm is correlated with stochastic programming; RNAi is correlated with Tissue Engineering; and Graphene is almost independent of all the others. Therefore, with this sample, we can measure the clustering performance under different conditions of relatedness. The sample is also limited to one of the authors being affiliated with the Georgia Institute of Technology, or Emory University in the case of RNAi (to increase the sample size). This gives the authors an opportunity to validate the sample. In addition, the sample provides a convenient size to run a variety of models. Our goal is to separate them with high accuracy and precision, without significant manual effort.

Table 1: Diverse Test Set from Web of Science

Category(Keyword)	**Document #**
MEMS	345
Solar cell or photovoltaic	178
Tissue Engineering	217
RNAi or RNA inference	127
Graphene	180
Genetic Algorithm	114
(stochastic or non-linear) programming	99

The data are gathered from the Web of Science database using simple search algorithms – i.e., one of seven different keywords representing the technologies appearing in the "topics" field, together with variants of the universities' name in the author affiliations field. In total, we retrieve bibliographical information in the form of 1,254 abstract records from Web of Science. Those are then mixed to form the corpus used as a single file in VantagePoint software [www.theVantagePoint.com]. The keywords and retrieved document counts are listed in Table 1.

The data are pre-processed prior to the actual topic modeling. First, we clean the data to only use the combined title and abstract information for each document. Then

we use two different processes to clean the data: 1) approaching the data in the form of "tokens" consisting of individualwords, or 2) tokens of noun phrases (extracted using VantagePoint's Natural Language Processing (NLP). At first we treat each word as a token and then remove words in a stopword list or words with very low frequency. With the second method we use term-clumping (Newman et. al, 2012), which aggregates noun phrases with significant commonality, and documents are tokenized as these consolidated phrases. The advantage of this preprocessing method is that it can generate more meaningful terms. For example, solar cell should not be separated into single words -- "solar" and "cell" -- but should be used as "solar cell." The results from both processes are analyzed separately with topic modeling algorithms.

No further preprocessing is used for the dataset. Since topic models represent an unsupervised learning algorithm, we do not have to provide the true category while training the model. Instead, after the topic model generates document clusters, we use the true categories to assess each cluster.

2.2 Verification sample

Based on the results of the first phase of the research, we create a verification sample to test the pre-processing method and topic modeling algorithm that performs best during the testing phase. In comparison to the more unnatural case of the test sample, where we deliberately select quite diverse fields that could show the potential of topic modeling, the verification sample is designed so that the authors will have a somewhat more focused, coherent science field to analyze.

For this purpose, we gather a dataset on energy technologies from EI Compendex. Gathered according to a similar process as the test sample, we download bibliographical records for seven energy-related science fields seen in Table 2.

Table 2: Energy-related Verification Set from EI Compendex

Category(Keyword)	Document #
Steam power	99
Solar	181
Fuel combustion	192
Energy management	96
Petrol	159
Mining	157
Fuel cells	122

For the overall verification set, containg the 1006 documents seen in Table 2, we use the highest-performing algorithm to create the topic distribution. We analyze the results with a qualitative approach, thus keeping the notion of knowing the sample (i.e., that 7 searches are combined therein) out of the analysis process.

3 Methodology

3.1 Assumptions

In this section, we review the models used in our experiment and their inference methods. However, before elaborating on each model, we have to be aware of the basic assumptions behind topic modeling. The assumptions include the following:

1. Documents are exchangeable in a corpus.
2. Words are exchangeable in a document (bag of words assumption, meaning no syntactic or proximity relationship information used).
3. A topic is modeled as a multinomial distribution on words from some basic vocabulary.
4. Words in a document are arising from a number of latent topics.

