Content-Based Retrieval of Web Pages and Other Hierarchical Objects with Self-organizing Maps

Mats Sjöberg and Jorma Laaksonen

Laboratory of Computer and Information Science*,
Helsinki University of Technology,
P.O.Box 5400, 02015 HUT, Finland
{mats.sjoberg, jorma.laaksonen}@hut.fi
http://www.cis.hut.fi/picsom/

Abstract. We propose a content-based information retrieval (CBIR) method that models known relationships between multimedia objects as a hierarchical tree-structure incorporating additional implicit semantic information. The objects are indexed based on their contents by mapping automatically extracted low-level features to a set of Self-Organized Maps (SOMs). The retrieval result is formed by estimating the relevance of each object by using the SOMs and relevance sharing in the hierarchical object structure. We demonstrate the usefulness of this approach with a small-scale experiment by using our PicSOM CBIR system.

1 Introduction

Large multi-modal databases, with objects of many different domains and formats, are becoming more common. Multimedia databases containing texts, images, videos and sounds require sophisticated search algorithms. Such algorithms should take into account all available semantic information including the actual contents of the database objects as well as their relationships to other objects.

We propose a content-based information retrieval (CBIR) method that models known relationships between objects in a hierarchical parent-child tree structure. We have used the CBIR system PicSOM [1] as a framework for our research and extended it to incorporate hierarchical object relationships. By mapping low-level features of the database objects, such as colour or word-frequency, to a set of Self-Organized Maps (SOMs) [2] we can index the objects based on their contents. The known relationships of objects, such as being a part of another object (eg. image attachment of an e-mail) or appearing near each other (eg. two images in the same web page) are modelled as object trees.

Section 2 presents the hierarchical object concept in more detail, Section 3 reviews the PicSOM CBIR system. Sections 4 and 5 discuss an experiment using a small set of web pages. Finally, conclusions are drawn in Section 6.

---

2 Multi-part Hierarchical Objects

A recent review of image retrieval from the World Wide Web [3] shows that most systems use only information of the images themselves, or in combination with textual data from the enclosing web pages. ImageRover [4] for example combines the visual and textual features into one unified vector. WebMars [5] is the only system in the review that allows multimodal browsing and it uses a hierarchical object model. Other systems include AMORE [6] that uses multiple media types in a single retrieval framework, and the Informedia project [7] that seeks to provide full content search and browsing of video clips by integrating speech, closed captioning and image recognition.

In this work, a *multi-part object* is a hierarchical object structure, organised in a tree-like manner modelling relationships between the objects. A hierarchical object tree can consist of objects of many different types and can, in principle, be of any depth. The trees are usually formed from natural relationships, like the child objects being parts of the parent object in their original context. For example an e-mail message as a parent object can consist of attachments as child objects. Likewise, an image can be parent of its segments.

As an example of the forming of a multi-part object, a web page with link information and embedded images is shown on the left in Fig. 1. The different parts have been enumerated and marked with a red rectangle. On the right we see the multi-part object tree structure created from this web page. The URL of the web page itself and links to other pages, images and other objects, are collected into one common "links" object, while the images and textual content of the web page are stored as objects by themselves.

The relevance of each object in a multi-part tree can be considered to be a property of not only the object itself, but to some extent also of its parents, children and siblings in the tree structure. We call this idea *relevance sharing*, which means that the relevance assessments originally received from user feedback will be transfered from the object to its parents, children and siblings. For example, if an e-mail message is considered relevant in a certain query, its attachments will also get increased relevance values. As a result of the *relevance propagation* performed by the PicSOM system, e-mail messages with similar attachments will then later get a share of that relevance.
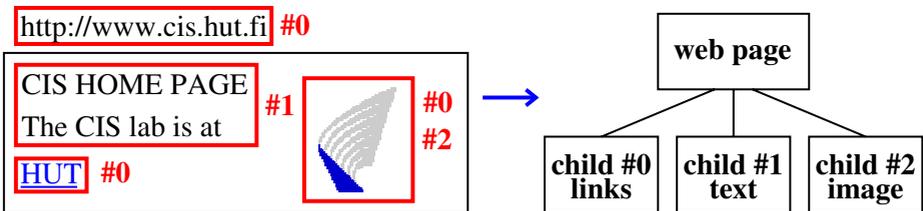
http://www.cis.hut.fi **#0**

CIS HOME PAGE
The CIS lab is at   **#1**   **#0**
                                #2
HUT **#0**

web page

child #0 links   child #1 text   child #2 image

Fig. 1. A web page (left) with its corresponding multi-part object tree (right)

3   PicSOM CBIR System

The content-based information retrieval system PicSOM [1] has been used as a framework for the research described in this paper. PicSOM uses several Self-Organizing Maps (SOMs) [2] in parallel to index and determine the similarity and relevance of database objects for retrieval. These parallel SOMs have been trained with different data sets acquired by using different feature extraction algorithms on the objects in the database. This results in each SOM arranging the objects differently, according to the corresponding feature.

