Knowledge Discovery and Data Mining Using an Electro-Optical Data Warehouse

P. Bruce Berra
PBB Systems and Wright State University, Box 246, Keene, NY 12942

Pericles A. Mitkas,
Electrical and Computer Engineering, Colorado State University, Fort Collins, CO 80523-1373

Raymond A. Liuzzi,
Air Force Research Laboratory/IFTB, 525 Brooks Road, Rome, NY 13441

Lorraine M. Duvall
Ramsey Ridge Enterprises, Box 246, Keene, NY 12942

Abstract
In this research we postulated an Electro-Optical Computer Architecture (EOCA) that could be used to evaluate the potential for increased performance and functionality of knowledge discovery and data mining systems that deal with very large multimedia data/knowledge bases. The postulated EOCA is composed of a number of individual holographic associative processors that could perform operations in parallel and could house terabytes of data. With regard to text and numeric data mining, we concentrated on association rules and a number of their variations since many of their operations can be common to other data mining techniques such as classification and clustering. We described these techniques mathematically as timing equations. Utilizing these equations as well as the equations that described the EOCA, we assessed the feasibility of implementing such data mining techniques on the electro-optical architecture. We concluded that great potential exists for orders of magnitude speedup in the data mining of very large text and numeric databases. In fact, some of our results indicate that the association rules algorithm can be evaluated in a matter of a few seconds for a terabyte database. In addition, we investigated the feasibility of the execution of image data mining on the postulated architecture. The results were comparable to those discussed above and therefore quite encouraging. While great potential exists, further research and development is required.

Introduction
In recent years considerable demand has developed for user oriented distributed multimedia management information systems that are able to manage terabytes of data. These systems must provide rich and extended functionality so that new, complex, and interesting applications can be addressed. The need for these systems exists in a multitude of fields including medicine, education and training, defense, business, manufacturing, arts and entertainment, space, as well as a number of other important areas. These applications place considerable importance on the management of diverse data types including text, images, audio and video. As these systems have developed, a wealth of data, information and knowledge has become resident within these vast repositories. This has given rise to a variety of new techniques that have as their objective the extraction of knowledge and information from these repositories [THU97].

Knowledge Discovery and Data Mining (KDDM) is the iterative process of efficiently and effectively finding patterns in data which are relevant to end users. The KDDM process incorporates many methods, tools and techniques from multiple fields to produce effective and usable results, ranging from machine learning techniques from the artificial intelligence field to visualization methods from the human computer interaction field to data warehousing techniques from the database world to provide multi-dimensional data analysis. Data mining is the major computational part of the process that provides algorithms for finding these patterns. There are a number of approaches to data mining including association rules, general characteristics and summaries, classification, clustering, temporal or spatial temporal and path traversal patterns [CHE96].

Optics may be able to help solve some of the very large multimedia data/knowledge base problems. Photons, which have some very attractive properties, such as high speed, non-interference, and a high degree of inherent parallelism can advantageously replace electrons in some processing operations. Optical systems can accommodate a large number of parallel, high-bandwidth channels, thus providing solutions to various interconnection problems. In addition, optical storage devices have very high storage densities and considerable research and development activities are underway to develop devices with read rates in the hundreds of megabytes per second range [MIT98a].

In the research reported here we postulated an Electro-Optical Computer Architecture (EOCA) that could be used to evaluate the potential for increased performance and functionality of knowledge discovery and data mining systems that
deal with very large multimedia data/knowledge bases. The postulated EOCA is composed of a number of individual holographic associative processors that could perform operations in parallel and could house terabytes of data. This system was used to assess the feasibility of implementing such data mining techniques on the electro-optical architecture and to obtain order of magnitude performance data.