The first two assumptions ignore document order in a corpus and word order in a document. This assumption is made, although models have been suggested that go beyond the "bag of words" assumption (Wallach, 2006). The third assumption defines the meaning of a topic, which is a distribution throughout a dictionary. For example, high probability words in a sports topic may be baseball, football, athlete, etc. Subsequently, we usually use the top few words to represent a given topic. The last one assumes that each word has one or several corresponding latent topics with which it is associated, and given the particular topic, the word is drawing from that topic's distribution. Next we review LDA and its variations.

3.2 Latent Dirichlet Allocation

Latent Dirichlet Allocation (LDA) is a generative probabilistic model widely used in the information retrieval field. The basic idea behind the model is that each document in a corpus is a random mixture over latent topics, and each latent topic is characterized by a distribution over words. A graph with notation for each element is shown as Figure 2. In Figure 2 we use the following notation for the LDA:

α parameters of topic Dirichlet prior

β parameters of word Dirichlet prior

W=$\mathbf{w_i}$ corpus, each w_i denote a word

Z=$\mathbf{z_i}$ latent topic assigned to words in W

$\Theta = \{\theta_{d,k}\}$: $\theta_{d,k} = P(z = k|d)$, the probability of topic z=k given document d

$\Phi = \{\phi_{k,v}\}$: $\phi_{k,v} = P(w = v|z = k)$, the probability of word w=v given topic z=k

$\Omega_{d,k}$: count of words in d assigned to topic k

$\Psi_{k,v}$: count of word v in corpus assigned to topic k

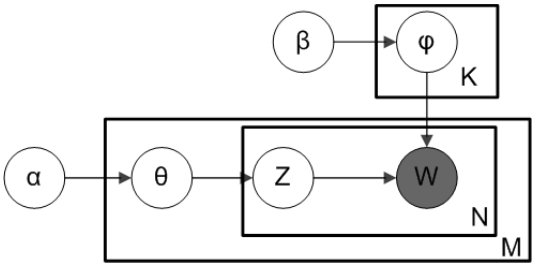

Figure 2: Graphical representation of the Latent Dirichelt Algorithm (LDA). Adopted from Blei and Lafferty (2009).

We can also describe the process of generating a document by the following process:

1. Choose document length $N \sim Poisson(\xi)$.
2. Choose Θ from Dirichlet distribution $\Theta \sim Dir(\alpha)$.
3. For each word in the document, choose w_n from a multinomial distribution $p(w_n|z_n, \beta)$ where z_n represents the topic of this word.

Since we are interested in the topic distribution for each document and the content of each topic, our goal is to find the τ and Φ, which requires inference methods. There are two major options for inference methods: MCMC (collapsed Gibbs sampling) (Griffiths and Steyvers, 2004) and variational EM algorithm (Blei et. al, 2003). In our approach, we use Gibbs sampling to sample the latent variable z. Based on z, we are able to infer the topic distribution for each document and the content of each topic.

Although LDA can provide very good estimation results, there are two important limitations. First, it is hard to measure the topic correlations between each of the topics. Second, LDA requires a fixed number of topics, which is usually unknown to the researcher. In following subsections we, describe three algorithms that address these shortcomings of LDA by modifying the prior distribution.

3.3 Correlated Topic Models

Correlated Topic Models (CTM) address the first limitation in LDA by replacing the Dirichlet distribution of topic proportions with a logistic normal distribution. This is illustrated in Figure 3, where we use the following notation for the CTM:

$\boldsymbol{\mu}$: K vector

$\boldsymbol{\Sigma}$: K × K matrix

Using the CTM approach we first draw topic proportion defined as follows:

$$\eta_d \; N(\eta, \Sigma) \tag{2}$$

followed by the topic assignment:

$$z_{d,n} \sim Multi \frac{exp(\eta)}{\sum_{i=1}^{K} exp(\eta_i)} \tag{3}$$

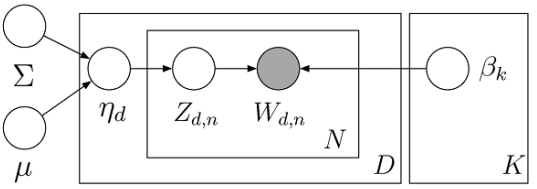

Figure 3: Graphical representation of the CTM algorithm. Adopted from Blei and Lafferty (2006).

