Science | DOI:10.1145/1378727.1378732

Kirk L. Kroeker

Finding Diamonds in the Rough

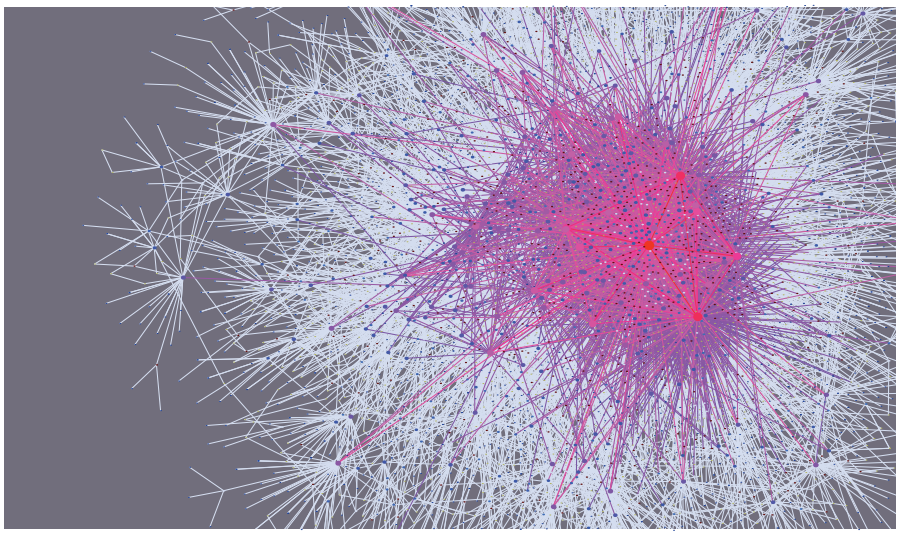
Spectral graph theory has proven to be very useful for text search and retrieval and for refining predictive-analysis systems.

THERE IS A little-known approach to information analysis that has built the foundation for many of the information technologies that we now consider to be givens of the 21st century. The strategy, called spectral graph theory, is well known among mathematicians and those working with massive data sets, but has not received a great deal of credit in the mainstream media as being an important method for understanding key relationships in data sets consisting of millions or even billions of nodes. With roots in the early 20th century, spectral graph theory and the corresponding interpretative method of spectral analysis were initially used as a theoretical approach to solving specialized math problems in which relationships between certain classes would otherwise be difficult to ascertain.

The origin of spectral graph theory can be traced to Markov chains, harmonic analysis, statistics, and spectral geometry, with mathematicians asking Zen-like questions such as "How can you hear the shape of a drum?" In the intervening years, spectral graph theory was applied in several areas, with IBM researcher Alan J. Hoffman even using the technique in the 1970s to partition circuitry on silicon chips. In the 1980s, Microsoft's Susan Dumais, the University of Colorado's Tom Landauer, and their colleagues used spectral analysis to cluster documents in a technique called latent semantic analysis. However, it wasn't until the 1990s that spectral graph theory became one of the most important tools in understanding how to conduct text search and retrieval more efficiently.

Spectral graph theory, which might be characterized as a kind of advanced Boolean logic on steroids (in the sense that it can help simplify extremely complex relationships among class members that are also members of other classes), has become useful not only to those working in Web search, but also to those refining predictive-analysis systems of the type used to determine what books you might find most interesting or what movies you might like to watch.

In her early work with spectral analysis, Fan Chung, an Akamai professor in Internet mathematics at the University of California, San Diego, applied the theory to light patterns to help determine the molecular composition of stars. Chung is now taking spectral analysis in new directions for Web search, such as how to analyze and improve Google's PageRank, which was introduced in 1998 and based on a math strategy called random walks. "One of the main applications of spectral graph



INTERNET BGP-AS ROUTING GRAPH REPRESENTATION COURTESY OF ROSS M. RICHARDSON, DEPT. OF MATHEMATICS, UNIVERSITY OF CALIFORNIA, SAN DIEGO.

theory is to analyze the rate of convergence of random walks," says Chung. "So you can see it takes just one further extension to analyze PageRank."

Chung points out that the algorithms used in spectral graph theory have become so efficient that it is now possible for graduate students to apply the theory to data sets consisting of millions of nodes. "Five years ago I would have never imagined my students would be able to deal with millions of data points," she says. "The power of the algorithm allows us to deal with this many data points on a single, dedicated PC."

