The infinite disk: challenges from no limitations

Mark Sanderson - University of Sheffield, m.sanderson@shef.ac.uk

Challenge

Managing and searching across multi-terabyte and potentially multi-petabyte personal stores of multimedia information.

Background

The ubiquitous doubling of processor speed every 18 months governed, by Moore's law is an accepted feature of computing. However, it is not the only such "law". There is one less noticed property of computers that is outstripping the rate of processor speed-up: disk storage is currently doubling in size every year. In 2002, a £1,000 desktop computer comes with 120Gb of storage as standard. If the rate of increase continues at the current pace, a terabyte of store should be easily affordable by 2005, a petabyte (2^{15}) by 2015, and possibly an exabyte (2^{18}) by 2025.

Traditionally, hard disks were never large enough for user's needs: many of their files either had to be moved to external storage or had to be deleted to make way for more recent items. With an annual doubling of capacity, storage growth has started to outstrip the growth in space required by items one might store on a computer. Conventional files, such as email, word processor documents, spreadsheets, source code, etc, grow at a pace of perhaps a few hundred megabytes per year and will occupy progressively smaller fractions of a hard disk.

The passing of this important milestone in computing went relatively unnoticed probably because at the same time, other types of files such as image, audio, and video became more prevalent, which caused new storage problems for most computer users. However, with both images and audio, current hard disks have the capacity to hold the collections of all but the most avid hoarder:

- 60,000 high quality still images (near to 35mm film standard) can be stored in 120Gb; and
- 30,000 hours (3.5 years) of CD quality music would fit on such a disk.

Only video remains as the medium that cannot be held in sufficient quantities. However, even it is starting to be manageable: MPEG-2 TV-quality video compresses to just over 1Gb per hour, a video of an entire lifetime in this format only occupies 1Pb. Quality of media such as the number of mega-pixels per still image may well increase as might the frame rate and definition of video, however, such increases do not keep up with current storage growth.

The grand challenge

If disk growth continues at its current pace, everything a person reads, sees, hears, or watches in their life time will be storable on their desktop PC. Their disk will be unfillable: an infinite disk. An immediate and trivial impact of this may be the demotion or even removal of the trashcan or recycle bin from the desktop, replaced perhaps with a shredder. Such a capacity may well present a series of challenges and opportunities to various computing communities, to the field of information retrieval (IR) however, it presents a grand challenge.

> Managing and making searchable a very large personal multi-media collection.

The challenge requires research from a broad range of computing disciplines reflecting the forms of data being stored such as unstructured and semi-structured text; strongly typed database information; erroneous text from scanned documents or speech recognised audio; music; images; video; other objects and files, such as vector drawings, software, etc.

There are two main aspects to IR: indexing and retrieval. For texts of all kinds, database content, and even music, means of indexing such objects are increasingly well understood; this is in contrast to methods for

managing images, video and other objects. While a long-term challenge might be to determine means of improving on content-based indexing of such files to determine semantic meaning from such objects, it is not clear what the chances of success will be. Solutions that allow searching of these media types via associated text or metadata are needed at least in the short-term and if the long-term challenge fails, will present the only solution.

Associating searchable data with unsearchable media objects to aid in index and retrieval presents challenges across a variety of computing disciplines.

- Users cannot be expected to put much effort into aiding searching systems. Consequently, the interface community will need to study the needs and expectations of user populations: understanding how they wish to organise multimedia as well as understand the lengths users will be willing to go in order to supervise indexing of their collection.
- Devices that record multi-media will need to record more peripheral information and embed it in any generated files: time, grid reference, an audio recording at the time of capture, etc. may all be valuable data for indexing the recording. Quite what metadata is feasible and how one searches effectively over such metadata needs to be better understood. Mapping of metadata to text will also need to be addressed potentially calling upon the database and text processing communities.
- Other forms of peripheral information held on a user's PC may aid in indexing media objects: emails sent or received at the time of capture or pertaining to the location of capture could be examined; equally electronic diary entries, or the web pages accessed by users could hold clues for indexing. To achieve such needs, language processing, data formatting standards, and interoperability of software all need to be addressed.
- The image or video content processing communities will also be required in this task providing means of determining forms of similarity across media objects: perhaps associating objects that have been manually indexed with similar objects that have not.
- With a range of media types each indexed using distinct and possibly multiple methods, the final challenge will be for information retrieval researchers to discover means of unifying the variety of indexing methods to create a single seamless media searching tool. Again, a range of challenges exist.
    - The theoretical models underlying IR systems will have to be adjusted to take equal notice of dispirit indexing components.
    - HCI research will again be called upon to better understand the ways that users wish to access, browse, and search across their personal information collection and how a user might query media objects without having to understand the indexing methods.

Conclusion

The whole basis for this challenge is the assumption that current growth in disk storage will continue for the next ten to fifteen years. While some have forecast hard limits in conventional magnetic storage technologies, previously predicted limitations have rarely proved correct and often alternative technologies have been found. Ultimately, there must be a limit to the amount of data that can be stored in a box the size of a desktop PC, however, what ever that limit is, it would appear that it is orders of magnitude greater than the amount of information a person could possibly observe in their lifetime. While this proposed grand challenge is technology driven, it is a driving force that challenges computer scientists to consider the impact of the coming of an infinite unfillable disk.