ORGANIZING WWW IMAGES BASED ON THE ANALYSIS OF PAGE LAYOUT AND WEB LINK STRUCTURE

Deng Cai, Xiaofei He, Wei-Ying Ma, Ji-Rong Wen, Hongjiang Zhang*

Microsoft Research Asia, Beijing 100080, China
*Department of Computer Science, University of Chicago

ABSTRACT

Due to the rapid growth of the number of digital images on the Web, there is an increasing demand for effective and efficient method for organizing and retrieving the images available. This paper describes a method for clustering and embedding WWW images. By using a vision-based page segmentation algorithm, a web page is partitioned into blocks, and the textual and link information of an image can be accurately extracted from the block containing that image. By extracting the page-to-block, block-to-image, block-to-page relationships through link structure and page layout analysis. we construct an image graph. With the image graph model, we use techniques from spectral graph theory for image clustering and embedding. Some experimental results are given in the paper.

1. INTRODUCTION

The emergence of World Wide Web (WWW) has created many new opportunities but also challenges for organizing and searching a large volume of images available publicly. The traditional image retrieval techniques based on content analysis, such as those content-based image retrieval (CBIR) systems, are usually focused on small, static, and close-domain image data such as personal photo album. When it comes to searching WWW images, these techniques may not be able to scale up to handle the large number of available images. Moreover, different from traditional image retrieval and browsing, there is a lot of additional information on the web, such as surrounding texts, page layout and hyperlinks which is useful to organize the images.

In this paper, we propose a method for clustering and embedding WWW images using link and page layout analysis. Here, by "embedding" we mean that each image on the web can be equipped with a vector representation in a Euclidean space such that the semantic relationships between images can be reflected from the Euclidean distances between them. Technically speaking, our method contains three parts, VIsion-based Page Segmentation (VIPS) [1], link and page layout based graph model, and spectral analysis for image clustering and embedding.

Link analysis [3][4] has received a lot of attention in recent years. Most of web-based search engines regard web pages as atomic units. However, it is the case that a single web page often does not contain pure content but also things like navigation, decoration, interaction elements. It is also often the case that a single web page contains multiple topics. By using the VIPS algorithm, each page can be segmented into a number of blocks which contain semantically related information. For image organization, we are interested in those blocks containing images (called image blocks hereafter). By page layout analysis, each page can be represented as tree. The hyperlinks are extracted in each block. With web page blocks, the link is considered from block to page, rather than from page to page. We consider three kinds of relationships, i.e. block-to-page (link structure), page-to-block (page layout) and block-to-image (inclusion relation), which ultimately results in an image graph. With the image graph, we use techniques from spectral graph theory [2] for image clustering and embedding.

The rest of this paper is organized as follows: Section 2 briefly describes the VIPS page segmentation algorithm. In Section 3, we describe how to build an image graph. We present our method for image clustering and embedding in Section 4. Some experimental results are provided in Section 5. Finally, we give concluding remarks and future work in Section 6.

2. VISION-BASED PAGE SEGMENTATION (VIPS)

The VIsion-Based Page Segmentation (VIPS) algorithm [1] extracts the semantic structure of a web page based on its visual presentation. Such semantic structure is represented as a tree; each node in the tree corresponds to a block. Each node will be assigned a value (*Degree of Coherence*) to indicate how coherent of the content in the block based on visual perception. For details about VIPS algorithm, please see [1].

3. IMAGE GRAPH MODEL

The Internet can be viewed as a forest and every web page can be viewed as a tree, as shown in Fig. 1. The links are from blocks to pages. In this section, we describe how to construct an image graph whose weights defined on the edges reflect semantic relationships between images. We first describe how to build page graph and block graph from which the image graph can be further induced. Let P denote the set of all the web pages, $P = \{p_1, p_2, \ldots, p_k\}$, where k is the number of web pages. Let B denote the set of all the blocks, $B = \{b_1, b_2, \ldots, b_n\}$, where n is the number of blocks. It is important to note that, for each block there is only one page that contains that block. Let $I = \{I_1, I_2, \ldots, I_m\}$ denote the set of all the images on the web, where m is the total number of the web images. $b_i \in p_j$ means the block i is contained in the page j. Similarly, $I_i \in b_j$ means the image i is contained in the block j.

