Running head: FINDING SCIENTIFIC TOPICS

Finding scientific topics

Thomas L. Griffiths

Department of Psychology

Stanford University

Mark Steyvers

Department of Cognitive Sciences

University of California, Irvine

Page count: 24

Figure count: 6

Table count: 0

Word count: 6358

Character count: 57445 (18120 is figures)

Address for correspondence:

Tom Griffiths

NE20-388

Massachusetts Institute of Technology

77 Massachusetts Avenue

Cambridge, MA 02139-4307

e-mail: gruffydd@psych.stanford.edu

phone: (617) 452 3636

Abstract

A first step in identifying the content of a document is determining which topics that document addresses. We describe a generative model for documents, introduced by Blei, Ng and Jordan (2003, Journal of Machine Learning Research, 3, 1107-1135), in which each document is generated by choosing a distribution over topics, and then choosing each word in the document from a topic selected according to this distribution. We then present a Markov chain Monte Carlo algorithm for inference in this model. We use this algorithm to analyze abstracts from the Proceedings of the National Academy of Sciences, using Bayesian model selection to establish the number of topics. We show that the extracted topics capture meaningful structure in the data, consistent with the class designations provided by the authors of the papers, and outline further applications of this analysis, including identifying "hot topics" by examining temporal dynamics, and tagging abstracts to illustrate semantic content.

Finding scientific topics

When a scientist decides to write a paper, one of the first things she or he does is identify an interesting subset of the many possible topics of scientific investigation. The topics addressed by a paper are also one of the first pieces of information a person tries to extract when reading a scientific abstract. Scientific experts know which topics are pursued in their field, and this information plays a role in their assessments of whether papers are relevant to their interests, which research areas are rising or falling in popularity, and how papers relate to one another. Here, we present a statistical method for automatically extracting a representation of documents that provides a first-order approximation to the kind of knowledge available to domain experts. Our method discovers a set of topics expressed by documents, providing quantitative measures that can be used to identify the content of those documents, track changes in content over time, and express the similarity between documents. We will use our method to discover the topics covered by papers in the Proceedings of the National Academy of Sciences in a purely unsupervised fashion, and illustrate how these topics can be used to gain insight into some of the structure of science.

The statistical model we will use in our analysis is a generative model for documents: it reduces the complex process of producing a scientific paper to a small number of simple probabilistic steps, and thus specifies a probability distribution over all possible documents. Generative models can be used to postulate complex latent structures responsible for a set of observations, making it possible to use statistical inference to recover this structure. This kind of approach is particularly useful with text, where the observed data – the words – are explicitly intended to communicate a latent structure – their meaning. The particular generative model we will use, called Latent Dirichlet Allocation, was introduced in [1]. This generative model postulates a latent structure consisting of a set of topics: each document is produced by choosing a distribution over topics, and then generating each word at random from a topic chosen using this distribution.

The plan of the paper is as follows. In the next section, we will describe Latent Dirichlet Allocation and present a Markov chain Monte Carlo algorithm for inference in this model, illustrating the operation of our algorithm on a small dataset. We will then apply our algorithm to a corpus consisting of abstracts from the Proceedings of the National Academy of Sciences from 1991-2001, determining the number of topics needed to account for the information contained in this corpus and extracting a set of topics. We will use these topics to illlustrate the relationships between different scientific disciplines, as well as assessing trends and "hot topics" by analyzing topic dynamics, and using the assignments of words to topics to highlight the semantic content of documents.

Documents, topics, and statistical inference

A scientific paper can deal with multiple topics, and the words that appear in that paper will reflect the particular set of topics it addresses. In statistical natural language processing, one common way of modeling the contributions of different topics to a document is to treat each topic as a probability distribution over words, and view a document as a probabilistic mixture of these topics [1, 2, 3, 4, 5, 6]. If we have T topics, we can write the probability of the ith word in a given document as

$$P(w_i) = \sum_{j}^{T} P(w_i|z_i = j)P(z_i = j), \tag{1}$$

where z_i is a latent variable indicating the topic from which the ith word was drawn and $P(w_i|z_i = j)$ is the probability of the word w_i under the jth topic. $P(z_i = j)$ gives the probability of choosing a word from topics j in the current document, which will vary across different documents.

Intuitively, $P(w|z = j)$ indicates which words are important to a topic, while $P(z)$ is the prevalence of those topics within a document. For example, in a journal which published only articles in mathematics or neuroscience, we could express the probability distribution over words with two topics, one relating to mathematics and the other to neuroscience. The content of the topics would be reflected in $P(w|z = j)$: the "mathematics" topic would give high probability to words like THEORY, SPACE, or PROBLEM, while the "neuroscience" topic would give high probability to words like SYNAPTIC, NEURONS, and HIPPOCAMPAL. Whether a particular document concerns neuroscience, mathematics, or computational neuroscience would depend upon its distribution over topics, $P(z = j)$, which determines how these topics are mixed together in forming documents. The fact that multiple topics can be responsible for the words occurring in a single document discriminates this model from a standard Bayesian classifier, in which it is assumed that all the words in the document come from a single class. The "soft classification" provided by this model, in which each document is characterized in terms of the contributions of multiple topics, has applications in many domains other than text [7].

Viewing documents as mixtures of probabilistic topics makes it possible to formulate the problem of discovering the set of topics that are used in a collection of documents. Given D documents containing T topics expressed over W unique words, we can represent $P(w|z)$ with a set of T multinomial distributions ϕ over the W words, such that $P(w|z = j) = \phi_w^{(j)}$, and $P(z)$ with a set of D multinomial distributions θ over the T topics, such that for a word in document d, $P(z_i = j) = \theta_j^{(d)}$. To discover the set of topics used in a corpus $\mathbf{w} = \{w_1, w_2, \ldots, w_n\}$, where each w_i belongs to some document d_i, we want to obtain an estimate of ϕ that gives high probability to the words that appear in the corpus. One strategy for obtaining such an estimate is to simply attempt to maximize $P(\mathbf{w}|\phi, \theta)$ following from Equation 1 directly, using the Expectation-Maximization [8] algorithm to find maximum likelihood estimates of ϕ and θ [2, 3]. However, this approach is susceptible to problems involving local maxima and is slow to converge [1, 2], encouraging the development of models which make assumptions about the source of θ.

