CSC 2508F
Spring, 2002

October 16, 2002

Due dates:

Proposal	October 30th
Presentations	November 25th to December 2nd (in class)
Written Version	December 18, 2002

CSC2508 -- Information: Quantification, Representation and Manipulation

Course Project

OVERVIEW

Pick a topic of your choice that is related to the representation and processing of information, and submit a report of approximately 10 to 15 pages discussing it. (A one-page project proposal should be submitted indicating what topic you have chosen.) While a variety of alternative topics are suggested below, originality is encouraged. If you have doubts about the suitability of a topic you are considering, discuss it with the instructor. If there are papers in the literature that are closely related to the topic you choose, then part of your report should be devoted to briefly summarizing the previous work (at least for the two or three most relevant references), and indicating how it relates to your work. A modest programming effort to support some experimentation may be appropriate for some of the projects.

SUGGESTIONS

Here are brief descriptions of some topics that have been or will be discussed in class. Exploring any one of them in more depth than is permitted by the time available in class would be a suitable project.

(1) Efficiency of Data Compression

Pick a certain class of homogeneous data sources susceptible to compression and try out alternative techniques experimentally. (For example, choose a certain class of data output files, or internal representations of spread sheets, or a set of bit-map image representations.) You could compare some form of Huffman encoding, and/or Arithmetic encoding with the Ziv-Lempel algorithm as provided in Unix. Unix provides two pairs of commands, "compact" / "uncompact" (which are based on Huffman coding) and "compress" / "uncompress" (which are based on Ziv-Lempel). Both of these work on arbitrary files. Pick some class of files (such as C programs, latex or troff input files, postscript files, or spread sheets, etc.) and experimentally measure the degree of data compression achieved by each of these. Then try to estimate how well they do with respect to an estimate of the theoretical information content of the files.

(2) Encoding Shapes

Consider closed curves in 2-space that define the boundaries of (the 2-d projection of) an object. A first order approximation to any shape is probably a square (or possibly a circle). Closer approximations can be obtained with rectangles (or ellipses). At the extreme, the boundary can be represented as a piecewise linear curve joining a large number of boundary points, or may be a smooth curve passing through all the points. Investigate the trade-off between precision of representation and size of representation. In this case the imprecision of representation is probably quantifiable as the sum of the area that is inside the true shape but outside the representation and the area that is inside

the representation but outside the true shape all divided by the total area inside the true shape. Overall, the goal here should be to retain the ability of generating recognizably correct representations of various shapes with as much storage economy as possible.

(3) Encoding Signatures

When banks process cheques, they are supposed to check that the signature on the cheque matches the signature on the signature card created at the time the account is opened. Current electronic processing of cheques completely ignores this requirement. Suggest a method of characterizing signatures so that they can be encoded in some reasonable number of bits and so that they can be used as the basis of distinguishing among the signatures of different people. (You only need to consider trying to catch cases where a truly different signature appears, as opposed to the vastly harder case of trying to detect the work of a forger who is attempting to duplicate the signature of another person.)

(4) Encoding Faces

Great strides have been made in storing fingerprints in a form that is concise and yet allows the fingerprints of a large number of different people to be accurately distinguished. Some related techniques might be used to store representations of peoples' faces. The goal would be to take a few pictures of each person in a group, store representations of the faces, and be able to later "recognize" the individuals when confronted with a new (not seen previously) picture. (And, better yet, to be able to say "I've never seen this person before", if that is indeed the case.) The question is what kind of information should be stored and how much of it is likely to be required (as a function of the number of people in the group to be distinguished).

(5) Encoding Concepts

Some designs for machines that exhibit high degrees of capability for associative retrieval and human-like memory characteristics require that all concepts be encoded as bit strings, preferably of some fixed length (say 1000 bits to allow for 2^{1000} distinguishable concepts). For some very limited domain, consider ways of mapping concepts into such long bit strings. The challenge comes from the fact that concepts that are similar or that are related closely should have "similar" encodings (in the sense of agreeing in a relatively large number of bit positions, like 20%, if possible).

(6) Encoding Pictures

Various hierarchical schemes allow the compact representation of pictures of various types. Investigate the use of these schemes for compact representation of pictures created by simple animation. This would involve trading off the degree of detail in representation among position, size and colour, for example. At the extreme, consider how you might encode art masterpieces in a manner such that similarities among them can be detected.