Here we use η and Σ to replace the original dirichlet distribution, which allows us to consider the topic correlation by Σ . The disadvantage is that the logistic normal distribution and multinomial distribution are no longer conjugate (Blei and Lafferty, 2006), which is computationally less efficient. Therefore, inference must be done by a variation inference method (Blei and Lafferty, 2007).

3.4 Hierarchical Latent Dirichlet Allocation

Hierarchical LDA is another variation of the original LDA. Instead of choosing a fixed number of topics, this approach uses a nested Chinese restaurant process (nCRP) as the prior distribution to learn topic hierarchies. In this model, each restaurant is a topic, and the topics of a document correspond to a path. For example, in Figure 5, there are four documents (1,..,4) and six topics $(\beta_1,...,\beta_6)$. All documents will share the root topic β_1 and choose their own path based on the nCRP. In detail, as seen in Figure 4, a document is generated as follows:

1. Let c_1 be the root restaurant.
2. For each level $\ell \in \{2,...,L\}$, $c\{\ell-1\}$ using CRP. Set c_ℓ to the restaurant at this level.
3. Draw an L-dimension topic proportion vector τ from *Dir*(α).
4. For each word $n \in \{1,...,N\}$, draw $z \in \{1,...,L\}$ from multinomial distribution *Mult*(τ). Based on z, draw w_n from the topic associated with restaurant c_z.

For inference, we can still use the collapsed Gibbs sampling as in LDA, but now we have to sample both $c_{m,\ell}$ and z.

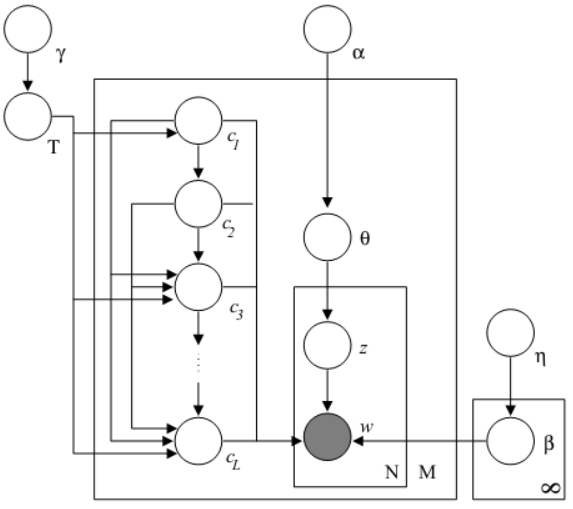

Figure 4: Graphical representation of the Hierarchical LDA algorithm. Adopted from Blei and Lafferty (2009).

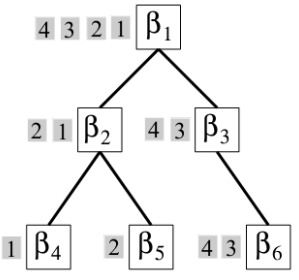

Figure 5: Process used in the Hierarchical LDA as a tree structure.

3.5 Hierarchical Dirichlet Process

The Hierarchical Dirichlet Process (HDP) is a nonparametric Bayesian model that has been widely used to implement admixture models. In its text mining application, HDP is an extension of LDA, which releases the fixed number of topics constrained by using a Dirichlet Process[? ? reference] as its prior distribution. The model is shown in Figure 6. In LDA, the topic distribution τ_m is sampled from a fixed dirichlet distribution $Dir(\alpha)$. In HDP, τ_m is sampled from $DP(\alpha_0 G_0)$, where G_0 is the base distribution. For inference, as with the LDA, collapsed Gibbs sampling can be used to sample as follows: *as a* as follows: as

$$p(z_i = k|\cdot) \propto (n_{m,k}^{-i} + \alpha\tau_k) \cdot \frac{(n_{k,t}^{-i} + \beta)}{(n_k^{-i} + V\beta)} \quad (4)$$

The dimension of τ_k is K+1, where K is the current topic number. When z_i is sampled into the (K+1)™s topic, the model will generate a new topic. Therefore, in HDP, the number of topics is based on the training data and the parameters in the dirichlet process.