Latent Dirichlet Allocation [1] is one such model, combining Equation 1 with a prior probability distribution on θ to provide a complete generative model for documents. This generative model specifies a simple probabilistic procedure by which new documents can be produced given just a set of topics ϕ, allowing ϕ to be estimated without requiring the estimation of θ. In Latent Dirichlet Allocation, documents are generated by first picking a distribution over topics θ from a Dirichlet distribution, which determines $P(z_i)$ for words in that document. The words in the document are then generated by picking a topic j from this distribution, and then picking a word from that topic according to $P(w_i|z_i = j)$, which is determined by a fixed $\phi^{(j)}$. The estimation problem becomes one of maximizing $P(\mathbf{w}|\phi, \alpha) = \int P(\mathbf{w}|\phi, \theta)P(\theta|\alpha)d\theta$, where $P(\theta)$ is a Dirichlet(α) distribution. The integral in this expression is intractable, and ϕ is thus usually estimated using sophisticated approximations – either variational Bayes [1] or expectation propagation [9].

Using Gibbs sampling to discover topics

Our strategy for discovering topics differs from previous approaches in not explicitly representing ϕ or θ as parameters to be estimated, instead considering the posterior distribution over the assignments of words to topics, $P(\mathbf{z}|\mathbf{w})$. We then obtain estimates of ϕ and θ by examining this posterior distribution. Evaluating $P(\mathbf{z}|\mathbf{w})$ requires solving a problem that has been studied in detail in Bayesian statistics and statistical physics: computing a probability distribution over a large discrete state space. We address this problem using a Monte Carlo procedure, resulting in an algorithm that is easy to implement, requires little memory, and is competitive in speed and performance with existing algorithms.

We will use the probability model for Latent Dirichlet Allocation, with the addition of a Dirichlet prior on ϕ. The complete probability model is thus

$$\begin{aligned}
w_i|z_i, \phi^{(z_i)} &\sim \text{Discrete}(\phi^{(z_i)}) \\
\phi &\sim \text{Dirichlet}(\beta) \\
z_i|\theta^{(d_i)} &\sim \text{Discrete}(\theta^{(d_i)}) \\
\theta &\sim \text{Dirichlet}(\alpha)
\end{aligned}$$

Here, α and β are hyperparameters, specifying the nature of the priors on θ and ϕ. While these hyperparameters could be vector valued as in [1, 9], for the purposes of this paper we will assume symmetric Dirichlet priors, with α and β each having a single value. These priors are conjugate to the multinomial distributions ϕ, θ, allowing us to compute the joint distribution $P(\mathbf{w}, \mathbf{z})$ by integrating out ϕ and θ. Since $P(\mathbf{w}, \mathbf{z}) = P(\mathbf{w}|\mathbf{z})P(\mathbf{z})$ and ϕ and θ only appear in the first and second terms respectively, we can perform these integrals separately. Integrating out ϕ gives the first term

$$P(\mathbf{w}|\mathbf{z}) = \left(\frac{\Gamma(W\beta)}{\Gamma(\beta)^W}\right)^T \prod_{j=1}^{T} \frac{\prod_w \Gamma(n_j^{(w)} + \beta)}{\Gamma(n_j^{(\cdot)} + W\beta)}, \tag{2}$$

in which $n_j^{(w)}$ is the number of times word w has been assigned to topic j in the vector of assignments $\mathbf{z}$ and $\Gamma(\cdot)$ is the standard gamma function. The second term results from integrating out θ, to give

$$P(\mathbf{z}) = \left(\frac{\Gamma(T\alpha)}{\Gamma(\alpha)^T}\right)^D \prod_{d=1}^{D} \frac{\prod_j \Gamma(n_j^{(d)} + \alpha)}{\Gamma(n_{\cdot}^{(d)} + T\alpha)}, \tag{3}$$

where $n_j^{(d)}$ is the number of times a word from document d has been assigned to topic j. Our goal is then to evaluate the posterior distribution

$$P(\mathbf{z}|\mathbf{w}) = \frac{P(\mathbf{w}, \mathbf{z})}{\sum_{\mathbf{z}} P(\mathbf{w}, \mathbf{z})}. \tag{4}$$

Unfortunately, this distribution cannot be computed directly, since the sum in the denominator does not factorize and involves T^n terms, where n is the total number of word instances in the corpus.

Computing $P(\mathbf{z}|\mathbf{w})$ involves evaluating a probability distribution on a large discrete state space, a problem that arises often in statistical physics. Our setting is particularly similar to the Potts model (e.g. [10]), with an ensemble of discrete variables $\mathbf{z}$, each of which can take on values in $\{1, 2, \ldots, T\}$, and an energy function given by $H(\mathbf{z}) \propto -\log P(\mathbf{w}, \mathbf{z}) = -\log P(\mathbf{w}|\mathbf{z})$ $-\log P(\mathbf{z})$. Unlike the Potts model, in which the energy function is usually defined in terms of local interactions on a lattice, here the contribution of each z_i depends on all $\mathbf{z}_{-i}$ through the counts $n_j^{(w)}$ and $n_j^{(d)}$. Intuitively, this energy function favors ensembles of assignments $\mathbf{z}$ that form a good compromise between having few topics per document and having few words per topic, with the terms of this compromise being set by the hyperparameters α and β. The fundamental computational problems raised by this model remain the same as those of the Potts model: we can evaluate $H(\mathbf{z})$ for any configuration $\mathbf{z}$, but the state space is too large to enumerate and we cannot compute the partition function that converts this into a probability distribution (in our case, the denominator of Equation 4). Consequently, we will apply a method that physicists and statisticians have developed for dealing with these problems: sampling from the target distribution using Markov chain Monte Carlo.