Optical Storage, Interconnection and Processing

The state of the art of electronic computing enjoys considerable maturity. In contrast, optics as applied to digital computing is very young and has yet to make its mark. One of the objectives of digital optics is to replace electrons with photons whenever appropriate in a computing environment. As discussed above, the motivation for this is that optics possesses some very attractive properties including massive parallelism, high speed, low power consumption and noninterference of light beams [BER89, 90, GUI96].

In terms of storage, optical disks of various types are in wide use because of their large storage densities even though their access times are slower than magnetic disks. However, with suitable modification to read multiple tracks simultaneously, data rates on the order of hundreds of Mbytes/s are possible [PSA90]. Since electronic computers are designed to deal with magnetic disk transfer rates, they will have difficulty with these increased rates. This dictates that we keep the data in optical form and process them to the fullest extent possible so that, on conversion to electronics, the data rate will be within the capabilities of the electronic computer but more content rich. In this way we hope to increase the performance of the system without disturbing the large investment in systems and user software.

The continuing interest in optical memories is well justified by the potential for high-density storage and for parallel access to two-dimensional pages of data. Optical memories can store as much as 8 terabits/cm², (i.e., approximately 931 GBytes of information). Using wavelength domain multiplexing this figure can be increased by 2-3 orders of magnitude.

Volume Holographic Memories

Most ultra large databases and knowledgebase systems used in knowledge discovery and data mining store data on magnetic or optical disks and employ indexing techniques to avoid or minimize disk accesses. Various clustering and accessing techniques are used to reduce response time. Even so, when the joint requirements of ultra large databases and very short response times are imposed, existing technologies degrade rapidly. In these cases, the ability to call forth and operate on large pages of data in parallel from a page-oriented holographic memory (POHM) would offer a profound advantage over serial operation. The basic concept of page-oriented holographic memory is quite simple. Many small spatially discrete holograms are recorded on a single substrate in a page format that can hold millions of bits per page. They are constructed in such a way that whenever a laser beam illuminates one of these small holograms, the data are read out in parallel in two dimensions. Volume holographic memories can store hundreds of thousands of these pages in photorefractive crystals using a combination of spatial, angular, peristrophic or wavelength multiplexing techniques [HON95, PSA95, PSA98]. An electrooptic or acoustooptic deflector can be used to address any of these stored pages within microseconds.

Since volume holographic memories have large storage capacities they are prime candidates for the storage of large amounts of data and information including multimedia as well as relational databases. Because of their associative nature [MIT94] they are well suited for accessing data at high speeds. The associative mode provides the ability to search the entire contents of the memory by presenting a search argument and receiving the location of the matching elements.

It is safe to assume that optical memories and especially holographic memories represent a promising solution for applications requiring high volume storage, such as: knowledge discovery, relational databases, image processing and in general, a number of research issues currently under consideration in the multimedia field. These applications typically require a high degree of parallelism for processing data. Most of the data operations required by these applications are single-instruction, multiple-data (SIMD) operations. Thus, optical memories and parallel computing have a common characteristic, namely parallelism.

In most conventional computer architectures the processing elements are separated from the data store. Usually a storage hierarchy is employed to move the desired data up the hierarchy to ultimate use by the processor. However, in data intensive processes fast memory is generally not available in abundant supply and large data transfer overhead is incurred. In order to mitigate these effects the processor in memory model offers considerable advantage. In this case processors are integrated with the memory and operations are performed in situ with results being the only data transferred out of memory. While this model is very desirable, it has not been fully realized primarily because of the high cost.
involved. Examples of systems that move in the direction of this model generally move multiple processors closer to the memory and employ some form of parallel processing. They do not, however, actually integrate processing capabilities with memory. In the case of holographic memory at least part of the desirable attributes of the processor in memory model are realized. That is, the memory tends to be very large which is very desirable for large data/knowledge base applications. In addition, the associative processing capabilities allow for some processing of data in memory; namely searching for data that match given search arguments exactly or, in some cases, finding the best match of images.