3.1   Relevance Feedback

Query by example (QBE) is the main operating principle in PicSOM, meaning that the user is presented with a set of objects of the desired target type, from which he selects the relevant ones. This *relevance feedback* information [8] is returned to the PicSOM system which expands it from parent objects to children, and from children to parents, and possibly also to siblings, depending on the types of the objects. This relevance sharing stage was added to the baseline PicSOM system for this work to gain an advantage from the dependencies between the objects.

For each object type, all relevant-marked objects in the database of that type get a positive weight inversely proportional to the total number of relevant objects of the given type. Similarly the non-relevant objects get a negative weight inversely proportional to their total number. The grand total of all weights is thus always zero for a specific type of objects. On each SOM, these values are summed into the best-matching units (BMUs) of the objects, which results in sparse value fields on the map surfaces.

After that the value fields on the maps are low-pass filtered or "blurred" to spread the relevance information between neighbouring units. This produces to each map unit a *qualification value*, which is given to all objects that are mapped to that unit (i.e. have it as the BMU). Map areas with a mixed distribution of positive and negative values will even out in the blurring, and get a low average qualification value. Conversely in an area with a high density of mostly positive values, the units will reinforce each other and spread the positive values to their neighbours. This automatically weights the maps according to relevance and coherence with the user's opinion.

The next processing stage is to combine the qualification values gained from each map to the corresponding objects. These values are again summed between parents and children of the object trees. The final stage is to select a specific number of objects of the desired target type with the highest qualification values. These will be shown to the user in the next query round.

4   Data and Features

We collected a set of web pages from the intranet of our institution. This resulted in a database of over 7000 web pages and almost 2900 images. In the hierarchical

model each web page forms a tree with the page itself as parent and the embedded text, images and links as children as illustrated in Fig. 1.

Two ground truth classes containing images as the target type were selected manually. *Tourist* class (907 images, *a priori* 31%) was from a conference or vacation and mainly outdoor tourist-type photography with attractions like monuments and buildings. *Face* images (253, 8.6%) were such that the main target was a human head.

4.1   Visual Features

From the images we extracted MPEG-7 still image descriptors, using the MPEG-7 Experimentation Model (XM) Reference Software [9]. As colour descriptors we used *Colour Layout* (dimension: 12) and *Scalable Colour* (256), as shape descriptor *Region-based Shape* (35), and as texture descriptor *Edge Histogram* (80), all calculated from the entire image area.

4.2   Weblink Feature

In our experiments we used a *weblink* feature calculated from all the URLs related to a web page. Each distinct URL can be regarded as one dimension in a very high-dimensional binary space of all valid URLs. Our weblink feature extraction algorithm is based on the idea initially presented in [10], and uses the *Secure Hash Algorithm* (SHA-1) [11] for performing *random mapping* to combine and to reduce the dimensionality of such vectors. Random mapping replaces an orthogonal base with a new base of lower dimensionality that is almost orthogonal. SHA-1 produces a condensed and nearly unique representation of a text string or message, called a *message digest.*

In the weblink feature extraction algorithm, each of the URLs related to a web page is recursively pruned into shorter URLs: first the original URL, then the web page directory and each higher level directory, and finally the bare domain part. We calculate an SHA-1 message digest for each of these generated URLs and form a 1024-dimensional binary random projection vector for each by looking at the first 32 bits of the digest. These bits are interpreted as four 8-bit indices into separate ranges of a 1024-dimensional vector where the corresponding components are set to unity and the others to zero. These vectors are then weighted, summed and normalised to unit length. Finally, the link feature vector for the web page is given as the sum of the normalised per URL vectors. The dimensionality is fixed at 1024 which is computationally sound.

4.3   Text Feature

We extracted the text from the HTML files and calculated a character *trigram* feature. For each character trigram in the text we calculate an SHA-1 message digest and form a 1024-dimensional vector from the first 32 bits in the same manner as with the weblink feature. The final feature vector is the sum of all these vectors from each trigram of the text document, normalised by their number. The text feature can thus be regarded as a random-projected histogram of character trigrams.

5   Experiments and Results

We trained a total of six SOMs, one for each of the weblink and character trigram features and one for each four MPEG-7 features. Every feature vector was used 100 times to train the corresponding SOM of size 256×256 map units.

The experiments were run in four ways: using only the MPEG-7 image features and combining MPEG-7 with weblink, trigram or both. Each query was initialised with one image of the pre-selected ground truth class. 50 query rounds were performed with 20 returned images in each round, where the relevance of each image could be automatically determined by using the ground truth data. The experiment was repeated so that each ground truth image was used once as the initialiser and the results were then averaged over all experiments.

