Active Disks for Data Mining

Alexander Thomasian
New Jersey Institute of Technology, USA

INTRODUCTION

Active disks allow the downloading of certain types of processing from the host computer onto disks, more specifically the disk controller, which has access to a limited amount of local memory serving as the disk cache. Data read from a disk may be sent directly to the host computer as usual or processed locally at the disk. Only filtered information is uploaded to the host computer. In this article, We am interested in downloading data mining and database applications to the disk controller.

This article is organized as follows. A section on background information is followed by a discussion of data-mining operations and hardware technology trends. The more influential active disk projects are reviewed next. Future trends and conclusions appear last.

BACKGROUND

Data mining has been defined as the use of algorithms to extract information and patterns as part of Knowledge Discovery in Databases –(KDD) (Dunham, 2003). Certain aspects of data mining introduced a decade ago were computationally challenging for the computer systems at that time. This was partially due to the high cost of computing and the uncertainty associated with the value of information extracted by data mining. Data mining has become more viable economically with the advent of cheap computing power based on UNIX, Linux, and Windows operating systems.

The past decade has shown a tremendous increase in the processing power of computer systems, which has made possible the previously impossible. On the other hand, with higher disk transfer rates, a single processor is required to process the incoming data from one disk, but the number of disks associated with a multiprocessor server usually exceeds the number of processors, that is, the data transfer rate is not matched by the processing power of the host computer.

Computer systems have been classified into shared-everything systems (multiple processors sharing the main memory and several disks), shared-nothing systems (multiple computers with connectivity via an interconnection network), and shared-disk systems (several computers sharing a set of disks). Parallel data mining is appropriate for large-scale data mining (Zaki, 1999) and the active disk paradigm can be considered as a low-cost scheme for parallel data mining.

In active disks the host computer partially or fully downloads an application, such as data mining, onto the microprocessors serving as the disk controllers. These microprocessors have less computing power than those associated with servers, but because servers tend to have a large number of disks, the raw computing power associated with the disk controllers may easily exceed the (raw) computing power of the server. Computing power is estimated as the sum of the MIPS (millions of instructions per second) ratings of the microprocessors. The computing power associated with the disks comes in the form of a shared-nothing system, with connectivity limited to indirect communication via the host. Amdahl's law on the efficiency of parallel processing is applicable both to multiprocessors and disk controllers: If a fraction F of the processing can be carried out in parallel with a degree of parallelism P, then the effective degree of parallelism is $1/(1–F+F/P)$ (Hennessy & Patterson, 2003). There is a higher degree of interprocessor communication cost for disk controllers.

A concise specification of active disks is as follows:

A number of important I/O intensive applications can take advantage of computational power available directly at storage devices to improve their overall performance, more effectively balance their consumption of system wide resources, and provide functionality that would not be otherwise available (Riedel, 1999).

Active disks should be helpful from the following viewpoint. The size of the database and the processing requirements for decision support systems –(DSS) is growing rapidly. This is attributable to the length of the history, the level of the detail being saved, and the increased number of users and queries (Keeton, Patterson, & Hellerstein, 1998). The first two factors contribute to the capacity requirement, while all three factors contribute to the processing requirement.

With the advent of relational databases (in the mid-1970s) perceived actual inefficiencies associated with

the processing of (SQL—structured query language) queries, as compared to low-level data manipulation languages (DMLs) for hierarchical and network databases, led to numerous proposals for database machines (Stanley & Su, 1983). The following categorization is given: (a) intelligent secondary storage devices, proposed to speed up text retrieval but later modified to handle relational algebra operators, (b) database filters to accelerate table scan, such as content-addressable file store (CAFS) from International Computers Limited (ICL), (c) associative memory systems, which retrieve data by content, and (d) database computers, which are mainly multicomputers.

Intelligent secondary storage devices can be further classified (Riedel, 1999): (a) processor per track –(PPT), (b) processor per head –(PPH), and (c) processor per disk –(PPD). Given that modern disks have thousands of tracks, the first solution is out of the question. The second solution may require the R/W heads to be aligned simultaneously to access all the tracks on a cylinder, which is not feasible. NCR's Teradata DBC/1012 database machine (1985) is a multicomputer PPD system.

To summarize, according to the active disk paradigm, the host computer offloads the processing of data-warehousing and data-mining operators onto the embedded microprocessor controller in the disk drive. There is usually a cache associated with each disk drive, which is used to hold prefetched data but can be also used as a small memory, as mentioned previously.

MAIN THRUST

Data mining, which requires high data access bandwidths and is computationally intensive, is used to illustrate active disk applications.

Data-Mining Applications

The three main areas of data mining are (a) classification, (b) clustering, and (c) association rule mining (Dunham, 2003). A brief review is given of the methods discussed in this article.

Classification assigns items to appropriate classes by using the attributes of each item. When regression is used for this purpose, the input values are the item attributes, and the output is its class. The k-nearest-neighbor (k-NN) method uses a training set, and a new item is placed in the set, whose entries appear most among the k-NNs of the target item.

K-NN queries are also used in similarity search, for example, content based image retrieval –(CBIR). Objects (images) are represented by feature vectors in the areas of object color, texture, and so forth. Similarity of objects is determined by the distance of their feature vectors. k-NN queries find the k objects with the smallest (squared) Euclidean distances.

Clustering methods group items, but unlike classification, the groups are not predefined. A distance measure, such as the Euclidean distance between feature vectors of the objects, is used to form the clusters. The agglomerative clustering algorithm is a hierarchical algorithm, which starts with as many clusters as there are data items. Agglomerative clustering tends to be expensive.

Non-hierarchical or partitional algorithms compute the clusters more efficiently. The popular k-means clustering method is in the family of squared-error clustering algorithms, which can be implemented as follows: (1) Designate k randomly selected points from n points as the centroids of the clusters; (2) assign a point to the cluster, whose centroid is closest to it, based on Euclidean or some other distance measure; (3) recompute the centroids for all clusters based on the items assigned to them; (4) repeat steps (2 through 3) with the new centroids until there is no change in point membership. One measure of the quality of clustering is the sum of squared distances (SSD) of points in each cluster with respect to its centroid. The algorithm may be applied several times and the results of the iteration with the smallest SSD selected. Clustering of large disk-resident datasets is a challenging problem (Dunham, 2003).

Association rule mining –(ARM) considers market-basket or shopping-cart data, that is, the items purchased on a particular visit to the supermarket. ARM first determines the frequent sets, which have to meet a certain support level. For example, s% support for two items A and B, such as bread and butter, implies that they appear together in s percent of transactions. Another measure is the confidence level, which is the ratio of the support for the set intersection of A and B divided by the support for A by itself. If bread and butter appear together in most market-basket transactions, then there is high confidence that customers who buy bread also buy butter. On the other hand, this is meaningful only if a significant fraction of customers bought bread, that is, the support level is high. Multiple passes over the data are required to find all association rules (with a lower bound for the support) when the number of objects is large.

Algorithms to reduce the cost of ARM include sampling, partitioning (the argument for why this works is that a frequent set of items must be frequent in at least one partition), and parallel processing (Zaki, 1999).

Hardware Technology Trends

A computer system, which may be a server, a workstation, or a PC, has three components most affecting its