Interesting KDD-News form SIGMOD'99

Daniel A. Keim
Alexander Hinneburg

The SIGMOD conference organized by the ACM Special Interest Group on Management of Data is one of the major conferences in the database area. Since databases play an important role in knowledge discovery, the SIGMOD conference is also an important forum for researchers in the KDD area, especially with respect to aspects of KDD which deal with very large data sets. At this years SIGMOD conference, topics of interest related to KDD were classical KDD topics: decision trees, association rules, and clustering. The focus of the papers presented at SIGMOD was on the efficiency for very large data sets (e.g., BOAT – Optimistic Decision Tree Construction) and the effectiveness and user feedback of the mining process (e.g., *OPTICS – Ordering Points to Identify the Clustering Structure* or *Online Association Rule Mining*). It is interesting that – compared to previous years – the *number* of KDD-related papers (not their quality!) decreased. This is probably an indication that KDD has become an area of its own and that SIGKDD has been accepted by the community as *the* major conference for KDD related issues!

Following a brief enumeration of KDD-related presentations at this years SIGMOD conference:

Association Rules & Decision Trees

- Christian Hidber: *"Online Association Rule Mining"*
- Laks V.S. Lakshmanan, Raymond T. Ng, Jiawei Han, Alex Pang: *"Optimization of Constrained Frequent Set Queries with 2-variable Constraints"*
- Johannes Gehrke, Venkatesh Ganti, Raghu Ramakrishnan, Wei-Yin Loh: *"BOAT-Optimistic Decision Tree Construction"*

Clustering

- Mihael Ankerst, Markus M. Breunig, Hans-Peter Kriegel, Jörg Sander: *"OPTICS: Ordering Points To Identify the Clustering Structure"*
- Charu C. Aggarwal, Cecilia Procopiuc, Joel L. Wolf, Philip S. Yu, Jong Soo Park: *"Fast Algorithms for Projected Clustering"*
- Alexander Hinneburg, Daniel A. Keim, *"Clustering Techniques For Large Data Sets: From the Past to the Future"* (Tutorial)

Let us now briefly discuss some KDD highlights from this years SIGMOD conference. In the area of **association rules**, an online association rule discovery algorithm called CARMA has been proposed. Main contributions are the continued user feedback and user interaction as well as the memory efficiency and accuracy of CARMA. In the area of **decision trees**, a new algorithm for a bottom-up construction of decision trees called BOAT has been proposed. BOAT allows an efficient construction of decision trees from very large data sets.

An other area of interest was **clustering**. The tutorial on "Clustering Techniques For Large Data Sets: From the Past to the Future" (a similar tutorial has been given at SIGKDD'99) gave an interesting overview of the existing clustering algorithms. Starting with the well-known k-means, linkage-based, and KDE-based approaches the tutorial focused on the techniques to improve the effectiveness and efficiency in clustering large high-dimensional data sets. A paper focusing on the effectiveness of the clustering results is the OPTICS paper. The basic idea of OPTICS is to find a linear ordering of the points such that a visual representation of the point distances allow an identification of the clusters. The idea of the second paper on clustering was to allow clusters to be defined in a subspace of the high-dimensional space. This new approach seems to be highly practically relevant and has some interesting potential to overcome some of the problems in clustering high-dimensional data.

An other KDD-related topic at SIGMOD was **similarity search**. Papers in this category dealt for example with image similarity, similarity of geometric data, and similarity of market basket data. The papers were also interesting with respect to KDD since they proposed new similarity measures which are also useful for KDD applications. A final hot topic at SIGMOD 99 were the database aspects of the **Web and XML** (extended markup language). Important topics in this area included storing and querying large collections of semistructured (XML) documents. A panel discussed the specific issues and problems occurring in mining the web - a topic which will certainly be also of high interest in the future.