In Markov chain Monte Carlo, a Markov chain is constructed to converge to the target distribution and samples are then taken from that Markov chain (see [10, 11, 12]). Each state of the chain is an assignment of values to the variables being sampled, in this case $\mathbf{z}$, and transitions between states follow a simple rule. We use Gibbs sampling [13], known as the heat bath algorithm in statistical physics [10], where the next state is reached by sequentially sampling all variables from their distribution when conditioned on the current values of all other variables and the data. In order to apply this algorithm we need the full conditional distribution $P(z_i|\mathbf{z}_{-i}, \mathbf{w})$. This can be obtained via a probabilistic argument, or by cancellation of terms in Equations 2 and 3, yielding

$$P(z_i = j|\mathbf{z}_{-i}, \mathbf{w}) \propto \frac{n_{-i,j}^{(w_i)} + \beta}{n_{-i,j}^{(\cdot)} + W\beta} \frac{n_{-i,j}^{(d_i)} + \alpha}{n_{-i,\cdot}^{(d_i)} + T\alpha}, \quad (5)$$

where $n_{-i,\cdot}^{(\cdot)}$ is a count that does not include the current assignment of z_i. This result is quite intuitive: the first ratio expresses the probability of w_i under topic j, and the second expresses the probability of topic j in document d_i. Critically, these counts are the only information necessary for computing the full conditional distribution, allowing the algorithm to be implemented efficiently by caching the relatively small set of non-zero counts.

Having obtained the full conditional distribution, the Monte Carlo algorithm is then straightforward. The z_i are initialized to values in $\{1, 2, \ldots T\}$, determining the initial state of the Markov chain. We do this using an on-line version of the Gibbs sampler, using Equation 5 to assign words to topics, but with counts that are computed from the subset of the words seen so far, rather than the full data. The chain is then run for a number of iterations, each time finding a new state by sampling each z_i from the distribution specified by Equation 5. Since the only information needed to apply Equation 5 is the number of times a word is assigned to a topic and the number of times a topic occurs in a document, the algorithm can be run with minimal memory requirements by caching the sparse set of non-zero counts and updating them whenever a word is reassigned. After enough iterations for the chain to approach the target distribution, the current values of the z_i are recorded. Subsequent samples are taken after an appropriate lag, to ensure that their autocorrelation is low [10, 11].

With a set of samples from the posterior distribution $P(\mathbf{z}|\mathbf{w})$, statistics that are independent of the content of individual topics can be computed by integrating across the full set of samples. For any single sample we can estimate ϕ and θ from the value $\mathbf{z}$ via

$$\hat{\phi}_j^{(w)} = \frac{n_j^{(w)} + \beta}{n_j^{(\cdot)} + W\beta} \tag{6}$$

$$\hat{\theta}_j^{(d)} = \frac{n_j^{(d)} + \alpha}{n_\cdot^{(d)} + T\alpha}. \tag{7}$$

These values correspond to the predictive distributions over new words w and new topics z conditioned on $\mathbf{w}$ and $\mathbf{z}$.[1]

A graphical example

In order to illustrate the operation of the algorithm and to show that it runs in comparable time to existing methods of estimating ϕ, we generated a small dataset in which the output of the algorithm can be shown graphically. The dataset consisted of a set of 2000 images each containing 25 pixels in a 5×5 grid. The intensity of any pixel is specified by an integer value between zero and infinity. This dataset is of exactly the same form as a word-document co-occurrence matrix constructed from a database of documents, with each image being a document, each pixel a word, and the intensity of a pixel being its frequency. The images were generated by defining a set of 10 topics corresponding to horizontal and vertical bars, as shown in Figure 1 (a), then sampling a multinomial distribution θ for each image from a Dirichlet distribution with $\alpha = 1$, and sampling 100 pixels (words) according to Equation 1. A subset of the images generated in this fashion are shown in Figure 1 (b). While some images show evidence of many samples from a single topic, it is difficult to discern the underlying structure of most images.

We applied our Gibbs sampling algorithm to this dataset, together with the two algorithms that have previously been used for inference in Latent Dirichlet Allocation – variational Bayes [1] and expectation propagation [9].[2] We divided the dataset into 1000 training images and 1000 test images, and ran each algorithm four times, using the same initial conditions for all three algorithms on a given run. These initial conditions were found by an online application of Gibbs sampling, as mentioned above. Variational Bayes and expectation propagation were run until convergence, and Gibbs sampling was run for 1000 iterations. All three algorithms used a fixed Dirichlet prior on θ, with $\alpha = 1$. We tracked the number of floating point operations per iteration for each algorithm and computed the test set perplexity for the estimates of ϕ provided by the algorithms at several points. Perplexity is a standard measure of performance for statistical models of natural language [14], and is defined as $\exp\{-\log P(\mathbf{w}_{test}|\phi)/n_{test}\}$, where $\mathbf{w}_{test}$ and n_{test} indicate the identities and number of words in the test set respectively. Perplexity indicates the uncertainty in predicting a single word: lower values are better, and chance performance results in a perplexity equal to the size of the vocabulary, which is 25 in this case. The perplexity for all three models was evaluated using importance sampling as in [9], and the estimates of ϕ used for evaluating Gibbs sampling were each obtained from a single sample as in Equation 6. The results of these computations are shown in Figure 1 (c). All three algorithms are able to recover the underlying topics, and Gibbs does so more rapidly than either variational Bayes or expectation propagation. A graphical illustration of the operation of the Gibbs sampler is shown in Figure 2. The log-likelihood stabilizes quickly, in a fashion consistent across multiple runs, and the topics expressed in the dataset slowly emerge as appropriate assignments of words to topics are discovered.

These results show that Gibbs sampling can be competitive in speed with existing algorithms, although further tests with larger datasets involving real text are necessary to evaluate the strengths and weaknesses of the different algorithms. The effects of including the Dirichlet (β)

prior in the model and the use of methods for estimating the hyperparameters α and β need to be assessed as part of this comparison. A variational algorithm for this "smoothed" model is described in [1], which may be more comparable to the Gibbs sampling algorithm described here, as well as an empirical Bayes procedure for estimating the hyperparameters. Ultimately, these different approaches are complementary rather than competitive, providing different means of performing approximate inference that can be selected according to the demands of the problem.

Model selection

The statistical model we have described is conditioned on three parameters, which we have suppressed in the above equations: the Dirichlet hyperparameters α and β, and the number of topics T. Our algorithm is easily extended to allow α and β to be sampled as well as $\mathbf{z}$, but this can slow the convergence of the Markov chain. Our strategy in this paper will be to fix α and β, and explore the consequences of varying T. The choice of α and β can have important implications for the results produced by the model. In particular, increasing β can be expected to decrease the number of topics used to describe a dataset, since it reduces the impact of sparsity in Equation 2. The value of β thus affects the granularity of the model: a corpus of documents can be sensibly factorized into a set of topics at several different scales, and the particular scale assessed by the model will be set by β. With scientific documents, a large value of β would lead the model to find a relatively small number of topics, perhaps at the level of scientific disciplines, while smaller values of β will produce more topics that address specific areas of research.