One of the metaphors that Chung uses to characterize spectral graph theory is a sandbox. The traditional way of identifying the relationships between grains of sand in a sandbox is to string them together, one by one. Spectral graph theory essentially makes such work much more efficient by instantly "cutting" to the relationships between all the grains of sand so it is possible to quickly identify, as Chung puts it, "diamonds in the rough."

Identifying Data Patterns

To understand how spectral graph theory cuts to these relationships, you must understand eigenvalues and eigenvectors, which are values placed on a group of data points. Unlike Boolean operators, which deal with either-or classes that have no relative weights, eigenvalues and eigenvectors enable mathematicians and others working with spectral graph theory to apply finely tuned weights to members of a population with the goal of identifying interesting data patterns. In a social network, for example, an eigenvalue is a number indicating a particular pattern or cluster. The pattern could be men and women, for example. The larger the eigenvalue, the more important the pattern. An eigenvector would be a particular weight applied to the men-women partition, with plus values applied to women and minus values applied to men.

One strength of spectral graph theory is the ability to apply multiple eigenvalues and eigenvectors to the equation, so you might have a value for, say, conservatives and liberals or those who read magazines and those who don't. In the end, every member of the social network has a numeric score. The idea is to cut to the eigenvalues that are important, then apply eigenvectors to determine what the clusters are. Rather than working with either-or relationships, which make it relatively difficult to efficiently apply gradient values, spectral graph theory enables researchers to tease out useful clusters by applying multiple values to each member of a population.

While the most well-known application of the technology today is in Web search, Milena Mihail, an associate professor in the College of Computing at the Georgia Institute of Technology, points out that spectral graph theory can be applied to just about any technology outside the Web to reveal which clusters in a massive data set are the most important. "The power of the method does not come only from the fact that it has some mathematical justification," she says. "It comes from the fact that, computationally, it is very easy to implement. Doing spectral analysis on a very large data set can be done in almost linear time, on the spot."

According to Mihail, spectral graph theory has a disadvantage in that it needs numeric stability. Mihail says the next major breakthrough in spectral analysis might be for distributed applications, such as peer-to-peer networks, in which little numeric stability exists because no central authority has total knowledge of the network.

In talking about spectral graph theory and the early days of Web search, Mihail points to the work of Prabhakar Raghavan, head of research at Yahoo!, and Cornell University's Jon Kleinberg as being fundamental to the modern-day concept of automating search results through the application of spectral analysis. Raghavan headed up the Computer Science Principles department at IBM's Almaden Research Center and led IBM's Clever Project on Web search and page popularity. The Clever Project has received a great deal of attention for developing efficient

Eigenvectors applied to pages for the search query jaguar*.

(jaguar*) Authorities: principal Eigenvector		
.370	http://www2.ecst.csuchico.edu/_jschlich/Jaguar/jaguar.html	
.347	http://www-und.ida.liu.se/_t94patsa/jserver.html	
.292	http://tangram.informatik.uni-kl.de:8001/_rgehm/jaguar.html	
.287	http://www.mcc.ac.uk/ dlms/Consoles/jaguar.html	Jaguar page
(jaguar jaguars) Authorities: 2nd nonprincipal vector, positive end		
.255	http://www.jaguarsnfl.com/	Official Jacksonville Jaguars NFL Web site
.137	http://www.nando.net/SportServer/football/nfl/jax.html	Jacksonville Jaguars home page
.133	http://www.ao.net/_brett/jaguar/index.html	Brett's Jaguar page
.110	http://www.usatoday.com/sports/football/sfn/sfn30.htm	Jacksonville Jaguars
(jaguar jaguars) Authorities: 3rd nonprincipal vector, positive end		
.227	http://www.jaguarvehicles.com/	Jaguar Cars Global home page
.227	http://www.collection.co.uk/	The Jaguar collection official Web site
.211	http://www.moran.com/sterling/sterling.html	
.211	http://www.coys.co.uk/	

ways to determine search results through measuring link popularity.

