Science | DOI:10.1145/1859204.1859210 Gary Anthes

Topic Models Vs. Unstructured Data

With topic modeling, scientists can explore and understand huge collections of unlabeled information.

TOPIC MODELING, AN amalgam of ideas drawn from computer science, mathematics, and cognitive science, is evolving rapidly to help users understand and navigate huge stores of unstructured data. Topic models use Bayesian statistics and machine learning to discover the thematic content of unlabeled documents, provide application-specific roadmaps through them, and predict the nature of future documents in a collection. Most often used with text documents, topic models can also be applied to collections of images, music, DNA sequences, and other types of information.

Because topic models can discover the latent, or hidden, structure in documents and establish links between documents, they offer a powerful new way to explore and understand information that might otherwise seem chaotic and unnavigable.

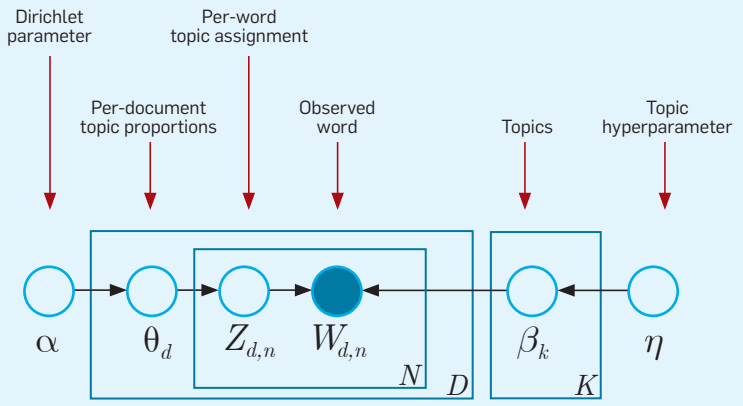
The base on which most probabilistic topic models are built today is latent Dirichlet allocation (LDA). Applied to a collection of text documents, LDA discovers "topics," which are probability distributions over words that co-occur frequently. For example, "software," "algorithm," and "kernel" might be found likely to occur in articles about computer science. LDA also discovers the probability distribution of topics in a document. For example, by examining the word patterns and probabilities, one article might be tagged as 100% about computer science while another might be tagged as 10% computer science and 90% neuroscience.

LDA algorithms are built on assumptions of how a "generative" process might create a collection of documents from these probability distributions. The process does that by first assigning to each document a probability distribution across a small number of topics from among, say, 100 possible topics in the collection. Then, for each of these hypothetical documents, a topic is chosen at random (but weighted by its probability distribution), and a word is generated at random from that topic's probability distribution across the words. This hypothetical process is repeated over and over, each word in a document occurring in proportion to the distribution of topics in the document and the distribution of words in a topic, until all the documents have been generated.

LDA takes that definition of how the documents to be analyzed might have been created, "inverts" the process, and works backward to explain the observed data. This process, called "posterior probabilistic inference," essentially says, "Given these observed data, and given the model for document-creation posited in the generative process, what conditional distribution of words over topics and of topics over documents resulted in the data I see?" It both defines the topics in a collection and explains the proportions of these topics in each document, and in so doing it discovers the underlying semantic structure of the documents.

LDA and its derivatives are examples of unsupervised learning, meaning that the input data is not labeled; the models work with no prior knowledge of the topics in the documents. The models can perform their inference by a number of different algorithms, but they all work by machine learning. They start with random assumptions about the probability distributions, try them out on the data to see how well they fit, then update them and try again.

Latent Dirichlet allocation.



Each piece of the structure is a random variable.

More Modular, More Scalable

LDA is essentially a technical refinement—making it more modular and scalable—of the topic modeling technique called probabilistic latent semantic indexing. Introduced in 1999 by Jan Puzicha and Thomas Hofmann, probabilistic latent semantic indexing was derived from Latent Semantic Indexing, which was developed in the late 1980s by Scott Deerwester, Susan T. Dumais, George W. Furnas, Thomas

K. Landauer, and Richard Harshman. Because of its modularity, LDA has become the springboard for a host of refinements and extensions. For example, David Blei, a computer scientist at Princeton University, who co-developed LDA with Andrew Ng and Michael Jordan, has inserted into LDA a stochastic method that relates topics to each other and displays them in potentially infinite tree-like hierarchies, with the various branches located at different levels of abstraction. Unlike LDA, hierarchical LDA does not have to be told the number of topics in advance, and each topic has the potential to break into smaller subtopics without limit.