Given values of α and β, the problem of choosing the appropriate value for T is a problem of model selection, which we will address using a standard method from Bayesian statistics [15]. To a Bayesian statistician faced with a choice between a set of statistical models, the natural response is to compute the posterior probability of that set of models given the observed data. The key constituent of this posterior probability will be the likelihood of the data given the model, integrating over all parameters in the model. In our case, the data are the words in the corpus, $\mathbf{w}$, and the model is specified by the number of topics, T, so we wish to compute the likelihood $P(\mathbf{w}|T)$. The complication is that this requires summing over all possible assignments of words to topics $\mathbf{z}$. However, we can approximate $P(\mathbf{w}|T)$ by taking the harmonic mean of a set of values of $P(\mathbf{w}|\mathbf{z}, T)$ when $\mathbf{z}$ is sampled from the posterior $P(\mathbf{z}|\mathbf{w}, T)$ [15]. Our Gibbs sampling algorithm provides such samples, and the value of $P(\mathbf{w}|\mathbf{z}, T)$ can be computed from Equation 2.

The topics of science

The algorithm outlined above can be used to find the topics that account for the words used in a set of documents. We applied this algorithm to the abstracts of papers published in the Proceedings of the National Academy of Sciences (PNAS) from 1991 to 2001, with the aim of discovering some of the topics addressed by scientific research. We first used Bayesian model selection to identify the number of topics needed to best account for the structure of this corpus, and then conducted a detailed analysis using the selected number of topics. Our detailed analysis involved examining the relationship between the topics discovered by our algorithm and the class designations supplied by PNAS authors, using topic dynamics to identify "hot topics," and using the topic assignments to highlight the semantic content in abstracts.

How many topics?

To evaluate the consequences of changing the number of topics T, we used the Gibbs sampling algorithm outlined in the previous section to obtain samples from the posterior distribution over $\mathbf{z}$ at several choices of T. We used all 28,154 abstracts published in the Proceedings of the National Academy of Sciences from 1991 to 2001, with each of these abstracts constituting a single document in the corpus (we will use the words abstract and document interchangeably from this point forward). Any delimiting character, including hyphens, was used to separate words, and we deleted any words that occurred in less than five abstracts or belonged to a standard "stop" list used in computational linguistics, including numbers, individual characters, and some function words. This gave us a vocabulary of 20,551 words, which occurred a total of 3,026,970 times in the corpus.

For all runs of the algorithm, we used $\beta = 0.1$ and $\alpha = 50/T$, keeping constant the sum of the Dirichlet hyperparameters, which can be interpreted as the number of virtual samples contributing to the smoothing of θ. This is a relatively small value of β, and can be expected to result in a fine-grained decomposition of the corpus into topics that address quite specific research areas. We computed an estimate of $P(\mathbf{w}|T)$ for T values of 50, 100, 200, 300, 400, 500, 600 and 1,000 topics. For all values of T except the last we ran eight Markov chains, discarding the first 1,000 iterations, and then took ten samples from each chain at a lag of 100 iterations. In all cases, the log-likelihood values stabilized within a few hundred iterations, as in Figure 2. The simulation with 1,000 topics was more time consuming, and thus we used only six chains, taking two samples from each chain after 700 initial iterations, again at a lag of 100 iterations.

Estimates of $P(\mathbf{w}|T)$ were computed based on the full set of samples for each value of T, and are shown in Figure 3. The results suggest that the data are best accounted for by a model incorporating 300 topics. $P(\mathbf{w}|T)$ initially increases as a function of T, reaches a peak at $T = 300$, and then decreases thereafter. This kind of profile is often seen when varying the dimensionality of a statistical model, with the optimal model being rich enough to fit the information available in the data, yet not so complex as to begin fitting noise. As mentioned above, the value of T found through this procedure depends on the choice of α and β, and it will also be affected by specific decisions made in forming the dataset, such as the use of a stop list and the inclusion of documents from all PNAS classes. By using just $P(\mathbf{w}|T)$ to choose a value of T, we are assuming very weak prior constraints on the number of topics. $P(\mathbf{w}|T)$ is just the likelihood term in the inference to $P(T|\mathbf{w})$, and the prior $P(T)$ might overwhelm this likelihood if we had a particularly strong preference for a smaller number of topics.

Scientific topics and classes

When authors submit a paper to PNAS, they choose one of three major categories, indicating whether a paper belongs to the Biological, the Physical, or the Social Sciences, as well as one of 33 minor categories, such as Ecology, Pharmacology, Mathematics, or Economic Sciences.[3] Having a class designation for each abstract in the corpus provides two opportunities: firstly, since the topics recovered by our algorithm are purely a consequence of the statistical structure of the data, we can evaluate whether the class designations pick out differences between abstracts that can be expressed in terms of this statistical structure; and secondly, we can use the class designations to illustrate how the distribution over topics can reveal relationships between documents and between document classes.

We used a single sample taken after 2,000 iterations of Gibbs sampling, and computed estimates of $\theta^{(d)}$ via Equation 7.[4] Using these estimates, we computed a mean θ vector for each minor category, considering just the 2,620 abstracts published in 2001. We then found the most diagnostic topic for each minor category, defined to be the topic j for which the ratio of θ_j for that category to the sum of θ_j across all other categories was greatest. The results of this analysis are shown in Figure 4. The matrix shown in the upper half of the Figure indicates the mean value of θ for each minor category, restricted to the set of most diagnostic topics. The strong diagonal is a consequence of our selection procedure, with diagnostic topics having high probability within the classes for which they are diagnostic, but low probability in other classes. The off-diagonal elements illustrate the relationships between classes, with similar classes showing similar distributions across topics.

The distributions over topics for the different classes illustrate how this statistical model can capture similarity in the semantic content of documents. Figure 4 reveals relationships between specific minor categories, such as Ecology and Evolution, as well as some of the correspondences within major categories – for example, the minor categories in the Physical and Social Sciences show much greater commonality in the topics appearing in their abstracts than do the Biological

Sciences. The results can also be used to assess how much different disciplines depend upon particular methods. For example, topic 39, relating to mathematical methods, receives reasonably high probability in Applied Mathematics, Applied Physical Sciences, Chemistry, Engineering, Mathematics, Physics, and Economic Sciences, suggesting that mathematical theory is particularly relevant to these disciplines.

