Report on the 18th British National Conference On Databases (BNCOD)

Carole Goble
Department of Computer Science
University of Manchester
Oxford Road
Manchester, M13 9PL, UK
carole@cs.man.ac.uk

Brian Read
Information Technology Department
CLRC Rutherford Appleton Laboratory
Chilton,
Didcot OX11 0QX,
Brian.Read@rl.ac.uk

Introduction

The annual series of the British National Conference on Databases has been a forum for UK database practitioners and a focus for database research since 1981. In recent years, interest in this conference series has extended well beyond the UK.BNCOD 2001, the 18th conference in the series, was held at the CLRC Rutherford Appleton Laboratory (RAL) from 9th -11th July 2001. RAL hosts national large-scale facilities for advanced scientific research. The Information Technology Department collaborates with the Laboratory's data centres that manage terabytes of data in remote sensing, high-energy physics and astronomy.

BNCOD 2001 included scientific papers, invited talks, a panel and a poster session. The BNCOD Programme Committee, chaired by Professor Carole Goble of Manchester University, selected for presentation at the meeting eleven papers, about one third of those papers submitted. Contributors were drawn from the Netherlands, Germany, Sweden, Canada and USA, as well as the UK. The audience of 60 attendees was chiefly drawn from the UK database community. The Proceedings are published by Springer-Verlag in the Lecture Notes in Computer Science series, and are available online at: http://link.springer.de/link/service/series/0558/tocs/t2097.htm.

Keynote addresses

The expanding growth of Information Technology continues to place fresh demands on the management of data. Database researchers must respond to new challenges, particularly to the opportunities offered by the Internet for access to distributed, semi-structured and multimedia data sources. Two specially invited speakers addressed two subjects of topical interest that are currently dominating distributed information management research and development activities, namely the Semantic Web and e-Science Grids.

Dr. Rudi Studer, Professor in the Institute for Applied Computer Science and Formal Description Methods (AIFB) at the University of Karlsruhe, Germany, addressed the Semantic Web. The Semantic Web aims to make information accessible to human and software agents on a semantic basis. The paper in the proceedings discusses the role that semantic structures, based on ontologies, play in establishing communication between different agents. Prof. Studer's presentation made a broader sweep of the area. He began with an argument for the Semantic Web, followed by a case for the important role played by ontologies, and their consequences. He specifically addressed appropriate languages for describing ontologies on the web, explaining the limitations of XML and RDF and referring to the DAML+OIL language adopted by the DARPA Agent Markup Language programme and its role in the forthcoming W3C Semantic Web Activity. He went on to discuss issues such as ontology learning, mining, merging and inference mechanisms such as F-Logic. The ontologies, or fragments of ontologies, are used as a vocabulary by metadata descriptions attached to resources. Prof. Studer went on to outline the challenges in representing, creating and managing metadata and presented the Ont-O-Mat tool developed at Karlesruhe that attempts to automatically mine web pages for their metadata annotations. He concluded with a description of the SEAL semantic portal, which has been used to develop the AIFB web site. In answer to a question, Prof. Studer made it clear that he did not believe that their would be single ontologies developed for specific domains, but rather there would be many ontologies and work would be needed to manage this heterogeneity.

The massive increase in data volumes from big science such as remote sensing, high-energy physics and high throughput biological data capture means that we now contemplate the storage

and processing of petabytes. Grid technology, specifically the "Data Grid" is seen as attractive. The second invited speaker addressed strategy in this field. Dr. Tony Hey was until recently Professor in the Department of Electronics and Computer Science at the University of Southampton, but has now taken a position of Director of the UK e-Science Programme. In this role he discussed at the conference the ambitious vision of a Grid infrastructure to support access to global data and processing resources. The UK government has recently invested £120 million pounds in e-Science.

Prof. Hey opened with a definition of e-Science, Grid and the relationship between the two. In the future, e-Science will refer to the large-scale science that will increasingly be carried out through distributed global collaborations enabled by the Internet. Typically, a feature of such collaborative scientific enterprises is that they will require access to very large data collections, very large scale computing resources and high performance visualisation back to the individual user scientists. Hey claimed that Web gave us access to information on Web pages written in html anywhere on the Internet, but now a much more powerful infrastructure is needed to support e-Science. Besides information stored in Web pages, scientists will need easy access to expensive remote facilities, to computing resources - either as dedicated Teraflop computers or cheap collections of PCs - and to information stored in dedicated databases.