3.1 Page Graph

Let G_P (V_P, W_P) denote the page graph, where V_P is the set of pages and W_P is the weight matrix. W_P can be simply defined as follows. W_P (i, j) is 1 if page i links to page j, and 0 otherwise. This definition is pretty simple yet has been widely used as the first step to many applications, such as PageRank [3], HITS [4], etc. However, different blocks in a page are of different importance. Therefore, those links in blocks with high importance value

0-7803-8603-5/04/$20.00 ©2004 IEEE

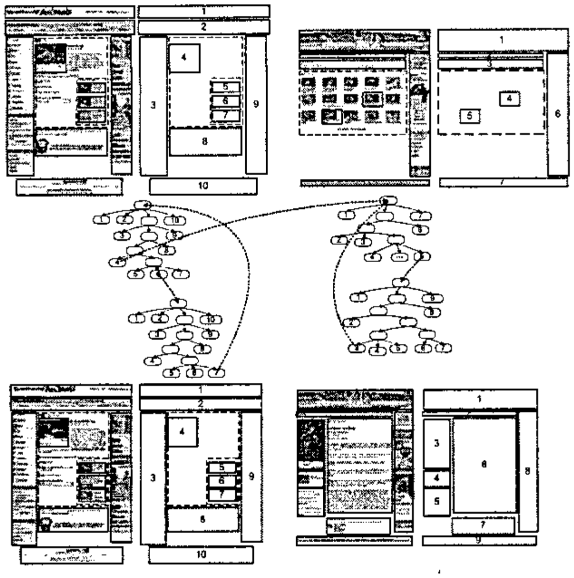


Fig. 1 The Internet can be viewed as a forest and every web page can be viewed as a tree. The hyperlinks (dashed line) are from blocks to pages rather than from pages to pages

should be treated more important than those in blocks with low importance value. In other words, in the random walk model, we assume the surfer might prefer to follow those links in important blocks. The block importance function can be simply defined as follows:

$$f_p(b) = \alpha \frac{\text{size of block } b \text{ in page } p}{\text{dist. from the center of } b \text{ to the center of screen}} \qquad (1)$$

where α is a normalization factor such that

$$\sum_{b\in p} f_p(b) = 1 \qquad (2)$$

The link structure can be described as a block-to-page matrix Z whose elements are defined as follows

$$Z_{ij} = \begin{cases} 1/s_i & \textit{if there is a link from block } i \textit{ to page } j \\ 0 & \textit{otherwise} \end{cases} \qquad (3)$$

Thus, the weight matrix W_P can be defined as follows

$$W_P(\alpha,\beta) = \sum_{b\in\alpha} f_\alpha(b) Z(b,\beta), \quad \alpha,\beta \in P \qquad (4)$$

The above equation can be rewritten as

$$W_P = XZ \qquad (5)$$

where X is a page-to-block matrix and its element is as follows

$$X_{ij} = \begin{cases} f_{P_i}(b_j) & \textit{if } b_j \in p_i \\ 0 & \textit{otherwise} \end{cases} \qquad (6)$$

3.2 Block Graph

Let $G_B(V_B, W_B)$ denote the block graph, where V_B is the set of blocks and W_B is the weight matrix. The block graph is constructed over the blocks. Let's first consider a jump from block a to block b. Suppose a user is looking at block a. In order to jump to block b, he first jumps to page β which contains block b, and then he focuses his attention on block b. Thus, a natural definition of W_B is as follows

$$\begin{aligned} W_B(a,b) &= Prob(b|a) \\ &= \textstyle\sum_{\gamma\in P} Prob(\gamma|a)Prob(b|\gamma) \\ &= Prob(\beta|a)Prob(b|\beta) \\ &= Z(a,\beta)X(\beta,b), \quad a,b\in B \end{aligned} \qquad (7)$$