The content of the diagnostic topics themselves is shown in the lower half of Figure 4, listing the five words given highest probability by each topic. In some cases, a single topic was the most diagnostic for several classes: topic 2, containing words relating to global climate change, was diagnostic of Ecology, Geology, and Geophysics; topic 280, containing words relating to evolution and natural selection, was diagnostic of both Evolution and Population Biology; topic 222, containing words relating to cognitive neuroscience, was diagnostic of Psychology as both a Biological and a Social Science; topic 39, containing words relating to mathematical theory, was diagnostic of both Applied Mathematics and Mathematics; and topic 270, containing words having to do with spectroscopy, was diagnostic of both Chemistry and Physics. The remaining topics were each diagnostic of a single minor category, and generally seemed to contain words relevant to enquiry in that discipline. The only exception was topic 109, diagnostic of Economic Sciences, which contains words generally relevant to scientific research. This may be a consequence of the relatively small number of documents in this class – only three in 2001 – which makes the estimate of θ extremely unreliable. Topic 109 also serves to illustrate that not all of the topics found by the algorithm correspond to areas of research – some of the topics picked out scientific words that tend to occur together for other reasons, like those that are used to describe data, or those that express tentative conclusions.

Finding strong diagnostic topics for almost all of the minor categories suggests that these categories have differences that can be expressed in terms of the statistical structure recovered by our algorithm. Since the topics discovered by the algorithm are found in a completely unsupervised fashion, using no information except the distribution of the words themselves, this implies that the minor categories capture real differences in the content of abstracts, at the level of the words used by authors. It also shows that this algorithm finds genuinely informative structure in the data, producing topics that connect with our intuitive understanding of the semantic content of documents.

Hot and cold topics

Historians, sociologists, and philosophers of science, as well as scientists themselves, recognize that topics rise and fall in the amount of scientific interest they generate, although whether this is the result of social forces or rational scientific practice is the subject of debate (e.g. [16, 17]). Since our analysis reduces a corpus of scientific documents to a set of topics, it is straightforward to analyze the dynamics of these topics as a means of gaining insight into the dynamics of science. If understanding these dynamics is the goal of our analysis, we can formulate more sophisticated generative models that incorporate parameters describing the change in the prevalence of topics over time. Here, we present a basic analysis based on a post-hoc examination of the estimates of θ produced by the model. Being able to identify the "hot topics" in science at a particular point is one of the most attractive applications of this kind of model, providing quantitative measures of the prevalence of particular kinds of research that may be useful for historical purposes as well as determining targets for scientific funding. Analysis at the level of topics provides the opportunity to combine information about the occurrences of a set of semantically related words with cues that come from the content of the remainder of the document, potentially highlighting trends that might be less obvious in analyses that consider only the frequencies of single words.

To find topics that consistently rose or fell in popularity from 1991 to 2001, we conducted a linear trend analysis on θ_j by year, using the same single sample as in our previous analyses. We applied this analysis to the sample used to generate Figure 4. Consistent with the idea that science shows strong trends, with topics rising and falling regularly in popularity, 54 of the topics showed a statistically significant increasing linear trend, while 50 showed a statistically significant decreasing linear trend, both at the $p = 0.0001$ level. The three hottest and coldest topics, assessed by the size of the linear trend test statistic, are shown in Figure 5. The hottest topics discovered through this analysis were topics 2, 134, and 179, corresponding to global warming and climate change, gene knockout techniques, and apoptosis (programmed cell death), the subject of the 2002 Nobel Prize in Physiology. The cold topics were not topics that lacked prevalence in the corpus, but those that showed a strong decrease in popularity over time. The coldest topics were 37, 289, and 75, corresponding to sequencing and cloning, structural biology, and immunology. All of these topics were very popular around 1991, and fell in popularity over the period of analysis. The Nobel Prizes again provide a good means of validating these trends, with prizes being awarded for work on sequencing in 1993, and immunology in 1989.

Tagging abstracts

Each sample produced by our algorithm consists of a set of assignments of words to topics. We can use these assignments to identify the role that words play in documents. In particular, we can tag each word with the topic to which it was assigned, and use these assignments to highlight topics that are particularly informative about the content of a document. The abstract shown in Figure 6 is tagged with topic labels as superscripts. Words without superscripts were not included in the vocabulary supplied to the model. All assignments come from the same single sample used in our previous analyses, illustrating the kind of words assigned to the evolution topic discussed above (topic 280).

This kind of tagging is mainly useful for illustrating the content of individual topics and how individual words are assigned, and was used for this purpose in [1]. It is also possible to use the results of our algorithm to highlight conceptual content in other ways. For example, if we integrate across a set of samples, we can compute a probability that a particular word is assigned to the most prevalent topic in a document. This probability provides a graded measure of the importance of a word that uses information from the full set of samples, rather than a discrete measure computed from a single sample. This form of highlighting is used to set the contrast of the words shown in Figure 6, and picks out the words that determine the topical content of the document. Such methods might provide a means of increasing the efficiency of searching large document databases, particularly since it can be modified to indicate words belonging to the topics of interest to the searcher.

Conclusion

We have presented a statistical inference algorithm for Latent Dirichlet Allocation [1], a generative model for documents in which each document is viewed as a mixture of topics, and shown how this algorithm can be used to gain insight into the content of scientific documents. The topics recovered by our algorithm pick out meaningful aspects of the structure of science, and reveal some of the relationships between scientific papers in different disciplines. The results of our algorithm have several interesting applications which can make it easier for people to understand the information contained in large knowledge domains, including exploring topic dynamics, and indicating the role that words play in the semantic content of documents.

The results we have presented employ the simplest model of this kind, and the simplest algorithm for generating samples. In future research, we intend to extend this work by exploring both more complex models and more sophisticated algorithms. While in this paper we have focused on the analysis of scientific documents, as represented by the articles published in the Proceedings of the National Academy of Sciences, the methods and applications we have presented are relevant to a variety of other knowledge domains. Latent Dirichlet Allocation is a statistical model that is appropriate for any collection of documents, from e-mail records and newsgroups to the entire World Wide Web. Discovering the topics underlying the structure of such datasets is the first step to being able to visualize their content and discover meaningful trends.