The Grid is an architecture proposed to bring all these issues together and make a reality of such a vision for e-Science. Ian Foster and Carl Kesselman, define the Grid as an enabler for Virtual Organisations: 'An infrastructure that enables flexible, secure, coordinated resource sharing among dynamic collections of individuals, institutions and resources.' Resource in this context includes computational systems and data storage and specialized experimental facilities. Prof. Hey gave a number of examples of e-Science applications drawn from Life Sciences, Physics, Environmental Science and Astronomy. He went to describe a three-layered abstraction for the Grid – a lower tier of data/computational connectivity; a middle tier of information management and an upper tier of knowledge-based discovery services. The central role of metadata and middleware was clear. Prof. Hey made the case that the US Grid efforts had delivered in part the bottom layer through Globus and other technologies, but that Europe had much to offer in the upper tiers. In addition, the information/knowledge tiers closely relate to Semantic Web activities pursued by industry and W3C. Hey also sketched the e-Science funding and delivery landscape in the UK, and the range of activities in the US and the rest of Europe. He concluded with a look at industrial perspectives on the Grid.

Both talks clearly indicated a major role for database researchers and practitioners in both the Semantic Web and e-Science, and were also a rallying call for our greater participation and engagement in these activities. Tony particularly made the point that much of the Grid activity had been done in ignorance of best computing practice and without database expertise.

Contributed papers

The contributed papers were presented in four groups: performance and optimisation; objects in databases and software engineering; query optimisation; and querying objects.

Performance and Optimisation

This issue has always been at the core of database technology. Regan, from IBM T J Hopkins, reported on a practical study of space management in logs. They evaluate a technique for reclaiming log space from short transactions while retaining recoverability for long running ones. Zhu and Lü (South Bank University, UK) proposed an algorithm for an effective storage placement strategy for XML documents that facilitates their efficient parallel processing. The trade-off between data quality and performance, was presented by Caine (University of Cardiff, UK). They presented algorithms for integrity checking delayed from when the system is too busy to off-peak, "lights out" hours.

Objects in databases and software engineering.

The great variety of CASE tools prompt the adoption of standardised meta-models and transfer formats. Gustavsson (University of Skovde, Sweden) proposed an extension to OCL further the interchange of models by defining a common, model independent notation for design transformations. Next, Ritter (University of Kaiserslautern, Germany) presented an investigation of the state of database support for software development using object-oriented programming languages. Their work highlights the

shortcomings, in this respect, of the current object-relational database paradigm, and suggests how it might beneficially be enhanced. The third talk returned to the engineering design environment and tackles concurrent version control. Al-Khudair (Cardiff University, UK) presented a generalised object-oriented model that captures the evolution of design configurations and their components by supporting versioning at all levels.

Query optimisation

In this session, contributors considered efficient querying in the newer domains of multimedia and distributed data sources. The requirements and techniques of the worlds of information retrieval and transactional databases are very different. The Dutch team of Blok, de Vries, Blanken and Apers (University of Twente) presented a case study on the "top-N" queries familiar in content retrieval in the context of a database approach to the management of multimedia data. The key issues addressed, such as speed and quality of answers and the opportunities for scalability, are supported by experimental results. A similar problem was addressed by Sattler (University of Magdeburg, Germany), whose work seeks control over the potentially excessive data returned from a query over heterogeneous data sources. By extensions to multi-database languages, they explore ways of asking for just the "first n" results, or of asking for a sample of the complete result. Still with the theme of information systems relying on database technology, Kersten, from CWI in The Netherlands, was concerned with a web multimedia portal based on the Monet database system, although little was specific to the web and memory aware caching would be a better description. Here the optimisation challenge is query throughput. He reported on the performance of a simple and robust scheme for the scheduling of queries in a large, parallel, shared-nothing database cluster and entertained us hugely with his analogies with dinosaurs.