or

$$W_B = ZX \qquad (8)$$

where W_B is a $n \times n$ matrix. By definition, W_B is clearly a probability transition matrix. However, there is still one limitation of this definition such that it is unable to reflect the relationships between the blocks in the same page. Two blocks are likely related to the same topic if they appear in the same page. This leads to a new definition,

$$W_B = (1-t)ZX + tDU \qquad (9)$$

where t is a suitable constant. D is a diagonal matrix, $D_{ii} = \sum_j U_{ij}$. U_{ij} is zero if block i and block j are contained in different pages; otherwise, it is set to the *DOC* (degree of coherence, see [1] for details) value of the smallest block containing both block i and block j. It is easy to check that the sum of each row of DU is 1. Thus, W_B can be viewed as a probability transition matrix such that $W_B(a, b)$ is the probability of jumping from block a to block b. Finally, it is worth noting that t is typically set to a small number between 0 and 0.1. Since in many cases, different blocks are likely related to different topics even though they appear in the same page.

3.3 Image Graph

Let $G_I(V_I, W_I)$ denote the image graph, where V_I is the set of images and W_I is the weight matrix. Once the block graph is obtained, the image graph can be constructed correspondingly by noticing the fact that every image is contained in at least one block. Let's consider the jump from image i to image j. From image i we first see the block α containing image i. By block graph, we get a jump from block α to block β containing image j. Finally, we stopped at image j. In this way, the weight matrix of the image graph can be defined as follows:

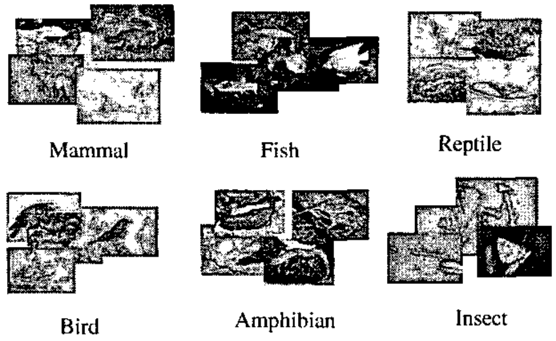
$$W_I(i,j) = \sum_{i\in\alpha, j\in\beta} W_B(\alpha,\beta) \qquad (10)$$

or

$$W_I = Y^T W_B Y \qquad (11)$$

where W_I is a $m\times m$ matrix. If two images i and j are in the same block, say b, then $W_I(i, j) = W_I(b, b) = 0$. However, the images in the same block are supposed to be semantically related. Thus, we get a new definition as follows:

$$W_I = tDY^TY + (1-t)Y^TW_BY \qquad (12)$$



Mammal Fish Reptile

Bird Amphibian Insect

Fig. 2. Our image database is composed of six categories.

where t is a suitable constant and D is a diagonal matrix, $D_{ii} = \sum_j (Y^TY)_{i\,j}$. Like W_B, W_I can be viewed as a probability transition matrix. t is typically set to be large, 0.7~0.9, since in many cases two images are related to the same topic if they appears in the same block.

4. CLUSTERING AND EMBEDDING WWW IMAGES

We have described how to obtain a weight matrix of the image graph, W_I. We first convert it into a similarity matrix S such that $S = 1/2\,(W_I + W_I^T)$ which is symmetric.

Suppose y_i is an one-dimensional vector representation of image i. The optimal $\mathbf{y} = (y_1, \ldots, y_m)$ is obtained from the following objective function:

$$\min_{\mathbf{y}} \sum_{i,j} (y_i - y_j)^2 S_{ij} \tag{13}$$

The objective function with the choice of S_{ij} incurs a heavy penalty if semantically related images are mapped far apart. Therefore, minimizing it is an attempt to ensure that if image i and image j are semantically related then y_i and y_j are close to each other. Let D be a diagonal matrix whose i^{th} element is the row (or column, since S is symmetric) sum of S, $D_{ii} = \sum_j S_{ij}$. By simple algebra formulation, the minimization problem reduces to finding:

$$\min_{\mathbf{y}^T D\,\mathbf{y}=1} \mathbf{y}^T L\mathbf{y} \tag{14}$$

where $L = D - S$. L is generally called Laplace matrix, or graph Laplacian. It is positive semi-definite. The solution is given by minimum eigenvalue solution to the generalized eigenvalue problem:

$$L\mathbf{y} = \lambda D\mathbf{y} \tag{15}$$

Let $(\mathbf{y}^0, \lambda^0)$, $(\mathbf{y}^1, \lambda^1)$, ..., $(\mathbf{y}^{m-1}, \lambda^{m-1})$ be the solutions to the above equation, and $\lambda^0 < \lambda^1 < \ldots < \lambda^{m-1}$. It is easy to check that $\lambda^0 = 0$ and $\mathbf{y}^0 = (1, 1, \ldots, 1)$. Therefore, we leave out the eigenvector $\mathbf{y}^0$ and use the next k eigenvectors for embedding in k-dimensional Euclidean space:

$$image\ \ j \leftarrow (\mathbf{y}^1(j), \cdots, \mathbf{y}^k(j)) \tag{16}$$

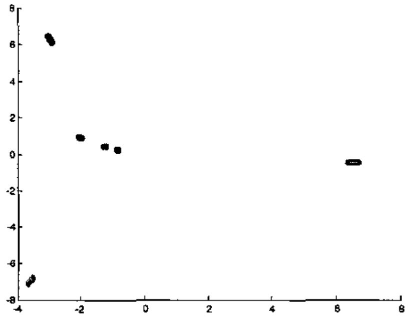


Fig. 3. 2-D embedding of WWW images using our techniques. Each color represents a semantic category. Clearly, they are well separated.

where $\mathbf{y}^i(j)$ denotes the j^{th} element of $\mathbf{y}^i$. In this way, we endow each image with a vector representation in the Euclidean space. Note that, all the matrices involved in this computation are sparse, and hence the computation can be performed very fast.

Once we obtain vector representations of the images, clustering is straightforward. The simplest way is to use $\mathbf{y}^1$ (called *Fiedler vector* in spectral graph theory [2]) to cut the image collection into several pieces. Another way is to use k-means clustering algorithm on the image vectors. Previous works demonstrated that spectral embedding followed by k-means can produce good result [5].

There are many applications of the embedding and clustering results. The major one should be browsing. The images can be grouped into semantic categories. Also, they can be visualized in a two-dimensional plane.

5. EXPERIMENTS

In this section, several illustrative examples are given. Due to the lack of sufficient resources, we are currently not able to perform image search on the whole Internet which is our ultimate goal. The purpose of this section is to provide people with an intuition on how our system works based on the techniques we described previously.

5.1 Data Preparation

All the data used in our experiments are crawled from the Web, starting from the following URL:

http://www.yahooligans.com/content/animals/

We crawled 1288 web pages in total. All the pages are restricted to be within this directory. From these web pages, 1710 JPG images are extracted. It can be clearly seen from the website that our database is composed of six categories, *i.e.* mammal, fish, reptile, bird, amphibian and insect, as shown in Fig. 2. Each category can be further divided into sub-categories since some spices are more related.

5.2 Clustering WWW images

We first construct the image graph which reflects the semantic relationships between images. With this image graph, the images were embedded in a 2-dimensional Euclidean space such that the Euclidean distances reflect their semantic relationships. Figure 3 shows the embedding results. Each data point represents an image. Each color stands for a semantic class. Clearly, the image database can be accurately clustered into six categories.

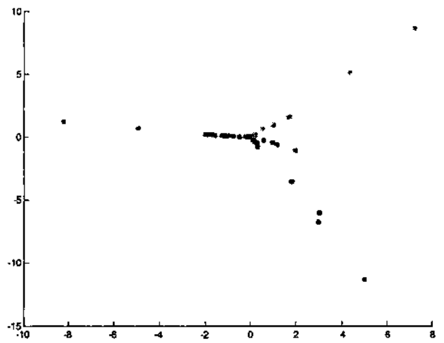


Fig. 4. 2-D embedding of WWW images. The image graph was constructed from traditional perspective that the hyperlinks are regarded as from page to page. The image graph was induced from the page-to-page and page-to-image relationships.