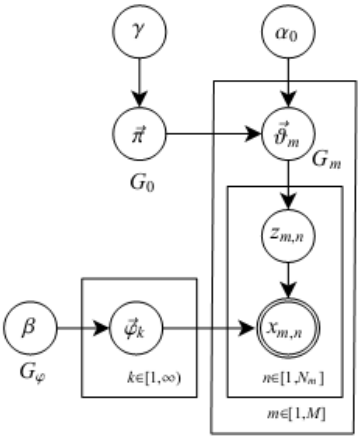

Figure 6: Graphical representation of the HDP process. Adopted from Blei and Lafferty (2009).

3.6 Manually assigning labels to topics

Topic modeling algorithms can be created from the word or phrase tokenized corpus using either a predefined or inferred number of topics. These topics are a collection of either the words or phrases used as tokens. Within a topic, the tokens have a weight, some being more dominant in their representation of the topic than others. Often the result of this is that, in practice, the five (or so) most dominant words enable the researcher to identify the topic's content and manually assign it a label. This is, however, a point in the research where the researcher impacts the results of the study by selecting the topic's name. This may result in challenges if we attempt to reproduce results. It should be noted that the researcher can use the token of phrases or clustering methods to clarify the meaning of an individual topic; however, labeling (topic naming) is done by the researcher. To some extent, different algorithms also give the researcher different possibilities to look at the topics created; thus we have discussed the manual assignment of topic names with the results of each algorithm.

4 Results of test dataset

We convert the test sample into topic distributions with the described topic modeling algorithms. We treat these distributions as different features and use them to generate document clusters. Since the topic distributions are based on an unsupervised learning algorithm, the methods require different levels of human effort to identify the clusters. Ultimately, it is up to the researcher to identify the significant topics from the results. In this study, our ultimate goal was to separate the data into their original seven clusters,

as shown in Table 1, with the least human effort and greatest accuracy and precision possible.

The first noteworthy observation—which is qualitative in nature—is that, in scientific document clustering, topic models showed good estimation results with high frequency terms. Thetokenized set of phrases created during the data preprocessing phase did not produce equally good results as did single word modeling. The method clearly lacked in the ability to model topics using phrases created by word clumping. Subsequently we will only show the results associated with the single word tokenized corpus.

4.1 LDA results

As mentioned before, LDA requires a fixed number of topics, which is usually unknown in a document corpus. In our model, we fix the number of topics at fifty and use different methods to match these fifty topics to the seven original categories. However, we try different numbers of topics, from 50 to 200, in the experiment, as a trial-and-error approach to using LDA. By comparing the results, we choose fifty topics since it is more manageable than the larger numbers of topics.

To illustrate the results from the LDA process, we extract the results for one document in the set. The bibliographical information of the document is seen in Table **Error! Reference source not found.**. Table **Error! Reference source not found.** shows that the four top topics in this document are topics 22, 24, 49 and 14. In order to know the content of each topic, the top five words are shown in Table **Error! Reference source not found.**. From the proportion of topics and top words in each topic, we can understand that the document in Table **Error! Reference source not found.** is a paper in the field of genetic algorithm or stochastic programming.

Table 3: Selected bibliographical information on one document in the test sample.