Kleinberg, a professor of computer science at Cornell, also worked on IBM's Clever Project. Kleinberg's research in this area relies on what he calls hubs and authorities: The way to find good hub pages is to find good authority pages that they link to, and the way to find good authorities is to find good hubs that link to the authorities. In a sense, this circular problem amounts to a mathematical version of Oroboros, the mythological serpent that swallows its own tail. And it is here that spectral graph theory made one of its most important contributions to Web search by offering a way to break that circularity.

Kleinberg explains that overcoming this problem involves using an eigenvector to determine the relative value of a node's endorsement. If one node endorses its neighbor, the quality of the endorsement would be equivalent to the sum of the qualities of every node endorsing that neighbor node, and so on. You can keep playing this hub-and-authority game all the way out to infinity. Eventually, though, things start to stabilize. "What they start to stabilize on is an eigenvector of the graph," Kleinberg says. "In Web search, an eigenvector of the graph is essentially the infinite limit of a sequence of increasingly refined estimates of quality."

The table here, derived from Kleinberg's seminal 1997 IBM research report titled *Authoritative Sources in a Hyperlinked Context,* serves as a good example for how automated semantic separation can be accomplished with eigenvectors. On the most positive weight nodes of the first eigenvector for the search "jaguar," spectral analysis turned up sites related to the Atari Jaguar video game console. On the second eigenvector, spectral analysis turned up sites related to the Jacksonville Jaguars football team. And on the third eigenvector, spectral analysis turned up Jaguar car dealers. The numbers in the table represent coordinates from the first few eigenvectors; thus, the pages with large coordinates in each of these eigenvectors correspond to results for different meanings of the query term. This is the sense in which the different eigenvectors are serving to distinguish between different meanings of a query.

According to Kleinberg, the ability to break through the problem of circularity with eigenvectors is why spectral graph theory is the natural mathematical language for talking about ranking on the Web. "But I think there is more basic research that needs to be done because it's not clear that we've seen the full power of the technique," he says.

Spectral graph theory is the natural mathematical language for talking about ranking on the Web, says Jon Kleinberg.

Kleinberg points out that most applications of spectral graph theory have been related to finding global, large-scale features, such as the highest-quality Web pages or the most significant user clusters in a customer database. He suggests that spectral analysis is not as naturally suited to finding needle-in-a-haystack details, such as 100 interesting Web pages in billions of Web pages or a few unusual purchases in terabytes of customer data. And when it comes to the theory itself, Kleinberg says that in some cases there are no guarantees for the quality of the solutions that spectral analysis offers. "So, trying to actually prove some guarantees on the performance of algorithms based on spectral analysis is a very nice open question," he says. "And in the process of trying to find proofs for the performance of spectral algorithms, we might in the process invent new algorithms."

Breakthroughs or not, it seems likely that spectral analysis will remain a favorite tool among mathematicians and computer scientists working with vast amounts of data, including those doing work in computer vision, image recognition, and any other area of research that might benefit from advanced pattern-recognition strategies. C

Kirk L. Kroeker works in communications and has written extensively about the impact of emerging technologies.

Science

Natural Storage

DNA is seen by many researchers as the classic information storage system. Just four bases (adenine, thymine, guanine, and cytosine) are required to code ongoing combinations of multiple amino acids that ultimately steer the cellular machinery. Inspired by the notion that DNA could be used to store nonbiological data, a team of researchers in Japan has created the first DNA strand made from artificial bases. The researchers, in a paper recently published in the *Journal of the American Chemical Society*, explain all the components of their DNA product are nonnatural, yet they spontaneously form duplexes with the corresponding opposite base, and these bonds are very similar properties to those found in natural DNA. The team hopes this artificial DNA could offer a range of real-world applications, from using DNA to store data, to using it in biomedical and nanotechnology settings.

Computer Science

CS Award Winners

IEEE Honors

Gordon E. Moore will receive the IEEE Medal of Honor at the upcoming IEEE Annual Honors Ceremony, Sept. 20 in Quebec City, Canada. Other individuals to be recognized for work that "improved the world in which we live" are: Ralph H. Baer, Sir Timothy Berners-Lee, Joseph Bordogna, Don Coppersmith, Teuvo Kohonen, Leslie Lamport, Chrysostomos L. Nikias, Richard F. Rashid, Raj Reddy, and Alan Jay Smith.

EATCS Award

Leslie G. Valiant received the 2008 EATCS Award from the European Association for Theoretical Computer Science in recognition of his distinguished career in theoretical computer science.