ei

Managing the **enterprise information** network

A picture is worth a thousand words

How new multimedia search engines can help you find everything you're looking for.

Published by

ark GROUP

A picture is worth a thousand words

Not just data. New multimedia search engines will help you find everything you're looking for, including images, audio and sound files.

By Professor Stefan Rüger and Adam Rae

For over 2,500 years humans have kept libraries of information in different forms. From ancient Egyptian archives to modern day computerised distributed databases, data has been gathered and people have found ways of handling this information. Librarians have long used indexes in different forms to help find the resource they are interested in, but search engines and the internet have changed this paradigm. No longer are we restricted to searching solely the data about resources or metadata – we can now search the entirety of the data itself.

The ever-increasing ease in producing and sharing digital content has lead to extensive databases of media ready to be exploited. But present systems tend to rely on rather unintuitive technical metadata to help people search through the collections to find what they want. Searching by esoteric file format information may be fine for graphic artist or computer scientist, but those hoards of everyday internet users would rather search by where images were taken, or by who is in a video clip.

In other words, people want to be able to search the content itself and not just the data associated with it.

The Multimedia & Information Systems group at the the Knowledge Media Institute, The Open University in the UK, focuses on addressing just these issues as well as other areas of information retrieval – from intuitive search engine interface design to high performance text retrieval algorithms, video search engine design to spoken document retrieval. Our core aims include more fully understanding and using information learnt from user interaction and gaining a fuller understanding of a user's information need.

A lot of our work has been based on extracting useful information from still images, video and audio, and using this in conjunction with search algorithms that combine this data to give results to a user that are as relevant as possible.

These features come in many forms: an account of the colours used in an image, to descriptions of textures, shapes and patterns of frames in a video clip. They can also include text comments and tags or keywords that, unlike the content-based features, can be updated and improved during the life-time of the item of media.

With this additional information, a search engine algorithm can interpret the query submitted by a user and find the media that most closely match it. These results are then displayed to the user and their interaction with these results can give useful feedback on how relevant they were. The system can then learn how to adjust its search algorithm and the metadata associated with the media to better perform in future.

Putting it together

Once a system has extracted these features, it needs to determine how best to use them. Judging whether a colour profile should be given a higher importance than text similarity, or what to do if different features give conflicting relevance values are examples of what is called the fusion problem. This can be partly overcome by using an algorithm to derive weightings for each contributing feature that gives greater importance to certain features over others.

This technique works well for fixed data sets of similar images, but doesn't handle diverse collections with many different users all that well.

Another approach is to take into account the feedback from the user. If those images returned from a search engine that have similar texture profiles are found to be most relevant (perhaps by the user inspecting them or explicitly denoting them as relevant) then that feature would have its weighting increased. The system can then tailor its function towards the user and the particular task.

Not all users are the same and have varying information needs. Someone searching the web for map images

The vertical search platform

Specialised search engines for specific applications may include:

- Identifying plants and trees from the shapes of their leaves in digital images. Allows large-scale automated cataloguing of species and ecosystem monitoring which can give insights into biodiversity and the effects of pollution and climate change;
- Searching and browsing though museum collections where metadata may be incomplete, inconsistent or non-existent. By being able to browse by visual similarity, related items that were previously lost amongst the collection can be rediscovered, or new trends in style can be found;
- In retail, online consumers can search for purchases by visual similarity. For example, upon finding the perfect pair of trousers, a matching shirt with the same colour, pattern or style could be automatically selected. By taking a webcam snapshot of oneself, the system could suggest items of clothing or new styles that suit your look and build;
- Searching through webcast recordings of university lectures, both the automatically transcribed audio and support materials like slides, would allow revising students to find the exact excerpt of the course that would help them with their current studies.

of their local area will have different requirements to someone browsing for their friends' recent birthday party images. These requirements lead to the need for an individual profile for a search engine user.

A cartographer might want to give high importance to geotags (latitudinal and longitudinal coordinates stored alongside an image or video of where it was taken) and certain textural criteria associated with clean and clear diagrams whereas a holidaymaker might prefer to give a higher weighting to personal identification tags and images that were taken indoors using a flash.

By treating users individually, a system can build such a profile and use it to return those results most suitable for the user based on his or her prior interaction with the system and what task is being performed. It is this new focus on personalisation and the influence of a user's wider social context that is beginning to affect the systems we use for modern knowledge management.

Someone searching the web for map images of their local area will have different requirements to someone browsing for their friends' recent birthday party images.

Fingerprinting the Rolling Stones

Extracting colour and shape information from digital images is relatively straightforward, but what information can you derive from moving video images, or even a continuous stream of audio? Searching through large collections of audio has normally been done by searching the associated, manually produced metadata. But by extracting a 'fingerprint' of a recording, a signature that can uniquely identify the excerpt of music based on its composition, melody and timbres involved, search engines can find and match recordings without relying on metadata.

Services like Shazam in the UK allow a user to query-by-example – by dialing a number on a mobile phone while listening to a piece of music on the radio or in a bar or club, the system can produce a fingerprint of the music and compare it against its database and respond with the title and artist. The user does not need to describe the music he is interested in, but merely gives an example for the system to find a match.

Birds of a feather

When asked who they turn to for advice on a potential new purchase or opinions on an as yet unexplored holiday destination, most people tend to ask people close to them – their friends, family and acquaintances. This social network surrounding a user is made up of people who might well share similar tastes, interests and characteristics in the way they look for information.

