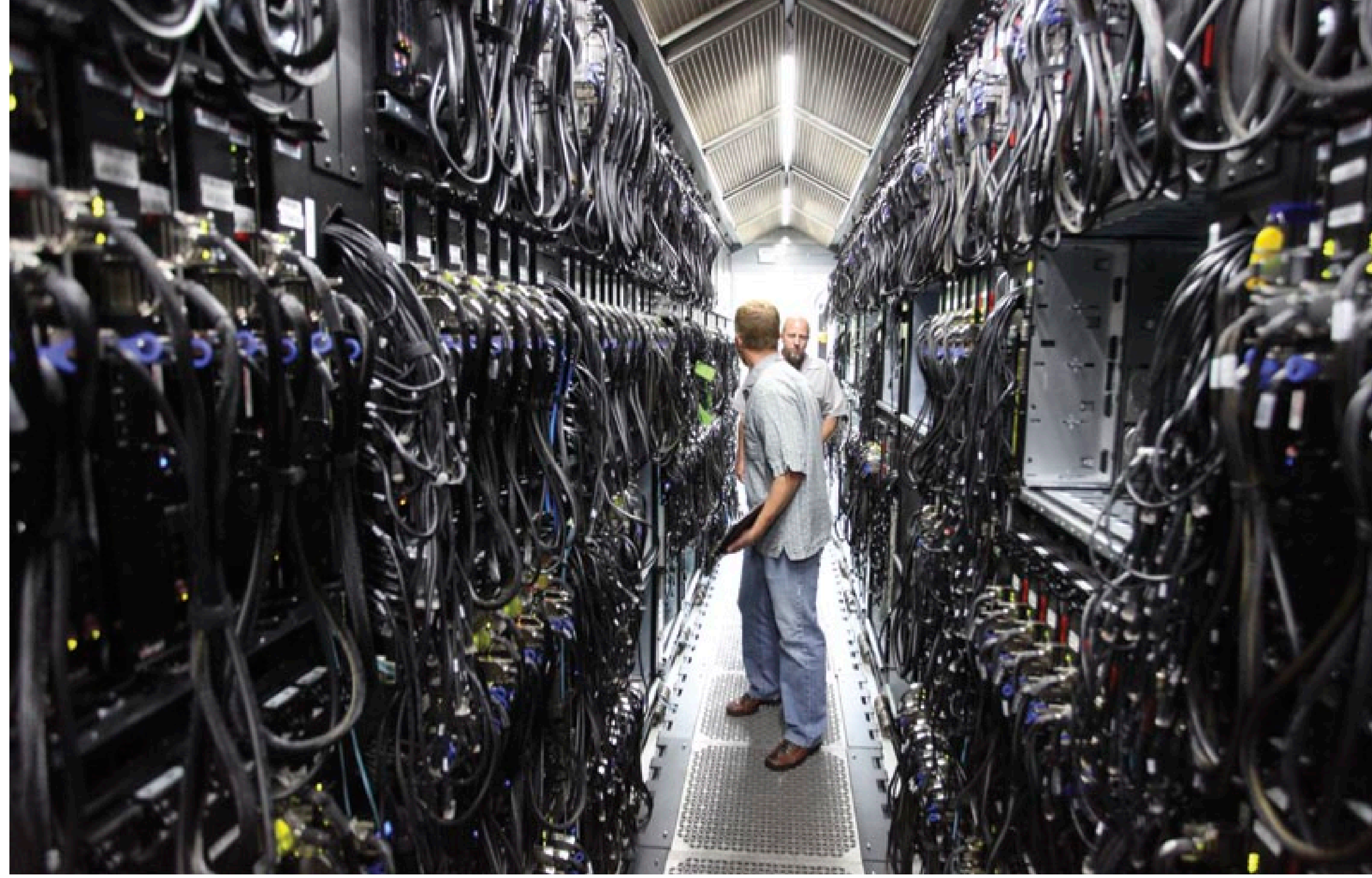
The internet – now with a geographic dimension

BY CHRISTIAN S. JENSEN

Christian S. Jensen (born 1963), PhD 1991, D. Tech. 2000, Professor of Computer Science, University of Aarhus, Denmark, 2010. Appointment at Aalborg University, Denmark, and research stays at the University of Maryland (USA), University of Arizona (USA) and Google Inc. (USA). Member of the Danish Academy of Technical Sciences and the Royal Danish Academy of Science and Letters, and recipient of several natŠnal and internatŠnal awards.

