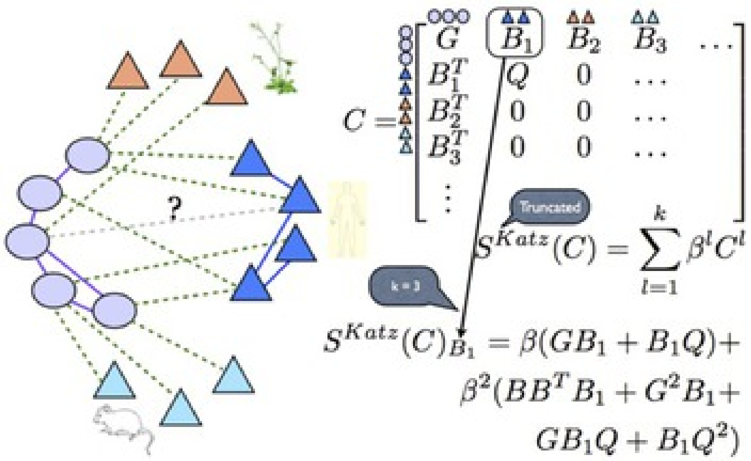
PHYS.ORG

Nomadic computing speeds up Big Data analytics

November 5 2015


$C = \begin{bmatrix} G & B_1 & B_2 & B_3 & \dots \\ B_1^T & Q & 0 & \dots & \\ B_2^T & 0 & 0 & \dots & \\ B_3^T & 0 & 0 & \dots & \\ \vdots & & & & \end{bmatrix}$
?
Truncated
$S^{Katz}(C) = \sum_{l=1}^{k} \beta^l C^l$
k = 3
$S^{Katz}(C)_{B_1} = \beta(GB_1 + B_1Q) + \beta^2(BB^TB_1 + G^2B_1 + GB_1Q + B_1Q^2)$

Schematic of the proposed approach to predicting gene-disease associations. First, the researchers construct gene and disease features using different sources. Then, they perform Inductive matrix completion using row and column features. The shaded region in the P matrix corresponds to genes or diseases with at least one known association. Credit: Nagarajan Natarajan and Inderjit Dhillon

How do Netflix or Facebook know which movies you might like or who you might want to be friends with?

Here's a hint: It starts with a few trillion data points and involves some complicated math and a lot of smart computer programming.

The ability to make sense of massive amounts of raw data—a process known as data analytics—has already brought benefits to consumers and long-lost friends and is beginning to have a real impact in medicine, law enforcement and public services.

Inderjit Dhillon, a professor of computer science at The University of Texas at Austin, is a leader in this new world of big data. He was named a 2014 Fellow of the Association for Computing Machinery for his contributions to large-scale data analytics, machine learning and computational mathematics.

"Nowadays, there is an abundance of massive networks," Dhillon says. "These networks may be explicit or implicit, and we want to use predictive analytics on these networks to see what they can tell us."

He is among those who have realized it's possible to tame highly complex data (or "data with high dimensionality," in the lingo of the field) by using machine learning to reduce data to its most meaningful parameters. His approaches are widely adopted in science and industry.

"People have come to realize over the last two decades that indeed, data that comes from different applications often has special structures," Dhillon says. "For example, in the case of a very high dimensional regression, it's only a small number of dimensions that may actually matter."

Computing anytime, anywhere

Imagine Netflix wants to recommend movies that customers might like based on their ratings. A customer who ranks several buddy comedies highly will probably like other films in that genre.

"If the user-movie matrix was general, there's no way that you could infer missing elements because they could be anything," Dhillon says.

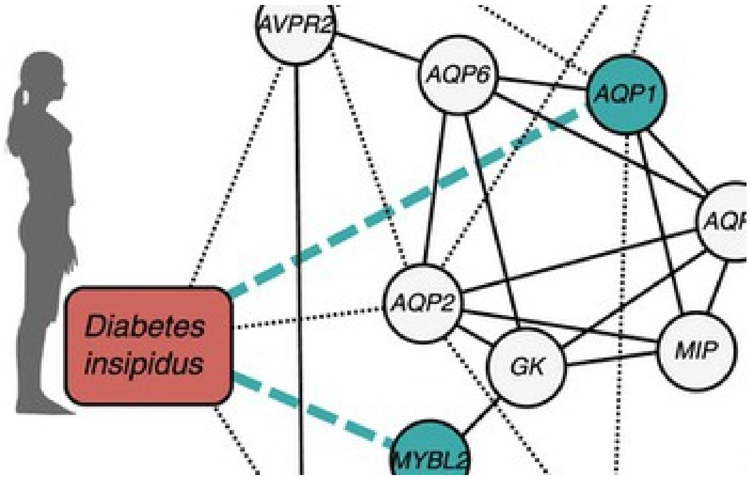
"But when you make the assumption that people have a finite number of tastes, or factors that determine what they like, the problem becomes tractable."