Sample document	
Authors	Kulankara Krishnakumar, Shreyes N. Melkote
Year	2000
Publication	International Journal of Machine Tools & Manufacture
DOI	10.1016/S0890-6955(99)00072-3
Title	Machining fixture layout optimization using the genetic algorithm
Abstract	Dimensional and form accuracy of a workpiece are influenced by the fixture layout selected for the machining operation. Hence, optimization of fixture layout is a critical aspect of machining fixture design. This paper presents a fixture layout optimization technique that uses the genetic algorithm (GA) to find the fixture layout that minimizes the deformation of the machined surface due to clamping and machining forces over the entire tool path. The advantages of the GA-based method over previously reported non- linear programming methods for fixture layout optimization are discussed. Two GA-based fixture layout optimization approaches are implemented and compared by applying them to several two-dimensional example problems

Table 4: The four most significant topics in the document seen in Table 3.

Topic	Proportion
22	58 %
24	17 %
49	16 %
14	5.4 %

Table 5: The five most significant words in each of the topics seen in Table 4.

Topic 22	Topic 24	Topic 49	Topic 14
Manually Assigned label to each Topic			
Genetic Algorithm	**Stochastic**	**Stochastic**	**Solar cell**
design	model	optimization	films
optimization	results	model	thin
genetic	experimental	approach	tio
algorithm	analysis	process	silicone
method	based	method	surface

We generate clusters by manually connecting topics to the actual research fields. Often the easiest approach to connecting topics to categories is to manually assign LDA topics into specific categories. Given the most common words in each topic, it is not hard to manually identify topics. For example, from the words shown in Table **Error! Reference source not found.**, we can assign topic 22 to genetic algorithm, topics 24 and 49 to stochastic programming, and topic 14 to solar cell. Similarly, we assign each topic a label. Then by aggregating the topic distribution within the same label, we can know how a document correlates with each label. For example, from Table **Error! Reference source not found.** and **Error! Reference source not found.**, we aggregate all topics and calculate that the document seen in **Error! Reference source not found.** contains 58 % genetic algorithm and 33% stochastic programming.

Based on this process, we set a threshold for each document. If the proportion of any given label is greater than 50 %, we assign this document to that label, or, in this case, to that research field. In our previous example, the document seen in **Error! Reference source not found.** will be assigned to the genetic algorithm category. However, if all of the labels in a document have a proportion of less than 50 %, we do not assign any label for the document.

To evaluate the labeling for the test sample, we use both precision and recall to measure the results. Table 6 shows the results of all seven research fields. The Graphene field, or label, provides the best result. This might be due to the fact that it is least correlated with the other research fields.

The problem of this method is that the performance is highly dependent on human judgment, and as the number of topics increases, the required human effort will increase as well. Also, obscure topics are hard to distinguish and may require expert knowledge.

Table 6: The results of the LDA analysis.

Category	actual	estimated	precision	recall	F-score
MEMS	345	270	0,85	0,67	0,748
Tissue Engineering	217	160	0,85	0,63	0,721
Graphene	180	149	0,84	0,69	0,76
Solar Cell	178	154	0,75	0,64	0,692
RNAi	127	116	0,73	0,67	0,699
Genetic Algorithm	114	104	0,65	0,59	0,624
Stochastic Programming	99	104	0,748	0,77	0,748

4.2 CTM results

Similar to LDA, CTM will generate word distribution for each topic and topic distribution for each document. The difference is that we can use lasso regression to find correlations between different topics (Blei and Lafferty, 2007). By connecting those highly correlated topics, we can get a correlation graph as seen in Figure 7. This graph can help us identify which topics should be grouped together. A link between two different topics represents a similarity between them. In our graph, we can find that most links correctly capture the relationship. However, these links are not perfect; thus we cannot use them for clustering. For example, in Figure 7, if we treat all connected topics as a cluster, this will create a wrong impression that a very large cluster contains MEMS, stochastic programming, and the genetic algorithm. With our test sample, this is, of course, not the case, and human effort is needed to analyze the graph and perform subsequent labeling.