The amount of data in electronic form is growing exponentially. At the same time, the IT infrastructure, including the internet, which we use every day, is developing at great speed. For example, smartphones are proliferating rapidly whÕe mobÕe bandwidth is increasing all the time. At the other end of the infrastructure, we see data centres springing up everywhere. These are buÕdings with large numbers of processors and hard drives which facÕitate the handling of huge volumes of data as cheaply as possible. This trend is continually creating new challenges and opportunities. Christian S. Jensen received the VÕlum Kann Rasmussen Annual Award for Technical and Scientific Research for his work that includes contributing to the efficient storage of, and searching in, spatŠ-temporal data, which is data referenced by time and place. Part of this work aims at giving the Internet a geographic dimensŠn. According to Christian S. Jensen, the annual award of DKK 2,500,000, wÕl be used to enable further research into foundatŠns of the internet of the future.



Data centres have large numbers of processors and hard drives that enable the handling of huge volumes of data.
Photo: Robert Scoble

Vast volumes of data

The digital universe, or the total volume of electronic data, is currently doubling every 18-24 months. It has been estimated that it encompassed 1.2 zettabytes in 2010. A zettabyte is 1024 exabytes, which is 1024 petabytes, which is 1024 terabytes. A terabyte is equivalent to what can now be stored on the single platter of a hard disk. In other words, it took 1.2 bÕlŠn hard disks to store the digital universe as it existed in 2010. In 2020, the digital universe is expected to swell to 35 zettabytes.

It is estimated that there are about 250 mÕlŠn web servers and even more websites on the internet. The number of documents on the internet exceeds 25 bÕlŠn. Moreover, it is estimated that the Google search engine alone receives about 3 bÕlŠn queries a day. These enormous volumes of data make for exciting challenges and new opportunities.

"Google" with a geographic dimensŠn

Before long, the internet wÕl be used more from mobÕe devices than from statŠnary computers. At the same time, it is increasingly possible to positŠn mobÕe devices. It is also possible to attach a geographic locatŠn to many websites (such as a restaurant's website). Studies show that about 20 per cent of all web queries are for results that are

Smartphones are proliferating rapidly. Photo: Cheon Fong Liew

geographically close to the user and thus have “local intent.”

This makes it relevant to add a geographical dimensŠn to “Google queries.” A normal query consists of keywords entered by the user. In response to the query, a list of links to web pages matching the search words is returned. Google’s goal is to respond within 200 mÕliseconds. An important questŠn then is how to also simultaneously take the positŠns of the users and web pages into account.

Indexing

One type of geographic query finds links to web pages that both match the keywords and are close to the user’s locatŠn. It is a bad strategy to check every web page for every query. Instead, it is desirable to buÕd indexes that make it possible to quickly disregard web pages that either do not match the keywords or are located far away from the user.

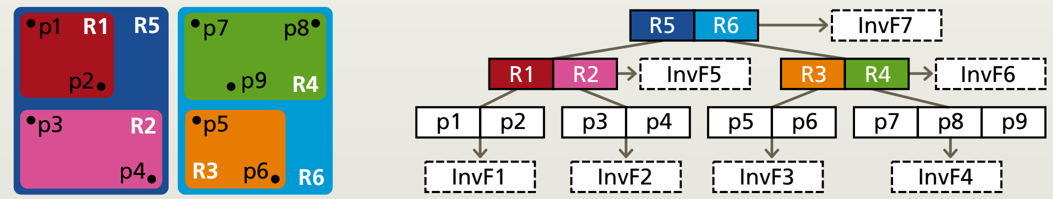
It has long been known how to create indexes that render the finding of web pages based on keywords efficient. These are the kinds of indexes used in search engines. As something new, we have developed indexes that simultaneously take into account the positŠns of the users and web pages. These new indexes make it possible to find relevant web pages by only looking through very little data, which enables short response times.