In order for holographic memory to completely meet the requirements of the processor in memory model considerable additional capability must be added so that arithmetic as well as logical operations can be performed. However, an intermediate system with a broad range of search capabilities would find wide application in the data/knowledge base field. And even with just the exact match capability many applications can be enhanced. For instance, many complex queries have exact match components that, with some query optimization, can be performed first thus reducing the size of the data/knowledge base needed for further processing. It is certainly true that one can construct queries that are void of exact match components, but the vast majority of queries do have one or more exact match components. And in the case of knowledge discovery and data mining many of the algorithms can be enhanced through the use of count data.

Significant advances in the field of page-oriented holographic memories have taken place over the last five years and several prototypes have been demonstrated. Companies such as IBM, Lucent Technologies, Rockwell, and others have pursued the technology, even though Universities continue to play a crucial role in new developments and innovations.

The team at IBM Almaden is heading the NSIC/DARPA/University/Industry Photorefractive Information Storage Materials (PRISM) and Holographic Data Storage Systems (HDSS) consortium. During the past five years a large variety of materials and system configurations have been tested in a specially designed holographic memory tester [BUR98]. Up to 10,000 data pages have been stored in a volume of 1 cm³. At resolutions of up to 1,000 x 1,000 (1 Mbit) per page, the total storage density reaches a significant 10 Gbits/cm³. A system that will employ spatial multiplexing may raise this capacity 50-100 times (with some increase in volume). Even more impressive are the data rates that have been demonstrated: 1 Tbit/sec burst and 100 Gbits/sec sustained. For 1 Mbit pages, the frame rate, that includes the (non-mechanical) access time, must range between 100 kHz and 1 MHz. At these rates, the detector array that receives the holographic memory output becomes the bottleneck. Charge-coupled devices (CCD) designed for display applications are a totally inadequate interface. Schaffer and Mitkas at Colorado State University have explored the use of CMOS smart photodetector arrays that can combine light detection and conversion with some preprocessing, such as demodulation, error control, and even some form of data selection [SCH97]. A prototype chip was fabricated capable of performing parallel error detection and correction of 2x2 cluster errors at frame rates of 5 MHz [SCH98]. A full size chip should be able to output corrected data at up to 100 Gbits/sec. Other research teams have considered and implemented CMOS arrays of active pixel sensors.

The media used most frequently include photorefractive crystals (iron-doped lithium niobate, barium titanate, stoichiometric lithium niobate, etc.) or photopolymers. Crystals can be used in a volumetric form while both crystals and polymers can be arranged on a disk form. Companies such as Holoplex, Rockwell, Optitek, and Lucent Technologies have all demonstrated working prototypes at small form-factors (down to a 3x4x5" black box).

Recording data holographically is invariably slower than reading them. In fact, writing cycles may be several times longer than readout cycles depending on the material and the available optical power.

The main advantage of holographic memories, that is, their ability to perform associative searches, has not been fully explored as yet. We know that associative recall with analog data works nicely and that recent experiments have demonstrated good associative recall when binary and other digital data are used. It is not known, however, to what extent, in terms of total capacity and search argument size, holographic associative processing is effective and reliable. In this work we have taken some small positive steps in the direction of showing that holographic associative processing can be effective.

**Volume Holographic Database System**

A computer-controlled angular-multiplexing photorefractive-based volume holographic memory has been used to store database records, search through the records, and recall the information stored in the memory [GOE96]. Figure 1 depicts the Volume Holographic Database System (VHDS) that was used in the experiments. To record information
we load the data into the spatial light modulator (SLM), create a unique reference angle through the reference beam generation arm, and then open shutters SH1 and SH2. After a predetermined time the shutters are closed, at which point the interference pattern of the two beams has been successfully recorded in the photorefractive crystal. This process is repeated until all the information has been stored. To recall pages we generate a unique reference angle, open shutter SH1, and then capture the data on the camera CCD1. This process can be repeated as needed. The most important aspect of this system is its ability to search every record stored in the memory in a single step; the associative property. To perform searching, we must first have multiple pages of data stored within the memory. With the data in place, we load the SLM with a search argument, open shutter SH2, and capture an image of the reference beam plane on CCD2. Using this image we can determine the angular “address” of the desired information. The search argument that is presented to the VHDS can range in size from an entire page of data to just a small section of a page. This gives us the ability to search for a very specific record, or to search for multiple records that contain similar information.