Blei also codeveloped the correlated topic model (CTM), which can find correlations between topics, which LDA can't do. CTM, for instance, could recognize that the topic "neuroscience" is more closely related to "biology" than it is to "geology."

Another LDA extension, called dynamic topic modeling (DTM), takes into account the evolution of a collection over time. It can recognize that a paper from 1910 on "phrenology" and one from 2010 on "neural networks" might both be classified in the evolving field of "neuroscience," even though they use vastly different vocabularies. "The DTM connects these topics across time so you can see how the lexicon evolved," says DTM pioneer John Lafferty, a professor of computer science, machine learning, and statistics at Carnegie Mellon University.

Researchers at the University of California, Irvine have developed two LDA derivatives that discover the relationships among the entities in news stories—people, organizations, and locations—and the topics found by classical LDA. They applied these "entity-topic models" to a collection of 330,000 *New York Times* articles and found they were especially adept at predicting the occurrence of entities based on the words in a document, after having been trained on test documents. For example, they report that an article about the Sept. 11, 2001 terrorist attacks was likely to include the entities "FBI," "Taliban," and "Washington."

Blei says three developments in the past 10 years have led to rapid advancements in topic modeling: the emergence of LDA as a kind of development platform, advancements in the use of machine learning to perform statistical inference, and "the emergence of very large, unlabeled data sets."

Latent Dirichlet allocation both defines the topics in a collection and explains the proportions of these topics in each document, thereby discovering the underlying semantic structure of the documents.

Despite the models' growing popularity, Blei offers several caveats about their use. He warns against the blind acceptance of results suggested by the models as conclusive. "It is important to be careful," he said. For example, running topic models multiple times, with the algorithms choosing different random initializations of the topics, can lead to very different results. "Also, it can be important to check sensitivity to different choices in the model," says Blei. "There are many dials to tune in topic modeling."

Mark Steyvers, a professor of cognitive sciences at the University of California, Irvine, is exploring how people can analyze documents when they have little idea of what is contained in them. Steyvers and his colleagues have recently used topic models in three real-world situations. In the first, a lawyer had received a large stack of papers related to a lawsuit, and needed a summary picture of their contents. In the second project, funded by a U.S. intelligence agency, the task was to examine huge feeds of email and documents and to provide analysts with lists of their topics. In the third, a government agency wanted to

Milestones

CS Awards

The Glushko-Samuelson Foundation, National Academy of Engineering, and ACM's Special Interest Group on Security, Audit and Control (SIGSAC) recently recognized leading computer scientists for their research and leadership.

DAVID E. RUMELHART PRIZE

Judea Pearl, director of the Cognitive Systems Laboratory in the department of computer science at the University of California Los Angeles, is the recipient of the eleventh David E. Rumelhart Prize. The prize is awarded annually by the Glushko-Samuelson Foundation to an individual or collaborative team making a significant contemporary contribution to the theoretical foundations of human cognition. Pearl has developed Bayesian networks that can be used to represent and draw inferences from probabilistic knowledge in a highly transparent and computationally tractable fashion.

ARTHUR M. BUECHE AWARD

Anita Jones, a university professor emerita in the computer science department at the University of Virginia, was awarded the Arthur M. Bueche Award from the National Academy of Engineering for "leadership in the development of U.S. science and technology policy and the development of technologies for national security, including technical contributions to high-performance computing and cybersecurity."

SIGSAC AWARDS

SIGSAC presented its top honors to Jan Camenisch of IBM Research-Zurich and Bhavani Thuraisingham of the University of Texas at Dallas for their contributions to the computer and communications security community. Camenisch received the SIGSAC Outstanding Innovation Award for his theoretical work on privacy-enhancing cryptographic protocols and his leadership in their practical realization. Thuraisingham received the SIGSAC Outstanding Contribution Award for her seminal research contributions and leadership in data and applications security over the past 25 years.

—Jack Rosenberger

understand the topics and inter-topic relationships among the hundreds of thousands of grants awarded by it and sister agencies.