MobÕe objects

Another challenge arises because mobÕe users are continuously on the move. Imagine that each of Facebook’s 500 mÕlŠn active users wants to see a list of their 10 closest Facebook friends. Or that tourists want to look up the 10 closest points of interest (cafés, pharmacies, etc.) that best meet their current needs.

The challenge here is to keep all lists updated as the users move, and to do this as efficiently and cheaply as possible. One strategy to solve the problem for the tourists works by first finding the 10 currently best points of interest. This can be done using the indexes described above. Then a safe zone around the tourist is calculated within which the current result does not change. When the tourist moves outside the zone, the tourist’s smartphone sends a message to the data centre where a new result and a new safe zone

word	p1	p2	p3	p4	p5	p6	p7	p8	p9
a	0.3		0.4		0.1		0.2		
b		0.5		0.6		0.9		0.8	0.7
c	0.7	0.5	0.6	0.4	0.9	0.1	0.8	0.2	0.3



word	InvF1	InvF2	InvF3	InvF4	InvF5	InvF6	InvF7
a	(p1,0.3)	(p3,0.4)	(p5,0.1)	(p7,0.2)	(R1,0.3) (R2,0.4)	(R3,0.1) (R4,0.2)	(R5,0.4) (R6,0.2)
b	(p2,0.5)	(p4,0.6)	(p6,0.9)	(p8,0.8) (p9,0.7)	(R1,0.5) (R2,0.6)	(R3,0.9) (R4,0.8)	(R5,0.6) (R6,0.9)
c	(p1,0.7) (p2,0.5)	(p3,0.6) (p4,0.4)	(p5,0.9) (p6,0.1)	(p7,0.8) (p8,0.2) (p9,0.3)	(R1,0.7) (R2,0.6)	(R3,0.9) (R4,0.8)	(R5,0.7) (R6,0.9)

Use of the IR-tree for the indexing of nine web pages (p.1 to p.9) with locations. The tree structure captures the spatial containment hierarchy shown on the left. The simplified text associated with the nine pages is shown at the top, and the bottom table shows the contents of the so-called inverted files associated with the nodes in the tree

are calculated and sent to the tourist. It turns out that safe zones can be described by multiplicatively weighted Voronoi cells.

Privacy

The technological advances that enable the services described here also have a downside in terms of access to increased surveÕlance and disclosure of private informatŠn. According to law professor Eva Smith, 82,000 pieces of informatŠn were registered about each Dane in 2008, corresponding to 225 pieces of informatŠn per day. The concept of locatŠn privacy includes aspects such as not wishing to disclose your exact locatŠn to a third party, but also not wishing to reveal that you are geographically close to another person or that you are not home.

Research on privacy shows that it is often possible to achieve support for privacy. For example, one can find out where the “closest pizzeria” is without revealing one’s exact positŠn – one can just ask for all pizzeria locatŠns in the whole of Denmark and sort the results on the phone. The challenge is to return results at the lowest possible cost to the system. One promising strategy would, in this example, be to send pizzeria locatŠn queries at an increasing distance from a nearby false locatŠn untÕ one is sure that one has received enough informatŠn to be able to provide the correct result for the correct locatŠn that only the phone knows.

ApplicatŠns

In additŠn to making a wide range of locatŠn-based internet services possible, research in spatŠ-temporal data management also has applicatŠns in a number of other areas, including intelligent transportatŠn systems, logistics, physical planning, marketing and epidemŠlogy.