In this work up to 800 pages were successfully recorded in one cm³ of Fe:LiNbO³ with each page comprising one record of a relation with data fields containing: last name, first name, affiliation, address, city, zip code, and telephone number. Records ranged in length from 98 to 210 characters. These characters were modulated to a binary format using a 2-out-of-15 encoding scheme and a multiblock row and column parity code. Tests were successfully performed on both modes of operation; addressed recall and associative recall. To test addressed recall the VHDS was presented with angles that corresponded to specific pages and then the output of the memory at CCD1 was examined to determine if the correct image was indeed recovered. The results showed that the 800 holograms were successfully recorded and that any page could be reconstructed.

In testing the associative recall they explored how both the search argument and the data stored in the memory affect reconstruction of reference beam planes [MIT98b]. How the number of characters in the search argument, the number of matches, the position, the orientation, and size of the search argument affect recall were also examined. It was determined that when the number of characters in the search argument decreased, the intensity of the correct hit dropped thereby setting a lower bound on the number of characters that are required in the search argument. However, this lower bound is well within the operational limits of the system. It was also shown that it is possible to find multiple pages containing similar data.

**Electro-Optical Computer Architecture**

Since we are interested in data mining applications, which are heavily based on content-based searches, a system similar to the VHDS forms the basic building block of the proposed Electro-Optical Computer Architecture (EOCA). We call this block a holographic associative processor (HAP) since it is an improved VHDS. The EOCA employs many HAP blocks arranged in groups. Each group will store related data (i.e., relations of the same database, images of the same collection, or video sequences). Certain HAP blocks are reserved for storing and searching index files for faster data access and more efficient data manipulation. Different data types can be stored in the pages of the same recording. For example, pages of binary alphanumeric data can be interleaved with pages of digitally encoded imagery or gray-scale images.

The need for data modulation and error coding to ensure industry acceptable corrected bit error rates (<10⁻¹⁴) will reduce the user capacity of the system. A 1 Mbit page with a 40% overhead for modulation and error control will be able to accommodate roughly 75,000 ASCII characters. This number can be contrasted with typical page sizes in electronic systems of .5, 1 and 2 Kbytes. Thus, with 75 Kbytes/page a variety of combinations can be accommodated from all tuples of the same relation to interesting mixes of various types of data.

In our analysis of the potential of EOCA, we select parameter values from the ranges given below. Other parameters are defined as needed.
Structure of the EOCA

Shown in Figure 2 is an overall block diagram of an Electro-Optical Computer Architecture. This architecture serves as the basis of our evaluation of the potential performance.

**Electro-Optical Computer Architecture**

![Figure 2. Electro-Optical Computer Architecture](image)

and functionality improvement that such a system can bring to the knowledge discovery and data mining environment. The optical system consists of many HAP blocks. These blocks are connected together in order to form an ultra large multimedia data warehouse that can house terabytes of data. In this section we characterize the system in terms of memory sizes, bandwidths, speeds, scalability, degree of parallelism, etc.