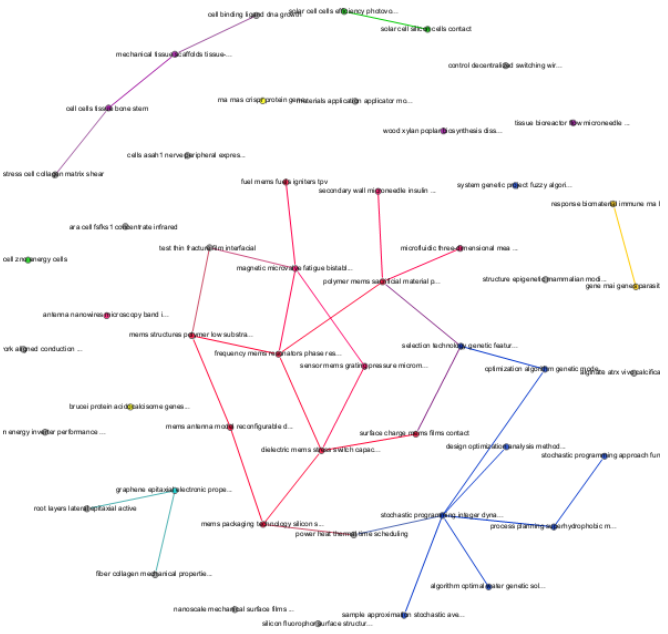

Figure 7: Correlation graph for the CTM algorithm results.

Just as with LDA, we can assign one label to each topic and aggregate topics with the same label. The results are shown in Table 7. The results are very similar to what we obtained from LDA. The Graphene field still achieves the best score while the genetic algorithm and stochastic programming show the worst match with the original data.

Table 7: CTM algorithm results.

Category	Actual	estimated	precision	recall	F-score
MEMS	345	355	0,757	0,779	0,768
Tissue Engineering	217	206	0,703	0,668	0,685
Graphene	180	183	0,863	0,877	0,87
Solar Cell	178	191	0,7	0,75	0,726
RNAi	127	104	0,75	0,614	0,675
Genetic Algorithm	14	132	0,613	0,71	0,658
Stochastic Programming	99	77	0,753	0,585	0,659

4.3 Hierarchical LDA with document path

In Hierarchical LDA, we know that each document will follow only one path in the hierarchical topic structure, and documents with similar word distribution should follow the same path. Therefore, we build a three-layer Hierarchical LDA, and check whether documents in the same path belong to the same research field. Figure 8 shows the hierarchical topic structure generated during this process.

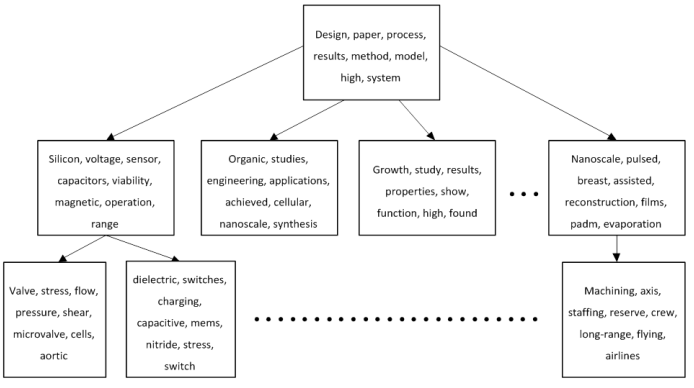

Figure 8: Three-layer hierarchical topic structure.

As seen from the results, the top words in the root node are common words such as "design," "paper" and "process." In the second layer, eight different and less general topics appear. In the third layer, we have 46 different topics. In other words, in this topic structure, we have 46 different paths for documents. To keep the table at a practical size, we show the paths that contain more than fifty documents in Table 8. We can see that documents that follow the same path belong to the same category and suggest better results than with the LDA algorithm. However, a challenge is that the number of paths is still large. The results require human judgment to distinguish which paths should be combined with each other. For example, in Table 8, both cluster 12 and 16 belong to the label "solar cell," but we cannot merge them automatically.

Table 8: Hierarchical LDA path clustering results.