Querying objects

The two presentations that closed the conference were both about querying objects. However, they were very different. Trigoni from the University of Cambridge, UK, presented an inference algorithm for OQL that identifies the most general type of a query in the absence of schema type information. This is relevant to where heterogeneity is encountered - for example, in any open, distributed, or even semi-structured, database environment. Distributed databases and virtual reality are combined in the ambitious work reported by Zaiane (University of Alberta, Canada). They explore data mining in a virtual data warehouse CAVE. Rendering multi-dimensional data aggregates as objects, the user flies through the data to explore and query different views. Zaiane pointed out that the user had great fun too, though results on the effectiveness of the approach are yet to be confirmed.

Posters & Demonstrations

Nine posters with demonstrations, one from Fame, a data mining company, were given a session where each gave a five minute advertisement. The poster session then commenced in earnest with a buffet lunch accompaniment. The posters, whose short papers were supplied in a separate booklet, covered Data mining, parallel query optimisation, multi-database consistency management, active rules and multimedia. Three were concerned with e-Science and Grid activities: the CLRC data portal, the management of time in the EU funded Data Grid project and a gene expression database developed as part of the Mouse Atlas programme. In particular the Data Grid project highlighted the desirability of database workers involvement in Grid projects in order to avoid hasty reinventions of the wheel by the Science communities.

Panel

A lively panel session, chaired by Prof. Alex Poulovassilis, addressed a particular issue that has challenged Hugh Darwen, a database specialist for IBM, and an active member of the SQL standard activity. Hugh is a long time collaborator with Chris Date, with whom he recently published a book entitled “The Third Manifesto”. In this book he makes a case for a different model of type inheritance, which he calls “agreeable”. Hugh presented his ideas with an on-going example based on ellipses and circles. His main argument was that a model to be agreeable had to reflect perceived reality, and with respect to inheritance as perceived by Maier and Zdonik’s desiderata (substitutability, mutability, static type checking and specialisation by constraint) only the final one is valuable. His presentation was then criticised by two panellists. Prof. Mike Jackson made the point that Hugh was confusing inheritance type systems as desired by programming environments (code reuse, a mechanism for reducing the inefficiency of strong typing) was different to that proposed. In fact Hugh was probably thinking about isa

relationships between collection classes rather than type inheritance. Prof. Jackson concluded that Hugh's model was a hybrid between a database view that reflects reality and a programmers view; and that specialisation by constraint as the sole mechanism for forming inheritance hierarchies was not enough. Dr Werner Nutt went on to draw the distinction between inheritance hierarchies and view hierarchies. He pointed out that grouping and classifying objects into classes based on their properties is an area well understood in knowledge management, and is the core tenet of description logic reasoning, for example, where it is called subsumption reasoning. Subsumption reasoning has been used for many years as a query containment mechanism, and forms the heart of ontology languages such as DAML+OIL. He also pointed out various complexity issues regarding Hugh's proposal. Dr Werner neatly drew the distinction between inheritance hierarchies where the subclass relationship is a pre-condition for class properties and view hierarchies where the subclass relationship is a consequence of class properties. He suggested that Hugh's hierarchies are not inheritance hierarchies, and he should not be confused by the tree-like abstraction. A lively debate ensued. Each panellist produced a position paper, published with the poster short papers.

Social Programme

The important social programme (a core activity in BNCOD conferences) was centred on Wadham College, Oxford, founded in 1611. Oracle sponsored the pre-banquet drink reception. At the banquet we practised the art of passing the port and juggling the five glasses required when dining in Oxford. Most of the party took part in a conducted tour of Oxford University, where highlights include Wren's first building, the library where Charles Dobson (Lewis Carroll) worked and the room where Bill Clinton didn't inhale.

Conclusion

The same issues dominated the database research agenda as ever, although often dressed in Web clothes. However, the main take home message of the meeting was the requirement for the database community to engage with the Semantic Web and the e-Science community. Our skills and track record are valuable and yet under-exploited. Moreover, there is a distinct danger of some rather poor solutions being adopted that we might have to live with for many years hence.