Automating the Linking of Content and Concept

Robert Tansley†, Colin Bird*, Wendy Hall†, Paul Lewis† and Mark Weal†

†IAM Research Group, University of Southampton, UK
*IBM UK Laboratories, Hursley Park, UK
rht96r@ecs.soton.ac.uk

ABSTRACT

In previous work we have described a multimedia system, MAVIS 2, supporting content and concept based retrieval and navigation. A central component of the system is a multimedia thesaurus in which media content is associated with appropriate concepts in a semantic layer. A major challenge is identifying and constructing these associations in a particular application without requiring a huge amount of manual effort. In this paper we propose a two phase approach to the problem. In the first phase, latent semantic analysis is used to associate metadata available for some media objects with concept class descriptions. This facilitates automatic associations to be made with the concept layer for those media objects. In the second phase, media content matching is used to classify media objects without metadata through their similarity to media objects classified in phase 1.

1. INTRODUCTION

Many multimedia information systems (MMISs) allow searching and navigating of multimedia objects based on associated textual *metadata*. The approach typically relies on the availability of suitably controlled metadata and the availability of people's time to enter it. Increasingly, we see MMISs which support retrieval of multimedia objects based on their content[2]. A small number also support hypermedia navigation based on content[6, 4]. However, it has become clear that, in many cases, the available low-level features of media objects are insufficient to determine whether or not two media objects pertain to the same real-world *concept*.

We have recently reported the development of an MMIS called MAVIS 2, Multimedia Architecture for Video, Image and Sound[1], which encapsulates both the information retrieval and hypermedia information discovery paradigms. It features a semantic layer component as part of a *multimedia thesaurus (MMT)* which is central to the MMIS architecture.

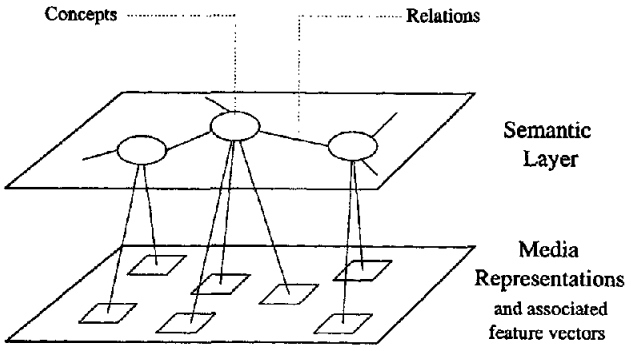

Figure 1: Multimedia Thesaurus Architecture

The semantic layer of the MMT consists of *concepts* connected by *relationships*. Each concept is an abstract entity corresponding to a real-world "object". Each concept is associated with one or more *media representations*, i.e. multimedia objects that represent the concept. These representations may be a text term or phrase, a portion of an image, a segment of video or any other medium. Thus, one concept may have many representations in many different media. This is illustrated in figure 1. Media representations are associated with feature vectors (signatures) extracted from the representation using media processing algorithms. Modules providing the media processing can be integrated in an incremental way as new techniques are developed and IBM's QBIC technology[3] has been incorporated to provide additional signatures and illustrate the extensible nature of the architecture.

The semantic layer architecture can support any number of arbitrary semantic relationships between concepts, but initially we implemented only the two common types used in existing thesauri: a hierarchical specialisation/generalisation relation, and a *related* relation. The *equivalence* relation for media representations is implicit in the architecture. Different media representations of the same concept are considered equivalent and are called *synonyms*, even if they are of different media types. They are linked to the same concept in the semantic layer.

Media representations may also constitute or contain source and destination anchors of hypermedia links. These links may be *generic links*; that is, the link may be followed not

only from the source anchor itself, but also from any other portion of media that matches the source anchor. Matching is achieved using the signatures mentioned above and the distances generated are combined using a normalised ranking system.

The MAVIS 2 system uses the multimedia thesaurus architecture in a variety of ways. Content-based navigation can be enhanced by supplementing the available links with links from *synonyms* or by widening or narrowing the scope through the concept generalisation and specialisation relations. The results of 'find similar object' queries may also be enriched using synonyms of retrieved media or query expansion through generalisation. Additionally, the concept layer provides a useful means of navigation in itself and a concept browser is provided as an additional entry point for navigation or retrieval.