In collaboration with Vishwanathan's group at the University of California, Santa Cruz, and with support from the National Science Foundation (NSF) and computing resources from the Texas Advanced Computing Center (TACC), Dhillon and his group have recently developed a new data analysis tool, called NOMAD. It stands for "non-locking, stochastic multi-machine algorithm for asynchronous and decentralized matrix completion."

NOMAD can pull insights from data much faster than other current state-of-the-art tools. It is also able to explore datasets, including some of the largest available, that break other leading software.

Among the problems Dhillon and his team are exploring with NOMAD are "topic modeling," where the system automatically determines the appropriate topics related to billions of documents, and "recommender systems," where, based on millions of users and billions of records, the system can suggest appropriate items to buy or people to meet.

In cases like these, fitting the data on a single computer is often impossible. Instead, users distribute data among a large number of host systems. At the heart of NOMAD is a new method for orchestrating computations among those hosts.



The local network around the human disease diabetes insipidus and two genes highly ranked by Catapult, AQP1 (top ranked candidate) and MYBL2 (ranked as number 40). AQP1 is ranked higher than MYBL2 because there are more paths from diabetes insipidus to AQP1 than to MYBL2, both through model organism phenotypes and through the gene-gene network. Only genes and phenotypes that are associated to both diabetes insipidus and the predicted genes AQP1 and MYBL2 are shown. Credit: U. Martin Singh-Blom, Nagarajan Natarajan, Ambuj Tewari, John O. Woods, Inderjit S. Dhillon, Edward M. Marcotte

"Suppose you have a massive computational problem and you need to run it on datasets that do not fit in a computer's memory," Dhillon says. "If you want the answer in a reasonable amount of time, the logical thing to do would be to distribute the computations over different machines."

Easier said than done.

Traditionally, systems have managed distributed computations through a process known as bulk synchronization. After computing a solution, each of the processors involved (often thousands) stops and communicates with the others, passing along the results of their computations.

In the program that Dhillon and Vishwanathan have developed, the communication is done asynchronously—the processors no longer stop and communicate at the same time.

"We are trying to develop an asynchronous method where each parameter is, in a sense, a nomad," he explained. "The parameters go to different processors, but instead of synchronizing this computation followed by communication, the nomadic framework does its work whenever a variable is available at a particular processor."

As soon as the work is done, the nomadic parameter travels to another processor. As a result, there is no waiting around for all the other processors in the system to finish computing.

Dhillon and his team used the Stampede supercomputer at TACC, the ninth most powerful in the world, and a system called Rustler that specializes in running machine learning algorithms to develop and test NOMAD.

Managing the process in this way, the team was able to get a superlinear speedup. That means when they ran the code on one thousand processors, they were able to solve the problem more than one thousand times faster. They were also able to seamlessly handle millions of documents with billions of occurrences of words—or millions of users and billions of ratings—in a reasonable time period.

The team reported their results in the Proceedings of the VLDB Endowment in July 2014 and at the World Wide Web conference in May 2015.

The research applies broadly to one of the most pressing challenges today, according to Amy Apon, a program director at NSF.

"Traditionally, machine learning inference algorithms run on a single large—and sometimes expensive—server, and this limits the size of the problem that can be addressed," she says. "This team has noticed a property of some machine learning algorithms that if a few slowly changing variables can be only occasionally synchronized, then the work can be more easily distributed across different computers.

"Their clever mathematical approach is opening doors to running machine algorithms on the kind of massive-scale, distributed, commodity computers that we find in today's cloud computing environment."

Thinking of genes as a social network

When we hear data analytics, we typically think of social networks like Facebook or LinkedIn, but Dhillon has applied his tools to a different type of network—the gene networks involved in disease.

Teaming up with Edward Marcotte in the biology department at The University of Texas at Austin, they have applied the methods and state-of-the-art algorithms from Dhillon' s group to the gene networking problems Marcotte is trying to solve.

"We thought about the relationships or linkages between genes and diseases as a network," Dhillon says. "From there, the question is: Can you do prediction from this sort of data to determine what genes have a propensity to be linked to which diseases? And you can do that by actually developing new mathematics."

Marcotte and Dhillon used evolutionary relationships to track down gene networks involved in human health. With this method, they showed it was possible to predict gene-disease associations for diabetes based on functional gene associations and gene-phenotype associations in model organisms.

The result of the work was published in *PLOS ONE* in 2013, and a further study by Dhillon and his group where they used a similar system appeared in *Bioinformatics* last year.

In October, Dhillon was awarded a three-year follow-up grant from NSF to continue work on nomadic algorithms for machine learning in the cloud, beginning in January 2016.

He and his group are now extending their study to medical informatics, where the problem might be to predict co-morbidities or the propensity for re-hospitalization.

"Data can tell a thousand stories," says Dhillon, "and the challenge is to develop new mathematics and methods that can extract the knowledge and discard the spurious. The possibilities are endless."

Provided by National Science Foundation

Citation: Nomadic computing speeds up Big Data analytics (2015, November 5) retrieved 21 May 2024 from https://phys.org/news/2015-11-nomadic-big-analytics.html