Published on *ScienceNordic* (http://sciencenordic.com)

On a mission to save search engines

Technology[1]
Technology[1]Information technology [2]The Future [3]Denmark [4]Videnskab.dk [5]

With the exploding amounts of data, search engines will not always work reliably in the age of Big Data. A Danish researcher has set out to find new methods of rescuing our search engines.

At a young age, Rasmus Pagh, now a professor at the IT University in Copenhagen, discovered that some searches get so heavy with data that the users become impatient waiting for the search result. This prompted him to start looking for smarter ways of searching for information online.

Today, a common alternative to a search engine seeping through every piece of data in the database is to base the search on random samples. The search methods used by Google, Facebook, insurance databases and corporate marketing departments are nowhere near as effective as they might appear on the surface. There are actually no guarantees that you will find everything you searched for.

This is the problem that Pagh has now set out to solve. He has attracted international attention with his research into Big Data – the huge amounts of data that researchers, businesses and authorities have had access to in recent years – and the special algorithms used for searching in these vast piles of data.

Better methods for similarity searches

He will head a comprehensive new research project backed with a €1.9 million ‘Consolidator Grant’ from the European Research Council.

“It’s a bit of a high-risk gamble, because we’re trying to do something that a number of pretty clever people have attempted before. But we believe we may have the additional clues that can lead us towards some kind of a breakthrough,” he says.

The project aims to find new and improved methods for so-called ’similarity searches’. These are searches where the user has one piece of information and is looking for something that is similar to it, but without knowing exactly what the searched-for object looks like.

He will also examine mathematically whether the methods that many of the search engines use in practice actually work. The methods are based on algorithms and ad-hoc solutions but have no theoretical basis.

No guarantees

As an example of a similarity search, Pagh uses a photo of a lion. You want to find other pictures that are similar to the one you have, but you don’t know whether the photo collection contains other lions, other cats or perhaps a painting of a lion.

It is far from certain that the search finds all results that are similar to the lion on your picture, since the methods used today are based on samples in which the quality of the search results is unknown. So even though you get some results, you cannot be certain that you have found all the relevant data.

”Imagine a doctor searching for cases of a disease similar to the one he wants to know about and he does not find a treatment that saved a patient somewhere else. Here you would want to know for certain that the software finds it if it’s there to be found,” he says.

You are likely to be unlucky

With the existing methods, the computer first converts our lion picture into electronic 1s and 0s, or yeses and nos. The software takes a sample of a variety of parameters: is it a hat? No. Is it an animal? Yes, etc.

The computer then finds other pictures where it is possible to answer yes or no to the same questions. However, it selects the parameters to be examined at random, which explains why the computer randomly goes looking for either other pictures of animals or other pictures without hats.

“If you’re lucky, the right parameters are selected, but there is a fairly high likelihood that you’re unlucky. These are the theoretically best algorithms available today – and they are only slightly faster than looking through all the data.”

Getting harder to use good methods

However, since this method – known as locality sensitive hashing – also places great demands on computing resources, many use methods where they do not know the mathematical quality of the search result.

In the future, it will become increasingly difficult to use the good methods.

The amount of data is growing at an exploding rate – every two years it doubles, according to market intelligence firm IDC. Although computing power is also growing, the ability to carry out complex computations is not growing anywhere near as fast as the amount of data is.

Searches that today can still go through all the data will be limited to the search methods that – for now – are statistically uncertain.

”Part of the motivation for the research project is to ensure that we will continue to be able to use similarity searches in the future. Another part is to describe the existing systems, which have some algorithms that work most of the time. But sometimes they do not work, and it is not always entirely clear when they do not work,” says Pagh.

Read the Danish version of this article at videnskab.dk [6]

Every two years, the amount of data doubles, and even though computing power is also growing, the ability to carry out complex computations isn?t growing anywhere near as fast as the amount of data. (Photo: <a href=" http://www.shutterstock.com/" target="_blank">Shutterstock</a>) [7]

Professor Rasmus Pagh has set out to find better and more secure methods for similarity searches ? i.e. searches in which the user is looking for information that is similar, but not identical, to the information they have. [8]

big data.jpg [9]

Fact box

‘Big Data’ refers to data sets so large and complex that it becomes difficult to process using on-hand database management tools or traditional data processing applications.

There are no fixed limits to when data sets are large enough to be called Big Data, but computer giant Intel suggests that when a company collects more than 300 terabytes per day, it qualifies as Big Data.

It?s a bit of a high-risk gamble, because we?re trying to do something that a number of pretty clever people have attempted before. But we believe we may have the additional clues that can lead us towards some kind of a breakthrough.
Rasmus Pagh

Software innovation: peripheral users are also key players [10]
Rasmus Pagh?s profile [11] "Near-Optimal Hashing Algorithms for Approximate Nearest Neighbor in High Dimensions", Communications of the ACM [12] IDC report: "The digital universe in 2020: Big Data, Bigger Digital Shadow s, and Biggest Growth in the Far East" [13]
"Finding Correlations in Subquadratic Time, with Applications to Learning Parities and the Closest Pair Problem", Foundations of Computer Science.Stanford University (2013), DOI: 10.1109/FOCS.2012.27 [14]

Rune Wriedt Larsen [15]
Dann Vinther

March 12, 2014 - 06:46
This field is not in use. The footer is displayed in the mini panel called "Footer (mini panel)"

Source URL: http://sciencenordic.com/mission-save-search-engines

Links:
[1] http://sciencenordic.com/category/section/technology
[2] http://sciencenordic.com/information-technology
[3] http://sciencenordic.com/category/keywords/future
[4] http://sciencenordic.com/category/countries/denmark
[5] http://sciencenordic.com/videnskabdk
[6] http://videnskab.dk/teknologi/dansk-forsker-skal-redde-sogemaskinerne
[7] http://sciencenordic.com/sites/default/files/big data.jpg
[8] http://sciencenordic.com/sites/default/files/rasmus pagh.jpg
[9] http://sciencenordic.com/sites/default/files/big data_0.jpg
[10] http://sciencenordic.com/software-innovation-peripheral-users-are-also-key-players
[11] https://pure.itu.dk/portal/en/persons/rasmus-pagh(5b58251e-1c93-4fd4-a765-8b7cdb8746d3).html
[12] http://mags.acm.org/communications/200801/#pg119
[13] http://idcdocserv.com/1414
[14] http://theory.stanford.edu/~valiant/papers/corrFull.pdf
[15] http://sciencenordic.com/content/rune-wriedt-larsen