2. CONSTRUCTION PROBLEMS

The usefulness of a semantic layer has been reasonably well established, but a major problem with the development of such systems is how to construct the semantic layer and create the appropriate associations between media representations and the concepts in the semantic layer which they represent, without a substantial amount of manual effort.

The basis of the semantic layer developed for this application is a subset of the Dewey Decimal Classification (DDC) system, a widely used classification system with a broad scope. The DDC is a large set of classes, designed for libraries with potentially millions of volumes to index. We used a suitable subset of the classification relating to the subject domain of the images in our test in order to establish the semantic layer. One of the advantages of using the DDC is that one subset can be 'attached' to another with a wider scope. In this way two subsets (and hence multimedia thesaurus assisted collections) can be merged however much or little the subject areas overlap.

The core of the multimedia collection used in this trial application were a set of 1023 images of artefacts from the Victoria and Albert Museum, London. This image collection was compiled during the first phase of a previous project concerning electronic access to distributed image collections called the Electronic Library Image Service for Europe, or ELISE [7]. A wide range of artefacts are depicted; they include paintings, sculptures, clothing, furniture and textiles. Some text metadata is associated with some of the images but not all. Only a fraction of the images had an appropriate amount of associated metadata, and not all images with metadata have the same fields.

Fortunately, related work in the previously mentioned ELISE project is also of use here. As part of the ELISE project, a selection of DDC classes was chosen for representing the contents of image collections, particularly for museums. Each class was given a number of associated keywords. Largely, these keywords (and indeed the class names) do not appear in the image metadata, so simple text matching cannot be used to tie image metadata to DDC classes.

The subset the ELISE project used was designed with several museum collections in mind, and is thus still rather widely-scoped for the Victoria and Albert collection used in this application. A subset has been chosen with the scope of this collection in mind, and not so deep that it will be sparsely represented. Part of the chosen subset is shown in table 1. The indentation shows the hierarchy.

```
700     The Arts; Fine and Decorative
730       Plastic Arts; Sculpture
736         Carving & Carvings
738         Ceramic Arts
738.2         Porcelain
738.4         Earthenware & Stoneware
739         Art Metalwork
740       Drawing and Decorative Arts
741         Drawing & Drawings
746         Textile Arts
748         Glass
749         Furniture & Accessories
```

Table 1: Dewey Decimal Classification

Each concept is made an abstract entity in the concept layer; the text label (for example, "Glass"), is held as a media representation in the media representation layer. In this way a suitable set of concepts is obtained for the semantic layer and for each concept a set of descriptive keywords was also available. We also have a set of images, of which only some have associated metadata. We now face the challenge of connecting the images to the concepts they represent without creating each association manually.

Our approach involves a two-phase process for associating the images with the appropriate concepts. Firstly, the images for which we have sufficient associated metadata are connected to the appropriate concepts by using Latent Semantic Analysis (LSA) to give a measure of correspondence. Secondly, low-level features are used to classify the remaining images, by using the image-concept associations created in the first stage as a ground truth.

3. PHASE 1: USING LATENT SEMANTIC ANALYSIS

Given two sets of text, the latent semantic analysis (LSA) technique developed at the University of Colorado[5] will return a value indicating how closely related the two pieces of text are, even if the terminology in each piece of text differs. The value is the cosine of the angle between vectors derived to represent the two pieces of text in a particular semantic space. Thus, to establish which concept is most appropriate for a particular image, the metadata associated with the image can be compared with the DDC class keywords to give a measure of correlation.

Initially, it might be assumed that since the Dewey class keywords and image metadata comprise a relatively small corpus of text, the effectiveness of the technique would be poor. However, several "ready-trained" semantic spaces are publicly available, in a variety of subject areas. Reviewing the subject areas revealed that the *encyclopedia* set holds the most relevant information on museum objects and artefacts and is likely to produce the best results.

The LSA Web site also offers on-line access to LSA software. We developed a simple Java based tool to send batches of pairs of text pieces and receive the resulting similarity values. Using the image metadata, concept keywords and the LSA tool, cosines were calculated indicating the degree of similarity between the images and the DDC classes (and hence, the appropriate concepts in the semantic layer). The next problem was how these values would be used to assign images to classes?