**Holographic Memory System (HMS) Response Time**

For the holographic system considered here the total storage capacity per module is determined from the following equation:

\[ S_{total} = N_{bits/char} \times N_{char/page} \times N_{pages} \times N_{SL} \]

where \( N_{bits/char} \) is the number of bits that it takes to represent a character, \( N_{char/page} \) is the number of characters that can be placed on a page, \( N_{page} \) is the number of pages that can be placed in a spatial location, and \( N_{SL} \) is the number of spatial locations. \( S_{total} \) represents the total capacity of the memory. However, the effective capacity is smaller since it will require more than 8 bits to store a character (byte) of data. Since we are most interested in very large data/knowledge bases we will assume a large system. Thus, if we assume 10,000 pages per spatial location, two spatial locations, 1000 HAP’s operating in parallel and one megabit/page, we will have a 20-terabit system. Allowing bits (40%) for parity and error correction and converting to bytes we would have a 1.5-terabyte capacity system. As with any system, design tradeoffs are required. For instance, in the case of increased spatial locations, we would be able to have fewer storage elements but search times would be increased.

Knowing that most operations in a database environment involve the retrieval of a record or group of records per request, it is more useful to discuss the response time of the system than the data rate, which is a commonly used performance metric. We define the response time here as the time between the point a request for data is made and the point when the desired information becomes available. This is directly affected by the system components, the type of data access (addressed or associative), and the possibility of having to reread a page of information due to double errors.

The response times of the system components are defined as \( T_{shutter} \), \( T_{angle} \), \( T_{SLM} \), \( T_{CCD} \), and \( T_{decode} \) for the shutter, generation of the angularly-encoded reference beam, SLM, CCD detector array, and decoder, respectively. \( T_{CCD} \) is the total response time of the CCD array which includes both the integration time (the time over which optical power is integrated on the array) and the time to read all pixels from the detector.

Address-based retrieval is performed by generating the reference beam (i.e. deflecting to the desired angle), illuminating the crystal, and then detecting and decoding the output. Thus, the addressed retrieval response time, \( T_{Addr} \), is

\[ T_{Addr} = T_{angle} + T_{shutter} + T_{CCD} + T_{decode} \]

A fast deflector (such as an acoustooptic device) can be set in only a few microseconds and decoding can be done in a parallel fashion within microseconds. The shutter, SLM, and CCD, however, have response times on the order of milliseconds. Thus, \( T_{angle} \) and \( T_{decode} \) can be eliminated from the equation and the equation for \( T_{Addr} \) is approximated by
\[ T_{Addr} = T_{shutter} + T_{CCD} \]

For associative retrieval, the search argument must first be generated on the SLM, the output reference beams must be detected, and finally each matching page retrieved by address. Thus, \( T_{Assoc} \), the associative retrieval response time, is the time to load the SLM with the search argument plus the time to detect the location(s) of the matching page(s) plus the time to retrieve and process those pages. Again ignoring, \( T_{angle} \) and \( T_{decode} \) we have

\[ T_{Assoc} = 2T_{shutter} + T_{SLM} + T_{CCD} + kN_{Rec} (T_{CCD} + T_{post}) \]

where \( k \) is the selectivity factor equal to the percentage of records which match the selection criterion (\( k \leq 1 \)), \( N_{Rec} \) is the total number of records in the database, and \( T_{post} \) is the time required to do any necessary post-retrieval processing to determine an exact match with the search argument.

This analysis is valid only for purely angularly multiplexed systems. If spatial multiplexing is also employed to increase capacity, the address based retrieval response time does not change, but the associative retrieval response time is directly affected. For spatio-angularly multiplexed systems, the search process must be carried out for each of \( N_{SL} \) locations. Thus, \( T_{Assoc} \) becomes:

\[ T_{Assoc} = 2T_{shutter} + T_{SLM} + N_{SL}T_{CCD} + kN_{Rec} (T_{CCD} + T_{post}) \]

where we have assumed that the response time of the deflector used to direct the search argument to the next location is on the order of \( T_{angle} \) and have neglected it in the third term.

The majority of retrievals in a database environment are content-based, so we are primarily interested in \( T_{Assoc} \). It is important to note that the search time in the HAP does not vary with the number of search criteria, unlike electronic database machines. That is, a search for the name 'Smith' and a search for both the name 'Smith' and the zip-code '68405' are performed equally fast since all records and attributes are searched simultaneously.