Cluster	**1**	**3**	**6**	**7**	**12**	**14**	**16**	**17**	**26**
MEMS	3	1	0	1	7	66	1	0	0
Tissue Engineering	1	0	0	60	0	0	0	1	57
Graphene	0	0	0	0	0	0	9	127	0
solar cell	1	0	0	0	54	1	66	1	0
RNAi	0	0	53	0	0	0	0	0	1
Genetic Algorithm	55	11	0	0	2	4	0	0	0
Stochastic Programming	3	91	0	0	0	0	0	0	0

# document in cluster	**63**	**103**	**53**	**61**	**63**	**71**	**76**	**129**	**58**

4.4 HDP results

The last model we applied was HDP. In HDP, the researcher controls the number of topics by changing the parameters in its prior dirichlet process. Then the algorithm will generate the topics based on the data. Here we use η=0.3 in the base distribution and the model generates twenty different topics. An advantage with this approach is that more than 98 % of the words are assigned to the top 6 topics. In other words, we can focus on the top 6 topics only. Table 9 shows top words in each topic, and we can easily find that each topic corresponds to a category in Table 1, except for topic 3. The problem with topic 3 is that genetic algorithm and stochastic research fields overlap, and the HDP process actually merges them together.

Table 9: Top 5 words in HDP topics.

Topic	Label	Top five Words				
1	MEMS	Mems	using	devices	fabricated	fabrication
2	Tissue engr	Tissue	cells	cell	engineering	collagen
3	Gen. Alg & Stochastic programming	Model	optimization	algorithm	design	method
4	Solar cell	Solar	cells	cell	efficiency	photovoltaic
5	Graphene	graphene	epitaxial	surface	properties	carbon
6	RNA	Rna	expression	genes	gene	cells

As seen from the clustering results in Table 10, the precision and recall with each of the topics is over 0.8, with F-scores ranging from 0.844 t0 0.957 for Tissue Engineering. Compared with our previous model, the F-score is even better than with the LDA. By using the HDP algorithm, we can obtain better results with less human effort. Although we still have to fine-tune some parameters for the best result, HDP provides an estimated topic number that is close to the true number of topics. The results suggest, however, that HDP can merge topics together or that one topic could be separated into a number of sub-topics.

Table 10: HDP results.

Category	actual	estimated	precision	recall	F-score
MEMS	345	317	0,91	0,84	0,876
Tissue Engineering	217	207	0,98	0,93	0,957
Graphene	180	185	0,87	0,894	0,882

Solar Cell	178	163	0,88	0,8	0,844
RNAi	127	120	0,97	0,913	0,939
Genetic Alg. & Stochastic prog.	213	225	0,88	0,93	0,904

5 Results of verification dataset on energy technologies

Based on the results with the test sample, we use the energy search sample described in Subsection 2.2 to verify the results, as illustrated in Figure 9. Approaching the set with the algorithm that performed the best, HDP, we capture the results outlined as follows.

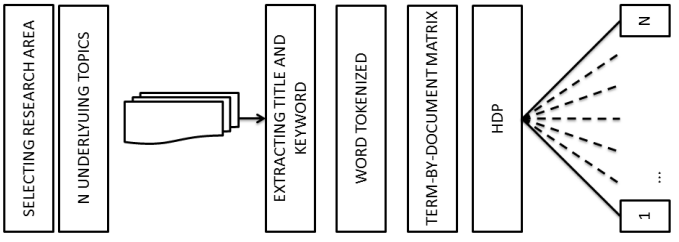

Figure 9: Verification process, based on test dataset findings, on energy technologies.

From Table 11, it is obvious that two general topics, topic 1 and topic 2, contain common words in this dataset. These were labeled "general technology" and "energy technology"—terms used to elaborate on the type of study and terms common to energy technology. Second, we are able to identify the original categories from the top words in each topic and subsequently label topics. In Table 11, we can easily identify Solar, Mining, Petrol, Fuel cells, and Fuel combustion. Steam power can be linked to two topics (6 and 10), with fairly distinct top word sets.

