Content –based searching of digital data.

Leslie Smith, David Cairns, Department of Computing Science and Mathematics, and Kevin Swingler, INCITE, Department of Psychology, University of Stirling, Stirling FK9 4LA. Email lss@cs.stir.ac.uk

The proposal

Imagine that your computer system had access to a large volume of data in many different formats: textual data, still and moving images, sound recordings, drawings etc. Now imagine that you could issue queries such as "Find me the sound of a baby crying" or "Find me images and architectural drawings of St. Paul's Cathedral". Clearly, if all the images were nicely labelled with meta-data in XML, this would present little challenge. But in reality, images and sounds are not nicely tagged: indeed, it is difficult to see how one might tag such data formats without a great deal of subjectivity. What we propose is the development of content-based searching techniques for digitally stored data. This is the key technology that underlies the capability of fulfilling queries like the above.

Background

We have come to expect to be able to search through huge quantities of text to find information on subjects of interest. We appreciate that this task would be easier if people provided meta-data in some standard form, or if they stored the original data in XML. But we also appreciate that, because of the way in which the data was generated, this is unlikely to happen, so we have developed free text searching tools.

In many ways the situation for non-textual digitally stored data is different: however, it shares the problem of rarely being tagged. Non-textual data takes many forms: some such data is at least partially symbolic (for example, MIDI sound files, or vector graphic files), but others are simply direct digital representations of the underlying analogue data (for example, .wav sound files, or bit-mapped still or moving image files).

How can we provide searching capabilities for these different type of files? We are not proposing a system that looks through the raw data for each query (indeed, free-text searching systems do not do this either). Rather, the data will be analysed and meta-data generated off-line, and only the search of the meta-data will be carried out at query time. Most PCs are far more powerful than required most of the time, and most are unused for many hours each day. Thus even now, there is plenty of available processing power for this task.

What is the nature of the challenge?

The types of file that such a system needs to be able to analyse fall into two categories. In the first category are files whose content is symbolic, and whose alphabet of symbols is the same as (or can be deterministically mapped to) the alphabet of symbols in the query. This is clearly the case for searching text for words or phrases (where the alphabet of symbols in both cases is words in the target language). It is also the case, for example, for searching MIDI files for patterns of musical notes. Even in these cases, however, there can be subtle problems caused by synonyms, both at the single symbol (word or note) level, and at the string of symbols (phrase) level.

The second category is the case where there is no deterministic translation from the query language to the alphabet of symbols in the file. Sound and image files provide clear examples for this case: with sound files, the symbols can be integers between –32768 and + 32767, each representing instantaneous pressure of the sound wave: with pixel-based images, each symbol is the mean colour and intensity of some small area of an image. There is a level mismatch between this and the queries (which are human language based), and which refer to extended periods of sound, or to whole areas of images). The primary challenge is to solve this mismatch.

A considerable amount of research has already been undertaken on many problems which can be seen as sub-problems of the proposed challenge. Into this category come projects such as finding faces in still and moving images, looking for keywords in telephone speech, or finding instantaneous pitches in musical

recordings. There are a number of different technologies which have been applied in these areas, ranging from statistical techniques, pattern matching techniques, neural network techniques for example. What this challenge does is to bring together the basic into an overarching project.

Tagging images is difficult. To assist the automated system, the user could provide examples of particular sounds or images in which they were interested, and provide the system with text referring to these images (such as "the sound of a bassoon" or "a picture of Uncle Ralph"). The tagging procedure could be task oriented: particular tags would be appropriate for particular applications. In addition, we could reasonably expect some user tagging to be provided: where and when an image was taken, or which band was playing on a sound file. There will always be a question of subjectivity: for example, is an image of Da Vinci's "Mona Lisa" a picture of a woman, of a face or of an enigmatic smile? Some aspects of this problem are unlikely ever to be able to be automated!

How does this proposal fit in with the Grand Challenge Criteria?

The proposal explores the limits of Computing (passes criterion 1), attempting to make it address directly the problem of interpreting (in the sense of generating meta-data about) sound and image data. Certainly, this is something that has not been built before (passes 2). The task is incremental, in the sense that it can be done well, or not very well (fails 3). It certainly has support from prospective users, though the research community's reaction remains to be seen (possible 4). There is a sense in which once the concept is visible, there will be a race to complete this: there are clear benefits for the computer using community from success of this project (passes 5). It is certainly comprehensible to the general public, as well as to scientists (passes 6). Its formulation is relatively recent, as it only became of interest once large volumes of data were digitally available (fails 7). Further, no-one knows how to solve the problems discussed in general: there are ideas for techniques which can solve some aspect of the problem, but new tools will need developed (passes 8). In order to deal with the many different forms that digital data can take, planned collaboration will be useful (passes 9). There is a sense in which it does encourage competition: this technology could be the basis for the next generation of "Killer Applications" for powerful PCs, so that there is a commercial prize to be won[1] (passes 10). Even partial success would have commercial benefit (passes 11). The project can be seen as part of a move away from computers as primarily handling human-supplied data, into automated handling of sound, vision and other signal-like data. This could be a part of a paradigm shift in Computing away from the desktop metaphor towards more ubiquitous computing (possible 12). Since there are commercial advantages to be gained, it could be met from commercial advance, but the problems are such that this is currently unlikely (possible 13).

[1] both for commercial and home users: how many home computer users have untagged digital still photographs that they would like to search?