In order to generate some insight into the capabilities of the HAP we assume some values for the terms in \( T_{Assoc} \). In the following calculations we will assume that \( T_{shutter} = 3 \) msec, \( T_{SLM} = 3 \) msec, \( T_{CCD} = 1 \) msec and \( T_{post} = 1 \) msec. In the case of performing any complex query for a count of the number of hits as described above \( T_{Assoc} = 11 \) msec. It is important to note that what is retrieved at this point are hologram locations that represent the pages that contain the search argument(s). The number of hits will yield the number of qualifying pages. In the association rules data mining technique, the algorithm can be executed by just counting the number of hits. We expect that this approach will yield two to five orders of magnitude reduction in time.

If we then desire the pages, we can estimate the time to retrieve them from the HAP by selecting a value for the selectivity factor \( k \) and knowing the number of records in the system. If we assume that the number of records is one per page then there are 20,000 records per HAP. With a selectivity factor of \( k = 0.01 \) then \( T_{Assoc} = 411 \) msec. With 75,000 characters per page this is an effective transfer rate of 36 megabytes per second. Standard magnetic disks have transfer rates on the order of five megabytes per second. It is important to note that with improved optical components the read out rate of the HAP can be increased considerably.

**Electro-Optical Computer Architecture Response Time**

The main strength of the EOCA is associative access. That is, we can search all pages in memory for responders to an arbitrarily complex query and determine page positions in one scan of the memory. Thus, we can search a terabyte database in a matter of milliseconds. From the mirror angles we can obtain the number of responses to the query and with multiple scans of the EOCA we can obtain all of the data we need to execute the association rules algorithm. With the EOCA the potential exists to render the time to execute the association rules algorithm negligible. From the peaks in the reference beam profile we can determine the pages in memory that have produced hits and they can be read out if needed or they can be accessed from secondary storage on the sequential front end computer and further processing performed.

Referring to Figure 2 note that all HAP units operate in parallel. Thus, for an arbitrarily complex query the electronic computer would broadcast the search argument to all HAP’s via the system bus. They would execute in parallel and collect the responding hologram position data at each HAP. The count could be determined at each HAP with a local processor or the hologram positions could be transferred to the electronic computer for determination of the count. In executing association rules, a local processor could collect the results of many passes and do some preprocessing prior to sending the results to the electronic computer.

The timing equation for executing a single search, \( T_5 \), on the EOCA for a complex query is composed of a) a query broadcast time, \( T_b \), b) a search, \( T_{Assoc} \), without readout (the first three terms), and c) a collection of the hologram positions from the
HAP’s and their transfer to the electronic computer for further processing $T_{trans}$. Thus,

$$T_S = T_B + T_{Assoc} + T_{trans}$$

$T_B$ takes a few microseconds and $T_{Assoc}$, based on previous calculations, is 11 msec. $T_{trans}$ will depend upon how many HAP’s have registered hits. However, suppose they all do. To transfer the hologram positions from a single HAP would require a few microseconds. Since there are 1,000 HAP’s in the EOCA we would expect the transfer time to take a few msec. Thus, the entire process would only take order of milliseconds to complete. For association rules, depending upon the number of queries to the EOCA that would be required, the algorithm could be executed in a matter of seconds.

### Database of Transactions

<table>
<thead>
<tr>
<th>Trans</th>
<th>$\Delta$</th>
<th>$\beta$</th>
<th>$\gamma$</th>
<th>$\delta$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>XX</td>
<td>X</td>
<td>X</td>
<td>D</td>
</tr>
<tr>
<td>2</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>XX</td>
<td>X</td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>XX</td>
<td>X</td>
<td>ABA</td>
<td>BD</td>
</tr>
<tr>
<td>5</td>
<td>X</td>
<td>X</td>
<td>6</td>
<td>4</td>
</tr>
<tr>
<td>6</td>
<td>XX</td>
<td>X</td>
<td></td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>XX</td>
<td>X</td>
<td>ABD</td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>XX</td>
<td>X</td>
<td>3</td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>XX</td>
<td>The significant relationships are:</td>
<td></td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>XX</td>
<td>X</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Figure 3. Association Rule Example