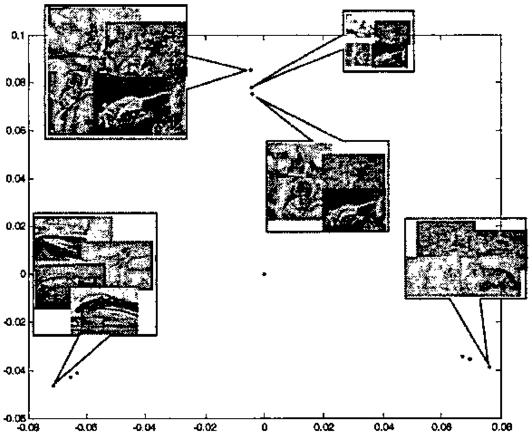


Fig. 5. 2-D visualization of WWW images using the second and third eigenvectors.

If we use the traditional link analysis methods that regard, hyperlinks as from page to page, the 2-D embedding result is shown in Fig. 4. As can be seen, the six categories are mixed together and can be hardly separated. This comparison shows that our image graph model is much more powerful than traditional methods as to describing the intrinsic semantic relationships between WWW images.

5.3 2-D Visualization of WWW images

For each category described in the previous section, it can be visualized in a two-dimensional plane such that the spatial structure reflects its semantic structure. We present the 2-D visualization result of the mammal category in Fig. 5 using the second and third eigenvectors. In our database, every kind of animal has three images with different sizes. As can be seen from Fig. 5, each point represents 4 images which belong to the same sub-category. It is interesting to note that the three points close to each other exactly represent three different sizes. The 2-D visualization using the fifth and sixth eigenvector is given in Fig. 6.

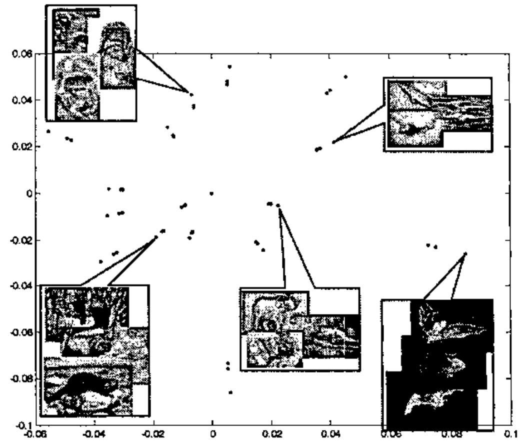


Fig. 6. 2-D visualization of WWW images using the fifth and sixth eigenvector

6. CONCLUSIONS

In this paper, we described a method for clustering and embedding WWW images. Different from traditional methods which regard the web pages as atomic units, we semantically partition web pages into blocks and the hyperlinks are extracted within blocks and the relationships between blocks are extracted. Three kinds of graph models, *i.e.* page graph, block graph and image graph, are constructed based on link and page layout analysis. By using techniques from spectral graph theory, the WWW images can be clustered into semantic classes. Also, each image can be endowed with a vector representation in Euclidean space. Thus, the semantic structure of the images can be reflected by the spatial structure of the images. Several experiments on real world data have demonstrated the effectiveness of our methods.

REFERENCES

[1] D. Cai, S. Yu, J.-R. Wen and W.-Y. Ma, "VIPS: a vision-based page segmentation algorithm", Microsoft Technical Report, MSR-TR-2003-79, 2003..

[2] Fan Chung. Spectral graph theory. *Regional conference series in mathematics*, no 92, American Mathematical Soceity, Providence, RI, 1997.

[3] S. Brin and L. Page, "The anatomy of a large scale hypertextual (Web) search engine", In *The Seventh International World Wide Web Conference, 1998.*

[4] J. Kleinberg, "Authoritative sources in a hyperlinked environment", *Proc. 9th ACM-SIAM Symposium on Discrete Algorithms*, 1998.

[5] Andrew Y. Ng, M. Jordan, and Y. Weiss, "On spectral clustering: Analysis and an algorithm", *Advances in Neural Information Processing Systems* 14, Vancouver, Canada, 2001.