The ultimate application may be to help understand how the human mind works. Steyvers is experimenting with topic modeling to shed light on how humans retrieve words from memory, based on associations with other words. He runs the models on educational documents to produce crude approximations of the topics learned by students, then compares the accuracy of recall, based on word associations, of the students and models. Sometimes the models make mistakes in their word and topic associations, which are shedding light on the memory mistakes of humans. What's needed, Steyvers says, is nothing less than "a model of the human mind."

Meanwhile, computer scientists are looking for ways to make algorithms more efficient and to structure problems for parallel processing, so that huge problems, such as topic modeling the entire World Wide Web, can be run on large clusters of computers.

Fernando Pereira, a research director at Google, says a number of experimental systems of probabilistic topic modeling are being investigated at the company. The systems could provide better Google search results by grouping similar terms based on context. A topic model might discover, for instance, that a search for the word "parts," used in an automobile context, should include "accessories" when it is also used in an automobile context. (The two words are seen as synonyms if both are used in the same context; in this case, automobiles.) Google does some of that now on a limited basis using heuristic models, but they tend to require a great deal of testing and tuning, Pereira says.

One of the advantages of the LDA framework is the ease with which one can define new models.

"I can't point to a major success yet with the LDA-type models, partly because the inference is very expensive," Pereira says. "While they are intriguing, we haven't yet gotten to the point that we can say, 'Yes, this is a practical tool.' "

But, says Tom Griffiths, director of the Computational Cognitive Science Lab at University of California, Berkeley, "We are seeing a massive growth in people applying these models to new problems. One of the advantages of this [LDA] framework is it's pretty easy to define new models."

Further Reading

Blei, D. and Lafferty, J.
Dynamic topic models. *Proceedings of the 23rd International Conference on Machine Learning*, Pittsburgh, PA, June 25–29, 2006.

Blei, D. and Lafferty, J.
Topic models, *Text Mining: Classification, Clustering, and Applications*, (Srivastava, A. and Sahami, M., Eds), Taylor & Francis, London, England, 2009.

Chang, J., Boyd-Graber, J., Gerrish, S., Wang, C., and Blei, D.
Reading tea leaves: How humans interpret topic models. Twenty-Third Annual Conference on Neural Information Processing Systems, Vancouver, British Columbia, Canada, Dec. 7–12, 2009.

Deerwester, S., Dumais, S.T., Furnas, G.W., Landauer, T.K., and Harshman, R.
Indexing by latent semantic analysis, *Journal of the American Society for Information Science 41*, 6, 1990.

Newman, D., Chemudugunta, C., Smyth, P., and Steyvers, M.
Statistical entity-topic models. The Twelfth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Philadelphia, PA, August 23–26, 2006.

Gary Anthes is a technology writer and editor based in Arlington, VA.

History

Building Babbage's Analytical Engine

A British programmer and author who wants to build a computer based on 19th century designs has about 3,000 people pledging donations to his project. John Graham-Cumming, author of *The Geek Atlas*, hopes to build an Analytical Engine invented by English mathematician Charles Babbage and first proposed in 1837. Babbage is often called "the father of computing."

The Analytical Engine, originally meant to be constructed of iron and brass and operated with steam power, will have the equivalent of 1.7 kilobyte of memory and be capable of four arithmetic operations: left and right shift, and comparison/jump operations. It will be very slow, with a single addition taking about 13,000 times as long as on a Z80, an 8-bit microprocessor from the mid-1970s.

"I think it would be an inspirational machine to see," Graham-Cumming said via email. "People could literally see how a computer operates since the architecture of the machine is close to the architecture of modern computers.

"It will answer the question: Could Victorians have had the computer? Others will have to answer the question: And what difference would that have made?" he says.

One challenge will be deciding what version of the machine to build as Babbage was continually refining his designs. Graham-Cumming is focusing his efforts on a version called Plan 28, but says more research is needed. Eventually, he plans to create a 3D virtual model of the Analytical Engine, work out all the bugs, and build it. As part of the project, Babbage's papers will be digitized and made available online.

The project could take up to five years and cost £1 million (about $1.6 million U.S.). Graham-Cumming has started a Web site, http://www.plan28.org, to solicit donations, and plans to start work once he has 10,000 pledges. Another Babbage machine, a calculator called the Difference Engine No. 2, was built by the London Science Museum in 1991.

In 2009, Graham-Cumming led an online petition demanding an apology from the British government for its treatment of World War II mathematician and code-breaker Alan Turing, who was prosecuted for being a homosexual. The petition, which attracted international media attention, was successful.

—Neil Savage