
Self-Organizing Maps of Web Link Information

Sami Laakso, Jorma Laaksonen, Markus Koskela, and Erkki Oja

Laboratory of Computer and Information Science
Helsinki University of Technology
P.O. Box 5400, Fin-02015 HUT

Summary. We have developed a method that utilizes hypertext link information in image retrieval from the World Wide Web. The basis of the method consists of a set of basic relations that can take place between two images in the Web. Our method uses the SHA-1 message digest algorithm for dimension reduction by random mapping. The Web link features have then been used in creating a Self-Organizing Map of images in the Web. The method has been effectively tested with our PicSOM content-based image retrieval system using a Web image database containing over a million images. The method can as such be used also in other Web applications not related to content-based image retrieval.

1 Introduction

There is a growing interest in the ability to search the World Wide Web for various data. Automated methods for information retrieval from the Web have in recent years attracted significant research interest and several search engines have been developed and made available for public use. These search engines allow the user to search for multimedia resources, including images, inside Web documents.

A recent approach to the problem of locating relevant images in huge databases such as the Web is content-based image retrieval (CBIR). It is based on visual features that can automatically be extracted from images without human intervention or interpretation. Such features can, for example, be the colors and textures found in the image as well as shape, structure and composition of the image scene. Besides visual features, the hyperlink structure of the Web can also be a rich source of information about the content of the environment, and some promising methods have already been proposed.

According to [1], useful documents in the Web can be categorized as either authorities or hubs. An authority is a source of specific information about a certain topic and a hub is a collection of links pointing to authorities. A good authority is recognized from the fact that it is pointed at by many hubs and a good hub is therefore a document that contains links to many good authorities. Another page ranking method, which is now used by the popular text-based search engine Google (<http://www.google.com>), was proposed in [2]. It uses a hyperlink graph constructed on the entire set of documents

retrieved from the Web. Given a query, Google retrieves the documents containing the query string in the order specified by the probability of a random walker to visit the page.

The problem with these methods is that they have been designed for text-based page search and cannot therefore be directly generalized into image search. One solution would be to make a normal query and then show images from those pages which were ranked best. However, this approach assumes that the images in a certain page are generally related to the surrounding text, which certainly is not always true.

2 Web Relation Feature Extraction

We have developed a mechanism which utilizes the information about the location of images in the Web and the hypertext link structure between them. The basis of the method consists of a set of basic Web relations that can take place between two images. For example, if one image acts as a hypertext link to another image (e.g. thumbnails) it can be assumed that the two images are closely related. Lesser but still highly informative image relation occurs if the images are located in the same Web page, or in the same directory, or at least in the same domain. The same relations can also be applied to Web pages that contain images. Although it would be possible to perform also a deeper study of links between image pages, we consider that images beyond two link steps are unlikely to be related. Another benefit of using only one level deep linkage information is that now only the links of image pages need to be saved. A deeper study would require the information about the entire indexed hypertext structure, which can be very space consuming.

Every object in the Web has its own unique URL. However, in order to efficiently exploit this location information and information concerning inter-location relations, we need to convert them into mathematical form. A trivial solution would be to form a relation vector whose dimensionality equals to the number of images in the database. Then a certain weight value would be set to each vector component according to the relation between the corresponding images. However, this would require storing N^2 relation values which is not feasible with large databases. In our case N was of the order of one million and therefore the dimensionality of the relation data had to be reduced.

Random mapping [4] provides a computationally feasible method for reducing the dimensionality of data so that mutual similarities between the data vectors are approximately preserved. The dimensionality of the representations is reduced by replacing the original orthogonal base with a lower-dimensional almost orthogonal base.

Secure Hash Algorithm (SHA-1) [3] is a powerful method for computing a condensed representation of a message. When a message of length $< 2^{64}$ bits is input, the SHA-1 produces a 160-bit output called a *message digest*. The

Table 1. Used Web relations and the corresponding weights.

Web relation	Weight
URL of an activation link	1.5
URL of the image	1.4
URL of the image's Web directory	1.3
URL of the image page	1.2
URL of the image page's Web directory	1.15
URL of the image page's domain	1.1
URL of a link to other image or page	1.0

SHA-1 has been designed so that it is computationally infeasible to find a message which corresponds to a given message digest, or to find two different messages which produce the same message digest. The latter property was the main inspirer for us to use this algorithm as an indexation tool.