By taking this social context of a user into account, a multimedia search engine can return results that are even more relevant. Taking this information and combining it with the behaviour of the user in past interactions with a system can let us produce a fingerprint not for the media in the system, but for the users themselves. By learning what an individual finds relevant and by learning how a query is put together can let the system better interpret the information need.

Time flies like an arrow

Using the tags that a user's friends have added to their photographs or video clips is indeed useful, but interpreting these for the benefit of a search engine can be tricky. One person may interpret an image of a football match as a joyful, happy scene if the team won, but it may have very different connotations for the supporters of the opposition.

Tags and captions written by end users can be very subjective and a search system has to take this into account. This ambiguity is a major problem in manual annotation of media, which helps explain why there is such a focus in current research on automated annotation. Keywords and phrases are derived directly from the data itself, without the need for a subjective human editor.

While such automatic systems may not be as skilled at interpreting images and audio as humans, their performance continues to improve. On the issue of interpretation, while a video can be analysed for textures, patterns and forms, it is a big leap of interpretation to go from these low level features to more abstract concepts such as 'car' or 'holiday'. This is known as the semantic gap.

The area is currently under intense scrutiny. As media collections that previously could be manually annotated continue to grow, automated solutions are increasingly needed to produce the kind of metadata humans would have created. This metadata is what can then be used in conjunction with search engines to allow users to navigate using keywords and relevant terms and phrases.

To attempt to shrink the gap between low level feature information and high level concept tagging, techniques are being explored that make use of the semantic analysis of tags to make inferences from the metadata of one item of data to see if they could justifiably be transferred to another, very similar item. Statistical classifiers can attempt to analyse the numerical low level feature data and see how they match up to higher level tags in annotated data sets and transfer these connections to other data sets.

Artificial intelligence techniques can be used to deduce new information about an image based on existing data, by extrapolating trends from clusters of similar images and mimicking the function of the human visual system to take advantage of the abilities we have developed over millions of years of evolution.

Good, but not good enough

Determining whether a multimedia search system performs well is measured by how relevant the end user feels the results are, and how the results match up to accuracy and precision metrics. To measure this performance, the system requires feedback, either implicit or explicit.

Asking a user directly whether the result set was what he or she was looking for, whilst giving valuable information, is extra work and can put the user off using the system. By designing the interface intelligently, every mouse click and key press can help indicate which of the returned items of media were ultimately useful and which were not.

The search process does not need to be a one-shot action – refining a result set is an excellent way of getting more useful results. This is known as the 'filter and refine paradigm'. By obtaining a crude initial set and honing it down by selecting additional criteria, the user can elaborate on the original information request in easily manageable steps.

For example, a neurosurgeon could highlight an area on an image of damaged brain tissue and query a database of MRI scans for other images (and therefore cases) of similar damage. He could then refine this subset of the image collection by selecting only images of patients who had similar symptoms described in text metadata. This could be further refined by selecting patients of a similar age and medical background as the query patient. The results should be useful and will have saved the surgeon trawling through the entire database in one go. This idea of iteratively exploring subsets also has connotations for very large scale media databases.

Another issue with handling and searching through such data is that it is very computationally expensive, requiring both time to process and storage space for indexes. Techniques that avoid having to search an entire data collection for the required results in one go tend to perform more quickly and can therefore save the user having to wait long for results.

What you see is what you get

Getting a set of results from a multimedia database is only part of the process. Displaying these results in a way that is efficient, pleasant and informative is also important. These different visualisations of data can be suited to particular tasks, media or users or can be more general.

Diagrams laid out in a tree-like structure, known as Dendro Maps, allow a user to easily navigate through a set of relevant results in a structured hierarchical fashion based on certain criteria, whereas geo-temporal browsing combines real-world spatial mapping and media to localise images, video or audio depend on the place they are associated with.

Vertical search

To date, the majority of systems have been based on horizontal frameworks, in that they attempt to use one single search method to explore a wide range of media – a library uses just one catalogue system to archive literature, books-on-tape, videos and maps.

Greater focus is being given to more vertical approaches, where systems are tailored to the particular tasks, users and media they are going to be used with. A media retrieval system based on searching through personal photographs has different requirement to other systems and can be better designed if it also does not have to be capable of performing well at, for example, medical X-ray image retrieval.

The main aim of the multi-million euro EU PHAROS (Platform for searcHing of Audiovisual Resources across Online Spaces) project is to develop a framework based on this vertical search paradigm, providing a system that will ultimately be able to handle any kind of multimedia data effectively in an innovative, open and distributed manner. As a partner in the project, the Knowledge Media Institute (along with institutions such as Fast Search & Transfer, France Telecom and Fraunhofer IDMT) contributes its expertise in handling content-based search tasks and works together with other groups in academia and industry to produce a system that will help consumers, business and organisations to unlock the potential in their digital media.

Getting what we want

Knowledge and media are important to societies that create and use digital information. Techniques for handling this ever-increasing amount of media are trying to keep up with the demands of producers, users and consumers whilst at the same time taking advantage of new advances in complementary fields like artificial intelligence, hardware design, psychology and statistics.

It is by gathering and intelligently analysing as much media data as possible that search systems will improve and our group at The Knowledge Media Institute is at the forefront of this endeavour. Getting what we want from the increasingly large sea of information is tricky, but we as researchers are working on making it easier, more efficient and more enjoyable. ■

Professor Stefan Rüger <s.rueger@open.ac.uk> and Adam Rae <a.rae@open.ac.uk> are with the Multimedia and Information Systems Group at The Knowledge Media Institute which is part of The Open University, Walton Hall, Milton Keynes.