References

[1] David M. Blei, A. Y. Ng, and M. I. Jordan. *Journal of Machine Learning Research*, 3:993–1022, 2003.

[2] Thomas Hofmann. *Machine Learning Journal*, 42:177–196, 2001.

[3] David Cohn and Thomas Hofmann. In *Advances in Neural Information Processing Systems 13*, pages 430–436, Cambridge, 2001. MIT Press.

[4] R. Iyer and M. Ostendorf. In *Proceedings of the International Conference on Spoken Language Processing 1*, pages 236–239, 1996.

[5] B. Bigi, R. De Mori, M. El-Beze, and T. Spriet. In *1997 IEEE Workshop on Automatic Speech Recognition and Understanding Proceedings*, pages 535–542, 1997.

[6] N. Ueda and K. Saito. In *Advances in Neural Information Processing Systems 15*, Cambridge, 2003. MIT Press.

[7] E. A. Erosheva. In *Bayesian Statistics 7*. Oxford University Press, Oxford, 2003.

[8] A. P. Dempster, N. M. Laird, and D. B. Rubin. *Journal of the Royal Statistical Society, B*, 39:1–38, 1977.

[9] Thomas Minka and John Lafferty. Expectation-propagation for the generative aspect model. In *Proceedings of the 18th Conference on Uncertainty in Artificial Intelligence*, 2002.

[10] M. E. J. Newman and G. T. Barkema. *Monte Carlo methods in statistical physics*. Oxford University Press, Oxford, 1999.

[11] W. R. Gilks, S. Richardson, and D. J. Spiegelhalter. *Markov chain Monte Carlo in practice*. Chapman & Hall, New York, 1996.

[12] J. S. Liu. *Monte Carlo strategies in scientific computing*. Springer, New York, 2001.

[13] S. Geman and D. Geman. *IEE Trans. on Pattern Analysis and Machine Intelligence*, 6:721–741, 1984.

[14] C. D. Manning and H. Schutze. *Foundations of statistical natural language processing*. MIT Press, Cambridge, MA, 1999.

[15] R. E. Kass and A. E. Raftery. *Journal of the American Statistical Association*, 90:773–795, 1995.

[16] T. S. Kuhn. *The structure of scientific revolutions*. University of Chicago Press, Chicago, 2nd edition, 1970.

[17] W. Salmon. In C. W. Savage, editor, *Scientific Theories, Minnesota Studies in the Philosophy of Science 14*. University of Minnesota Press, Minneapolis, 1990.

Author Note

We are grateful to Josh Tenenbaum, Dave Blei and Jun Liu for thoughtful comments that improved this paper. Kevin Boyack provided the PNAS class designations, Shawn Cokus wrote the random number generator, Tom Minka wrote the code used for the comparison of algorithms, and several simulations were performed on the BlueHorizon supercomputer at the San Diego Supercomputer Center. This work was supported by funds from the NTT Communication Sciences laboratory and a Stanford Graduate Fellowship to the first author.

Footnotes

[1]These estimates cannot be combined across samples for any analysis that relies upon the content of specific topics. This issue arises due to a lack of identifiability: since mixtures of topics are used to form documents, the probability distribution over words implied by the model is unaffected by permutations of the indices of the topics. Consequently, there need be no correspondence between individual topics across samples – just because two topics have index j in two samples, there is no reason to expect that similar words were assigned to those topics in those samples. However, statistics insensitive to permutation of the underlying topics can be computed by aggregating across samples.

[2]The implementations of variational Bayes and expectation propagation were provided by Tom Minka, and are available at
`http://www.stat.cmu.edu/~minka/papers/aspect.html`

[3]Anthropology and Psychology can be chosen as a minor category for papers in both Biological and Social Sciences. We will treat these minor categories as distinct for the purposes of our analysis.

[4]In this and other analyses, similar results were obtained by examining samples across multiple chains, up to the permutation of topics, and the choice of this particular sample to display the results was arbitrary.

Figure Captions

Figure 1. (a) Graphical representation of 10 topics, combined together to produce "documents" like those shown in (b), where each image is the result of 100 samples from a unique mixture of these topics. (c) Performance of three algorithms on this dataset: variational Bayes (VB), expectation propagation (EP), and Gibbs sampling. Lower perplexity indicates better performance, with chance being a perplexity of 25. Estimates of the standard errors are smaller than the plot symbols, which mark 1, 5, 10, 20, 50, 100, 150, 200, 300 and 500 iterations.

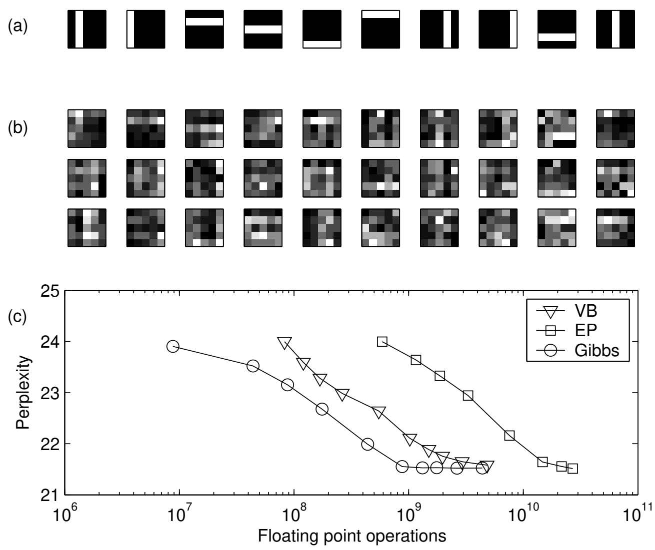
Figure 2. Results of running the Gibbs sampling algorithm. The log-likelihood, shown in the left panel, stabilizes after a few hundred iterations. Traces of the log-likelihood are shown for all four runs, illustrating the consistency in values across runs. Each row of images on the right hand side of the figure shows the estimates of the topics after a certain number of iterations within a single run, matching the points indicated on the left hand side. These points correspond to 1, 2, 5, 10, 20, 50, 100, 150, 200, 300 and 500 iterations. The topics expressed in the data gradually emerge as the Markov chain approaches the posterior distribution.