(7) Quantifying Information

Assess and compare the information holding capacity of various neural network structures. A question of interest is how the network size (number of nodes) and shape (arrangement of nodes in layers) is related to its capacity for learning and retaining some specific identifiable set of information. Some types of networks to consider are symmetric neural networks (Hopfield nets), layered neural nets, and Kanerva's Sparse Distributed Memory.

(8) Encoding Representation of 3-d Structures
In both graphics and robotics, it is important to be able to represent and compare and manipulate representation of 3-d objects and scenes. One approach for doing this is using Superquadrics. How complex an object or scene can be represented in a limited number of bits using this or other techniques?

(9) Encoding Tunes and Melodies and Chord Progressions
People easily recognize familiar songs despite shifts in key, different musical instruments, changes in tempo, variations in the durations of notes, and imprecise pitch (by poor singers). Develop an encoding scheme for songs that would allow a computer to recognize similarities with the same degree of robustness.

(10) Encoding Maps
Maps, such as those provided by the National Geographic Society, are very detailed and contain an immense amount of information. Consider some simpler forms of maps (e.g., maps of the London subway system) or else subsets of the information provided in a complex map, and determine how it could be compactly stored in a computer so that the computer representation could be used by appropriate algorithms to provide answers to the same kinds of questions that can be answered by looking at a map. (E.g., what state is south of Saskatchewan?, or which Great Lakes are accessible from Canada?, or what rivers pass through or by three or more Provinces?)

(11) Encoding and Indexing Video Clips
Consider TV news programs, which are composed of a sequence of segments interleaved with video clips and commercials. How could such programs be abstracted, stored and indexed to permit subsequent queries and retrievals based on the content of the parts of the program?

(12) Encoding Faxes
There are several levels of Fax transmission standards that involve varying degrees of compression. Study and evaluate schemes at the most sophisticated end of the spectrum where the page image is decomposed into a number of parts (letterhead, address, text, tables, figures, etc. all the way to doodles and smudges) and each part is represented and transmitted in the most appropriate way.

(13) Inference in Statistical Databases
In databases, it is desirable to allow statistical queries (about average salaries, etc.) without divulging individual information (i.e., an individual's salary). Take an information theory point of view, and examine how the answers to statistical queries reduce the uncertainty about individual information.

(14) Data Mining for Associations and Causality
Techniques and algorithms for mining association rules from large databases are well-established. Consider how causal relationships can be identified and established based on the extracted association rules.

(15) Indexing Structures for Point Data in High-Dimensional Spaces
There is a growing need for supporting range queries, similarity queries, and nearest-neighbor queries in multidimensional spaces. Survey a few of the indexing structures / strategies that have been proposed recently to address this problem.

(16) Indexing Schemes for Spatial Objects in Multidimensional Spaces
In a k-dimensional space, it is desirable to be able to represent objects of dimensionality

k or less. For example, in models of our 3-dimensional world, we would like to be able to represent objects of 0-dimension (points), 1-dimension (line segments), 2-dimension (surfaces), as well as 3-dimension (e.g., buildings). Survey and/or propose methods for supporting such representations.

(17) Quantization and Clustering
In high-dimensional spaces (say, 100-dim or 1000-dim), the precise representation of a single point can be quite large (e.g., 4KB if each of a 1000 attribute values is represented with 32-bit precision). Concise (but approximate) representations of data sets composed of such points may involve both quantization (representing attribute values with lower precision, say 4 bits instead of 32) and clustering (representing a set of points with a set of possibly quantized attribute values). Investigate how these two techniques interact in attempting to achieve concise approximate characterizations of the structure of a multidimensional data set.

(18) Datacubes and Histograms
Both Datacubes and Multidimensional Histograms have been proposed and the basis for providing quick approximate answers to certain types of queries. With histograms, the elements of the multi-dimensional data set are partitioned into non-overlapping and mutually exhaustive "buckets", and summary information about each bucket is retained. Approximations of the values of any data element in the range of a bucket can then be produced from the summary information for that bucket. With datacubes, each cubelet defines the entire dataset after projection on some subset of the dimensions. A subset of cubelets is retained, but the full data set is not. Approximations for individual data element values are then produced from some combination of cubelets that summarize the element in question. Examine the comparative advantages and disadvantages of each approach. (Note also that datacube cubelets themselves might also be represented by histograms rather than exactly.)

(19) Digital Library Issues
Explore some aspect(s) of the large digital library project (that is based at several prominent U.S. universities), from the viewpoint of information representation as discussed in class. [See the February, 1999 issue of IEEE Computer, which contains several papers that have emerged from the project.]

(20) Self-designed Project
Pick some project that relates both to your own major area of interest and to the content of this course.