First, all URLs were extracted from the collected images pages. These URLs contained all external links, the locations of the images and image pages themselves, and the directories and domains of the image pages. Then, SHA-1 message digests were calculated for each URL. The first 8 characters (32 bits) of the digests were used to determine the random projection by interpreting them as four 8-bit values. The first value was used as an index in the range $[0, 255]$, the second in $[256, 511]$, the third in $[512, 767]$, and the fourth in $[768, 1023]$. These four indices were used in setting four components of otherwise zero 1024-dimensional vector to value one. These 1024-dimensional vectors were then considered as being random base vectors of an almost orthogonal base whose dimensionality was 2 622 472, the count of unique URLs extracted from the image pages.

For each image in the database, the random base vectors corresponding to the extracted URLs in the image's Web page were multiplied with a weight value that depended on the type of the URL in question, as shown in Table 1. All the resulting random projection vectors for one particular image were combined into a relation feature vector so that for each index the maximum value was chosen. By forming relation feature vectors from all images, images which have close Web relations will also have small Euclidean distances between the feature vectors.

An example of the forming of one random base vector follows. Consider an imaginary URL of *http://www.cis.hut.fi/images/image1.jpg*. From that URL the SHA-1 algorithm results a 160-bit message digest which has the hex form of 4CDF3EEB45A12D6044A96A911E7559084B3F037F. The first 32 bits of the message digest are then used to determine the index values and the result is a random projection vector which has weight values in indices 76 ($= 4C_{16}$), 479 ($= 256 + DF_{16}$), 574 ($= 512 + 3E_{16}$) and 1003 ($= 768 + EB_{16}$), as shown in Table 2. The final relation feature vector would then be the combination of this vector and the random projection of the image directory URL (that would be *http://www.cis.hut.fi/images/*) and other previously defined URLs.

Table 2. An example of using SHA-1 message digest for random mapping.

Bit positions	0...7	8...15	16...23	24...31	32...159
SHA-1 hex value	4C	DF	3E	EB	45A12D6...
Decimal value	76	223	62	235	
Random base indices	76	479	574	1003	

3 PicSOM

The PicSOM system [5] is a framework for research on algorithms and methods for content-based image retrieval. PicSOM implements *relevance feedback* by using Tree Structured Self-Organizing Map (TS-SOM) [6] in storing the user responses and in selecting the images. The TS-SOM differs from the original SOM [7] in that TS-SOM consists of a stack of SOM layers. The BMUs are first searched for on the topmost layer and the search is then continued on the next layer in a restricted area centered below the BMU on the above map. This makes the BMU search much faster, otherwise the properties of SOM and TS-SOM are similar.

Images that are similar to each other with respect to a particular feature extraction method are clustered together on the corresponding TS-SOM map. When the user's relevance feedback is marked on the maps and this spatial relevance function low-pass filtered, what results is automatic adaptation to the user's conception of image similarity and relevance. The mutual weighting of different feature types is performed simultaneously, as features that map relevant images in tight clusters are automatically given more weight than the others. A genuine characteristic of PicSOM is thus its ability to automatically adapt to the user's perception of similarity of images based on their low-level visual content even though humans perceive image similarity on abstract semantic level. An on-line demonstration of PicSOM and comprehensive documentation of it can be found at <http://www.cis.hut.fi/picsom/>.

4 Web-fi Image Database

Our experimental Web image database was retrieved in summer and autumn 2000. The original plan was to collect all images existing in the registered domains in Finland (i.e. all the domains whose name end in **.fi**). It was assumed that there exists a path from the host's main page to a majority of pages in that domain, and therefore the search was started from every domain's root URL. To avoid retrieving too many thumbnails and icons, our Web robots retrieved only images whose width and height were both more than 50 pixels and the image also had to contain at least five distinct colors. Accepted image formats were JPEG, GIF, TIFF, and PNG. The collection process was stopped when the count of retrieved images exceeded one million. Overall, the

Web robots examined 2 176 261 pages in 12 991 different domains. The total count of unique images was 1 008 844. Included in the examined domains are all domains owned by Finnish cities, municipalities, universities, and polytechnics. During the collection, link structure information corresponding to each retrieved image was also collected. After the collection process was completed, the Web relation feature vectors for all images were calculated. Then a five-layer TS-SOM, with layers of sizes 4×4 , 16×16 , 64×64 , 256×256 , and 1024×1024 , was constructed.

5 Results

Figure 1 shows a partial surface of the 1024×1024 -sized SOM formed from Web-fi database. The benefits of the link information can be seen, as there are certain clearly visible clusters of images, for example a group of trucks in the bottom left corner of the map and a group of images from technical books in the bottom right corner. Although for example some of the truck images are visually quite dissimilar, they are still mapped close to each other. This can be regarded as a promising achievement, as the system uses no visual nor textual data, but only graph information about the Web structure.