Shown in Figure 3 is an example database of transactions that is used to illustrate the capabilities of the EOCA in solving the association rules problem. There are ten transactions that have $\Delta$, $\beta$, $\gamma$, and $\delta$ as possible values. In relational database parlance we have a single relation with the transaction number as primary key and the presence or absence of the values $\Delta$, $\beta$, $\gamma$, and $\delta$ in the four domains. One can view this in a commercial application as the fact that the customer purchased $\Delta$, $\beta$, and $\gamma$ in transaction 1, another customer purchased $\beta$ and $\delta$ in transaction 2, and so on. In mining for association rules we would like to know the strength of the relationship between and among the items purchased in all of the transactions.

We first set the level of support or strength of relationship that we are interested in. Here we choose 50%. That is, if the percentage of transactions that include an item is 50% or greater, then we look further for associations between and among all of those items. In this case we see that $\Delta$, $\beta$, and $\gamma$ meet our criteria. We now look for associations between products and find that $\Delta$ and $\beta$ qualify, but $\Delta$ and $\gamma$ does not. Finally, the relationship $\Delta\beta\gamma$ does not qualify.

In executing this algorithm using sequential computing, the database would have to be accessed many times or multiple indexes would have to be established depending upon the approach taken to solving the problem. Using the EOCA, the timing equation given below would determine the time to produce all of the necessary count data and then it would be a simple matter to determine all possible associations.

$$T_{AR} = T_B + k \sum_{i=1}^{n} \left( \binom{n}{i} (T_{assoc} + T_{trans}) \right) T_{Calc}$$

In this equation $k$ is the number of tuples per page since we will have to perform multiple searches if we have more than one tuple per page; $n$ is the number of domains in the transactions (four in the above example), while the sum of combinations gives all possible combinations of the domain values ($\Delta$, $\beta$, $\gamma$, $\delta$, $\Delta\beta$, $\Delta\gamma$, $\beta\gamma$, $\Delta\beta\gamma$), $T_{assoc}$ is as before and $T_{trans}$ is the transfer time from each HAP to the sequential computer under the assumption that results are transferred after each search of the EOCA. If the results are all collected first and then transferred this term would be larger but outside the parenthesis yielding a smaller value overall. However, the calculation of the associations $T_{Calc}$ would be impacted since this operation could not commence until all the data in the EOCA were collected and transferred. In the above equation it is assumed that the transfer of the partial results will be provided to the sequential computer for processing as they become available and the transfer time and calculation time can be overlapped. Thus, the time required for $T_{Calc}$ is just the time to process the results from a single interrogation of the EOCA.

If we assume that there are four domains, 1.5 terabytes in the EOCA, 10 tuples per page and $T_{Calc}$ is 10 msec, then $T_{trans}$ is 10 msec. $T_{AR}$ for this example is about 3.6 seconds. Although not a valid comparison, just to transfer 1.5 terabytes of data from magnetic disks would take days.

Thus, it is clear that the use of the associative property in the EOCA has great potential for speeding up association rule processing. However, we must still bear in mind that holographic memories are not yet widely available, they take a long time to load, and, of course, have other problems that must be solved before they can become a main stream computer system reality. But, nonetheless, great potential exists which clearly warrants continued investigation.
Similar difficulties arise with clustering, so additional research needs to be performed to more completely measure the effectiveness of the HAP in executing these data mining algorithms.

References


Acknowledgement

This research was partially supported by AFRL/IFTB SBIR Contract No. F30602-98-C-0076 entitled “Dynamic Data Mining Using an Electro-Optical Data Warehouse.”