Fig. 2 shows the recall–precision graphs where the precision has been normalised relative to the *a priori* of the class. In all plots the precision initially increases and then begins to decline when a clear majority of the relevant images has been found. The additional non-visual features can be seen to increase the precision of the retrieval in all the three combinations. In those cases the recall level where the precision starts to decline is also substantially higher. Using the non-visual features seems to bring the final recall very close to unity already when only one third of the images has been retrieved.
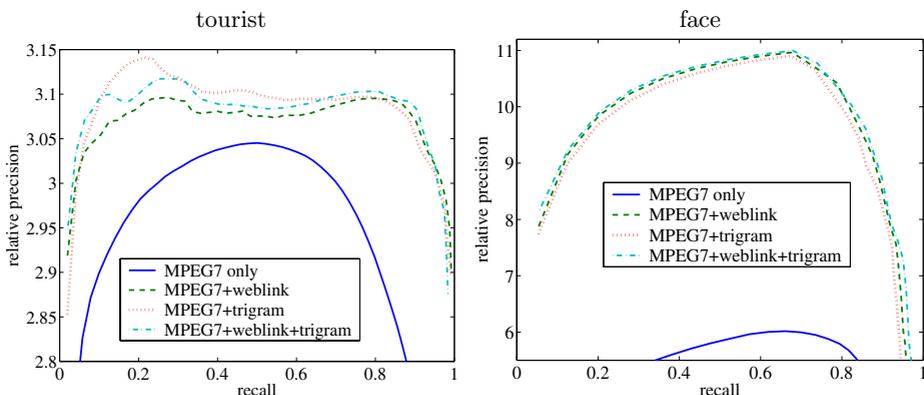

Fig. 2. Recall–relative precision graphs for classes tourist and face

6   Conclusions

In this paper, we have studied the use of hierarchical object trees to represent relationships between multimedia objects in a content-based information retrieval system. We have demonstrated that such structures can improve the performance of the system by implementing parent-child relevance sharing in the object trees. The novel idea is that while we are searching for a certain object type, for example images, other related object types in the database, like

web page links, can implicitly contribute to the retrieval. This technique can be used whenever one has a database with multiple object types with hierarchical interrelations.

Our approach can be seen to complement the *semantic web* paradigm [12], where semantic information is explicitly embedded in web documents. The hierarchical object structures used in our work incorporate certain forms of semantic knowledge in an automated way, which will reduce the required manual annotation work. On the other hand, future developments in our system could utilise semantic web information as an additional feature in the hierarchical structure.

References

1. Laaksonen, J., Koskela, M., Oja, E.: PicSOM—Self-organizing image retrieval with MPEG-7 content descriptions. IEEE Transactions on Neural Networks, Special Issue on Intelligent Multimedia Processing **13** (2002) 841–853
2. Kohonen, T.: Self-Organizing Maps. Third edn. Volume 30 of Springer Series in Information Sciences. Springer-Verlag (2001)
3. Kherfi, M., Ziou, D.: Image retrieval from the world wide web: Issues, techniques and systems. ACM Computing Surveys **36** (2004) 35–67
4. M. La Cascia, S. Sethi, S.S.: Combining textual and visual cues for content-based image retrieval on the world wide web. IEEE Workshop on Content-Based Access of Image and Video Libraries (1998) 24–28
5. Ortega-Binderberger, M., Mehrotra, S., Chakrabarti, K., Porkaew, K.: Webmars: A multimedia search engine. In: Proceedings of the SPIE Electronic Imaging 2000: Internet Imaging, San Jose, CA (2000)
6. Mukherjea, S., Hirata, K., Hara, Y.: Amore: A world wide web image retrieval engine. World Wide Web **2** (1999) 115–132
7. Wactlar, H.D., Kanade, T., Smith, M.A., Stevens, S.M.: Intelligent access to digital video: Informedia project. IEEE Computer **29** (1996) 46–52
8. Salton, G., McGill, M.J.: Introduction to Modern Information Retrieval. Computer Science Series. McGraw-Hill (1983)
9. MPEG: MPEG-7 visual part of the eXperimentation Model (version 9.0) (2001) ISO/IEC JTC1/SC29/WG11 N3914.
10. Laakso, S., Laaksonen, J., Koskela, M., Oja, E.: Self-organizing maps of web link information. In Allinson, N., Yin, H., Allinson, L., Slack, J., eds.: Advances in Self-Organising Maps, Lincoln, England, Springer (2001) 146–151
11. FIPS: Secure hash standard (1995) PUB 180-1, `http://www.itl.nist.gov/fipspubs/fip180-1.htm`.
12. Berners-Lee, T., Hendler, J., Lassila, O.: The semantic web. Scientific American **284** (2001) 28–37