Figure 3. Model selection results, showing the log-likelihood of the data for different settings of the number of topics, T. The estimated standard errors for each point were smaller than the plot symbols.

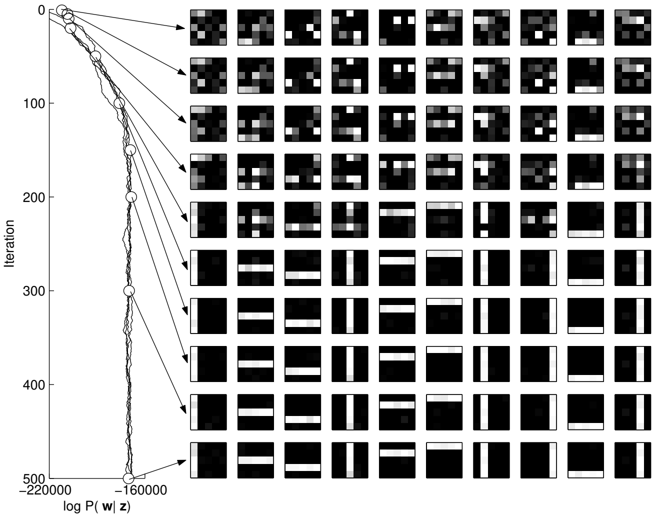
Figure 4. The upper panel shows the mean values of θ at each of the diagnostic topics for all 33 PNAS minor categories, computed using all abstracts published in 2001. Higher probabilities are indicated with darker cells. The five most probable words in the topics themselves are shown in the lower part of the figure, in the same order as on the horizontal axis.

Figure 5. The plots show the dynamics of the three hottest and three coldest topics from 1991 to 2001, defined as those topics that showed the strongest positive and negative linear trends. The 12 most probable words in those topics are shown below the plots.

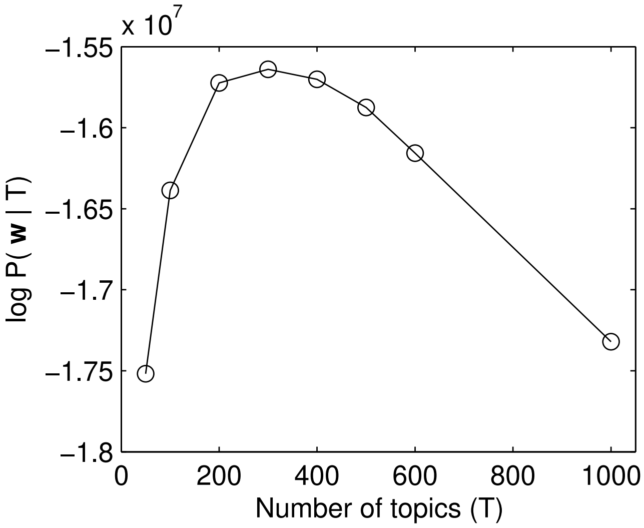
Figure 6. A PNAS abstract, tagged according to topic assignment. The superscripts indicate the topics to which individual words were assigned in a single sample, while the contrast level reflects the probability of a word being assigned to the most prevalent topic in the abstract, computed across samples.

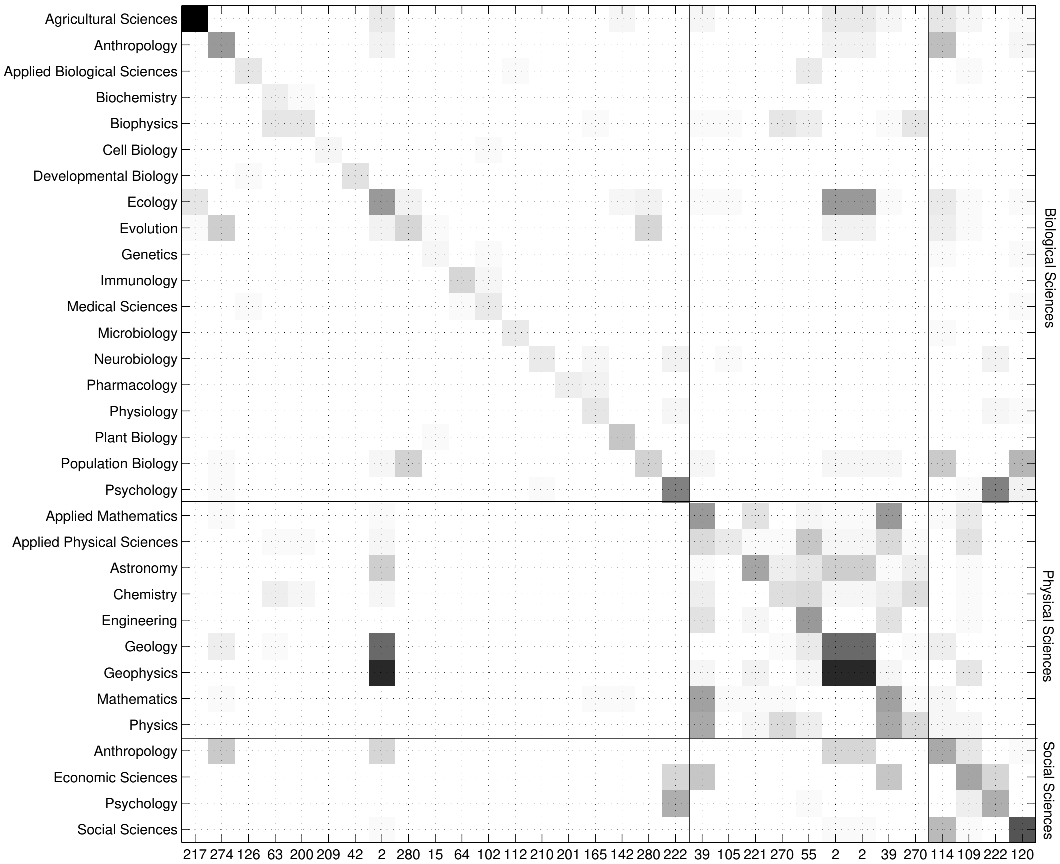


(a)
(b)
(c)
VB
EP
Gibbs
Perplexity
25
24
23
22
21
10^6
10^7
10^8
10^9
10^{10}
10^{11}
Floating point operations