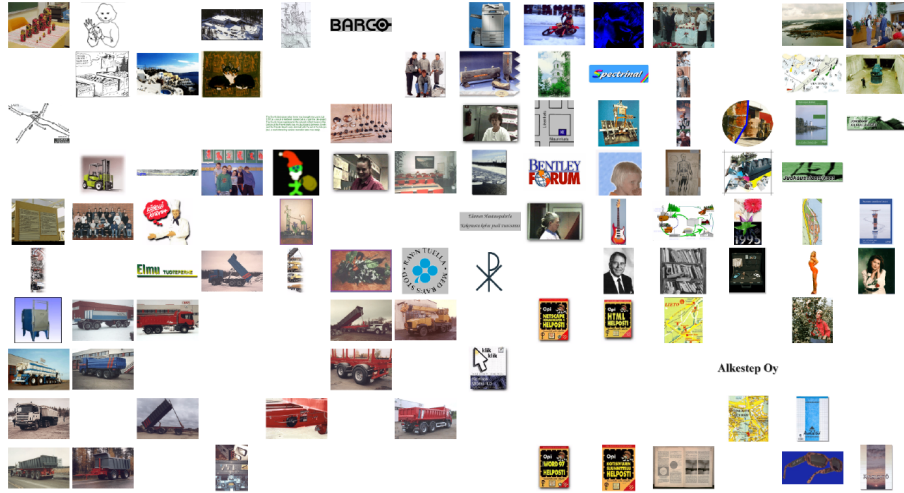


Fig. 1. Partial surface of 1024×1024 -sized SOM formed from Web structure data.

Figure 2 shows another example of the discrimination power of our method. The figure shows the mappings of all images from four distinct domains. The selected domains were *hut.fi* (Helsinki University of Technology), *oulu.fi* (University of Oulu), *ouka.fi* (City of Oulu) and *utu.fi* (University of Turku), and the corresponding image counts were 103 613, 47 845, 21 786 and 21 495 images, respectively. The resulting image densities show that some of the domains are very tightly clustered, whereas others are spread more widely.

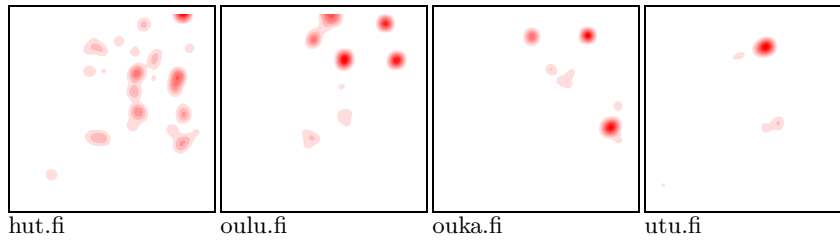


Fig. 2. Mappings of different domains on the lowest-level SOM (1024×1024). The distributions have been low-pass filtered to ease inspection.

6 Conclusions and Future Plans

In this paper, we have presented a method for utilizing the Web link information in CBIR. Our method requires no information about the entire hypertext structure, but only the relevant data of the image pages themselves. The preliminary results are very promising and indicate that the method can as such be used also in other applications not related to content-based image retrieval. However, the relation feature is not intended to be a stand-alone feature, but rather to be used in combination with other features. Especially in the CBIR, the link structure information may greatly enhance the search.

References

1. Kleinberg, J. (1997) Authoritative sources in a hyperlinked environment, *IBM Technical Report RJ 10076*, May 1997.
2. Brin, S. and Page, L. (1998) The Anatomy of a Large-Scale Hypertextual Web Search Engine, *Proceedings of Seventh International World Wide Web Conference, April 1998, Brisbane, Australia*.
3. FIPS PUB 180-1 Secure Hash Standard (1995)
<http://www.itl.nist.gov/fipspubs/fip180-1.htm>
4. Kaski, S. (1998) Dimensionality Reduction by Random Mapping: Fast Similarity Method for Clustering, *Proceedings of IEEE International Joint Conference on Neural Networks, May 1998, Anchorage, Alaska*.
5. Laaksonen, J., Koskela, M., Laakso, S. and Oja, E. (2000) PicSOM - Content-based image retrieval with self-organizing maps, *Pattern Recognition Letters* **21**(13-14): 1199–1207.
6. Koikkalainen, P. and Oja, E. (1990) Self-organizing hierarchical feature maps, *Proceedings of 1990 International Joint Conference on Neural Networks, January 1990, Washington, USA*, Vol. 2, pp. 279–284.
7. Kohonen, T. (2001) *Self-Organizing Maps*, Vol. 30 of *Springer Series in Information Sciences*, Springer-Verlag, Third Edition.