What's Wrong with High-Dimensional Similarity Search?

Stephen Blott
School of Computing
Dublin City University
Dublin, Ireland
sblott@computing.dcu.ie

Roger Weber
Credit Suisse
Zurich
Switzerland
roger.weber@credit-suisse.com

ABSTRACT

Similarity search in high-dimensional vector spaces has been the subject of substantial research, motivated in part by the need to provide query support for images and other complex data types. The paper VLDB 1998 paper "Quantitative Analysis and Performance Study for Similarity-Search Methods in High-Dimensional Spaces" analyses why this search problem can be so tricky, and shows with intuitive yet formal proofs that nearest-neighbour search is fundamentally linear beyond a certain dimensionality. Consequently, the paper proposes a new, linear search structure (the VA-File) which focuses on accelerating the indispensable sequential scan with approximations and computational schemes to reduce both CPU and IO efforts. Experiments with both synthetic and image data showed -- surprisingly, at the time -- that such schemes outperform hierarchical methods in all cases where the dimensionality is greater than five. In this paper, we review that work and identify both what we got right in the paper and its impact, and also (with the benefit of hindsight) those elements of the work for which we were off the mark. The lessons learned are relevant not just to the narrow area of similarity search, but also more broadly across the fields of databases and computing.

BIOGRAPHY OF THE SPEAKER

Dr Stephen Blott has been a Senior Lecturer at the School of Computing at Dublin City University in Dublin, Ireland, since 2002, and currently serves as Head of School. He obtained his doctorate from the University of Glasgow in 1992 with a thesis entitled "An approach to overloading with polymorphism" (supervisor, Prof Phil Wadler). From 1993 until 1996 he was a Senior Research Assistant ("Oberassistent") with the Database Research Group led by Prof Hans-Jorg Schek at ETH in Zurich, Switzerland. During that period, his research focussed on technologies to allow the functionality of advanced applications such as GIS or image databases to be embedded within database management systems and exploited for the purposes of storage and retrieval. From 1997 until 2002, Dr Blott was a Member of Technical Staff at Bell Labs Research in Murray Hill, New Jersey, and a member of the Information Systems Research Center led by Prof Avi Silberschatz. His research focussed on embedding database management systems and advanced applications (such as billing for voice and IP services) within communication networks. His current research interests include network data management, payment for network services and digital content, stream data management and identity management.

CITATION OF THE VLDB 1998 PAPER

Roger Weber, Hans-Jörg Schek, Stephen Blott: A Quantitative Analysis and Performance Study for Similarity-Search Methods in High-Dimensional Spaces. VLDB 1998: 194-205

PVLDB '08, August 23-28, 2008, Auckland, New Zealand
Copyright 2008 VLDB Endowment, ACM 978-1-60558-305-1/08/08