Iteration
0
100
200
300
400
500
−220000
−160000
log P(**w**| **z**)


x 10^7
−1.55
−1.6
−1.65
−1.7
−1.75
−1.8
log P(w | T)
0
200
400
600
800
1000
Number of topics (T)



217
INSECT
MYB
PHEROMONE
LENS
LARVAE

274
SPECIES
PHYLOGENETIC
EVOLUTION
EVOLUTIONARY
SEQUENCES

126
GENE
VECTOR
VECTORS
EXPRESSION
TRANSFER

63
STRUCTURE
ANGSTROM
CRYSTAL
RESIDUES
STRUCTURES

200
FOLDING
NATIVE
PROTEIN
STATE
ENERGY

209
NUCLEAR
NUCLEUS
LOCALIZATION
CYTOPLASM
EXPORT

42
NEURAL
DEVELOPMENT
DORSAL
EMBRYOS
VENTRAL

2
SPECIES
GLOBAL
CLIMATE
CO2
WATER

280
SPECIES
SELECTION
EVOLUTION
GENETIC
POPULATIONS

15
CHROMOSOME
REGION
CHROMOSOMES
KB
MAP

64
CELLS
CELL
ANTIGEN
LYMPHOCYTES
CD4

102
TUMOR
CANCER
TUMORS
HUMAN
CELLS

112
HOST
BACTERIAL
BACTERIA
STRAINS
SALMONELLA

210
SYNAPTIC
NEURONS
POSTSYNAPTIC
HIPPOCAMPAL
SYNAPSES

201
RESISTANCE
RESISTANT
DRUG
DRUGS
SENSITIVE

165
CHANNEL
CHANNELS
VOLTAGE
CURRENT
CURRENTS

142
PLANTS
PLANT
ARABIDOPSIS
TOBACCO
LEAVES

222
CORTEX
BRAIN
SUBJECTS
TASK
AREAS

39
THEORY
TIME
SPACE
GIVEN
PROBLEM

105
HAIR
MECHANICAL
MB
SENSORY
EAR

221
LARGE
SCALE
DENSITY
OBSERVED
OBSERVATIONS

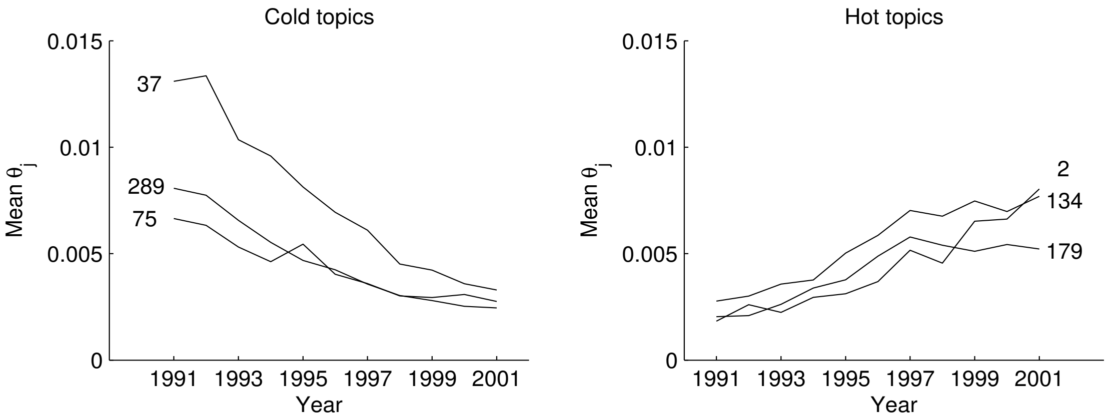
270
TIME
SPECTROSCOPY
NMR
SPECTRA
TRANSFER

55
FORCE
SURFACE
MOLECULES
SOLUTION
SURFACES

114
POPULATION
POPULATIONS
GENETIC
DIVERSITY
ISOLATES

109
RESEARCH
NEW
INFORMATION
UNDERSTANDING
PAPER

120
AGE
OLD
AGING
LIFE
YOUNG



37	289	75	2	134	179
CDNA	KDA	ANTIBODY	SPECIES	MICE	APOPTOSIS
AMINO	PROTEIN	ANTIBODIES	GLOBAL	DEFICIENT	DEATH
SEQUENCE	PURIFIED	MONOCLONAL	CLIMATE	NORMAL	CELL
ACID	MOLECULAR	ANTIGEN	CO2	GENE	INDUCED
PROTEIN	MASS	IGG	WATER	NULL	BCL
ISOLATED	CHROMATOGRAPHY	MAB	ENVIRONMENTAL	MOUSE	CELLS
ENCODING	POLYPEPTIDE	SPECIFIC	YEARS	TYPE	APOPTOTIC
CLONED	GEL	EPITOPE	MARINE	HOMOZYGOUS	CASPASE
ACIDS	SDS	HUMAN	CARBON	ROLE	FAS
IDENTITY	BAND	MABS	DIVERSITY	KNOCKOUT	SURVIVAL
CLONE	APPARENT	RECOGNIZED	OCEAN	DEVELOPMENT	PROGRAMMED
EXPRESSED	LABELED	SERA	EXTINCTION	GENERATED	MEDIATED

(PNAS, 1991, vol. 88, 4874-4876)

A generalized[3] fundamental[146] theorem[267] of natural[280] selection[280] is derived[233] for populations[280] incorporating[149] both genetic[280] and cultural[280] transmission[25]. The phenotype[3] is determined[17] by an arbitrary[3] number[287] of multiallelic[3] loci[3] with two[271]-factor[60] epistasis[280] and an arbitrary[149] linkage[3] map[3], as well as by cultural[280] transmission[25] from the parents[280]. Generations[280] are discrete[69] but partially[275] overlapping[146], and mating[280] may be nonrandom[280] at either the genotypic[280] or the phenotypic[280] level[199] (or both). I show[25] that cultural[280] transmission[25] has several[173] important[173] implications[17] for the evolution[280] of population[280] fitness[280], most notably[230] that there is a time[72] lag[72] in the response[213] to selection[280] such that the future[287] evolution[280] depends[105] on the past selection[280] history[280] of the population[280].