Two associators were built. One implements a *knn*-style classifier, which assigns images to concepts based on the highest cosine generated by the LSA process. The second implements a simple decision tree, propagated down to the descendent (narrower concept) with the highest cosine. Some preliminary testing established that the *knn*-style associator produces the best results.

The LSA based classification was performed with 106 images that had associated ELISE metadata. The relevant images were associated with the appropriate concepts in MAVIS 2 using a batch process that sent relevant messages to import the images to the MAVIS system and create the associations with the appropriate concept in the semantic layer. Once the main categorisation has been done, the resulting network was browsed using the concept browser to find and move any images that were obviously out of place. Ninety five of the 106 images were correctly associated with concepts using the automatic LSA technique. A script that quickly allowed the reassigning of an image to another class was used and the process of correctly associating the 11 misclassified images took about 15 minutes.

4. PHASE 2: USING IMAGE FEATURE MATCHING

The remainder of the Victoria and Albert Museum images were classified using only image features. No text metadata associated with either the images or the concepts was used. To facilitate this, another "batch classifier" facility was developed. The process is given a number of media objects to classify and these are used to form CBR queries to the MAVIS 2 system. The corresponding results are used to determine a best-matching concept. Associations are then be made between the media representations and their appropriate concepts. The process is fully automatic.

The batch classification had a high rate of success in identifying *Glassware* and *Furniture & Accessories*. Images depicting paintings were more prone to misclassification. On the whole the content matching stage resulted in less reliable results than the LSA stage and an explanation for this is the great visual variability, particularly of the painting images, and the relatively small number of LSA classified images we used to act as prototypes. The signatures used were essentially, colour, spatial colour and texture measures and the classification would clearly benefit from the use of more pertinent features if they were available. Finally, the classification used a basic 'nearest neighbour' approach and more sophisticated techniques should produce more robust results.

5. CONCLUSIONS AND FUTURE WORK

We have presented a brief overview of an approach to the automatic building of a multimedia thesaurus in a small trial application of our MAVIS 2 multimedia information system. While this procedure was not entirely robust for the reasons cited above, it demonstrates that such an approach is feasible with the semantic layer architecture adopted and accelerates the creation of a the MMT to support both content and concept based retrieval and navigation. In future trials, we plan to use larger numbers of media objects in the LSA stage in order to provide a larger set of prototypes for phase 2 of the classification process. We are also investigating more robust classification techniques and improving the range and pertinence of signatures available for content matching.

Acknowledgements

The authors are grateful to the Victoria and Albert Museum for the use of their image collection, the ELISE project for their metadata and to the EPSRC for support through research grant GR/L03446. The authors also wish to thank David Dupplaw, Dan Joyce and Mark Dobie for useful discussions.

6. REFERENCES

[1] M. Dobie, R. Tansley, D. Joyce, M. Weal, P. Lewis, and W. Hall. A flexible architecture for content and concept based multimedia information exploration. In *Proceedings of the Challenge of Image Retrieval (CIR'99)*, pages 1–12, Newcastle, UK, Feb. 1999.

[2] J. Eakins and M. Graham. Content based image retrieval. Technical Report 39, U.K. JISC Technology Application Programme, Oct. 1999. Available at http://www.jtap.ac.uk/.

[3] M. Flickner, H. Sawhney, W. Niblack, J. Ashley, Q. Huang, B. Dom, M. Gorkani, J. Hafner, D. Lee, D. Petkovic, D. Steele, and P. Yanker. Query by image and video content: The QBIC system. *IEEE Computer*, 28(9):23–32, Sept. 1995.

[4] K. Hirata, S. Mukherjea, Y. Okamura, W.-S. Li, and Y. Hara. Object-based navigation: An intuitive navigation style for content-oriented integration environment. In *Proceedings of ACM Hypertext '97*, pages 75–86, Southampton, UK, Apr. 1997. ACM, ACM Press.

[5] T. K. Landauer, P. W. Foltz, and D. Laham. An introduction to latent semantic analysis. *Discourse Processes*, 25:259–284, 1998.

[6] P. H. Lewis, H. C. Davis, S. R. Griffiths, W. Hall, and R. J. Wilkins. Media-based navigation with generic links. In *ACM Hypertext 96 Proceedings*, pages 215–223, 1996.

[7] A. Seal. The creation of an electronic image bank: Photo-CD at the V&A. *Managing Information*, 1(1):42–44, 1995.