Table11: top 5 word in HDP topics (Energy data)

topic	category	top 5 words				
1	**General technology**	model	data	results	simulation	Study
2	**Energy technology**	energy	system	power	fuel	technology
3	**Fuel combustion**	fuel	engine	emissions	combustion	diesel
4	**Fuel cells**	inf	cell	fuel	water	temperature
5	**Petrol**	reservoir	oil	gas	production	fracture
6	**Steam power**	water	seismic	basin	level	drainage
7	**Energy management**	concrete	test	roof	strength	materials
8	**Solar**	solar	radiation	spacecraft	mission	orbit
9	**Mining**	mining	dose	injuries	exposure	blast
10	**Steam power**	heat	water	boiler	steam	system

In comparison to the test sample, we are unable to identify a one-to-one match for the sample to the topics, but the model identifies several other topics outside the verification sample. Topic modeling, in addition to the research fields searches for the dataset, including two general areas. The emergence of general topics might in a more correlated document set might result either from the common structure of writing scientific texts or from the more general notions expressed within the text.

6 Discussion

List of discussion items to be written:

1. Method is most usable with a dataset that has undergone only a modest cleaning process. A short section on DSSCs could serve as a future research topic.
2. Topic modeling shows promise in being able to categorise documents with minimal human interaction from text rather than giving predetermined categories.
3. Ultimately, the topic model needs to adapt to both time series and communities (Ferrara, 2012).

7 Conclusion

A few final summary remarks pending...

References

Blei DM, Lafferty JD (2007) A correlated topic model of science. The Annals of Applied Statistics pp 17–35, URL http://www.jstor.org/stable/10.2307/4537420

Blei DM, Lafferty JD (2009) Text Mining: Classification, Clustering, and Applications, 10th edn, Taylor and Francis, chap Topic Models, pp 71–94

Blei DM, Ng AY, Jordan MI (2003) Latent dirichlet allocation. the Journal of machine Learning research 3:993“1022, URL http://dl.acm.org/citation.cfm? id=944937

Blei MD, Lafferty JD (2006) Correlated topic models. In: Advances in Neural Information Processing Systems, Proceedings of the 2005 Conference, p 147–155

Borgman C, Furner J (2002) Scholarly communication and bibliometrics. Annual Review of Information Science and Technology 36:3–72

Daim T, Rueda G, Martin H, Gerdsri P (2006) Forecasting emerging technologies: Use of bibliometrics and patent analysis. Technological Forecasting & Social Change 73(8):981–1012

Feldman R, Sanger J (2006) The Text Mining Handbook: Advanced Approaches in Analyzing Unstructured Data. Cambridge University Press

Griffiths TL, Steyvers M (2004) Finding scientific topics. vol 101, pp 5228–5235, URL http://www.pnas.org/content/101/suppl.1/5228.short

Grün B, Hornik K (2011) Topicmodels: An r package for fitting topic models. Journal of Statistical Software 40(13):1–30

Hofmann T (1999) Probabilistic latent semantic indexing. In: Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval, p 50“57, URL http://dl.acm.org/citation.cfm? id=312649

McCallum A (2002) Mallet: A machine learning for language toolkit. URL http://mallet.cs.umass.edu

Newman NC, Porter AL, Newman D, Trumbach CC, Bolan SD (2012) Comparing methods to extract technical content for technological intelligence. In: Technology Management for Emerging Technologies (PICMET), 2012 Proceedings of PICMET’12:, p 1279–1285

Suominen A (Forthcoming) Analysis of technological progression by quantitative measures: a comparison of two technologies. Technology Analysis and Strategic Management

Wallach H (2006) Topic modeling: Beyond bag-of-words. In: In Proceedings of the 23rd International Conference on Machine Learning, Pittsburgh, Pennsylvania, U.S., p 977–984