Multi-Model Similarity Propagation and its Application for Web Image Retrieval

Xin-Jing Wang[1,2] wxj01@mails.tsinghua.edu.cn
Wei-Ying Ma[1] wyma@microsoft.com
Gui-Rong Xue[1,3] grxue@sjtu.edu.cn
Xing Li[2] xing@cernet.edu.cn

Microsoft Research Asia[1]
Department of Electronic Engineering, Tsinghua University, China [2]
Shanghai Jiao Tong University, China[3]

ABSTRACT

In this paper, we propose an iterative similarity propagation approach to explore the inter-relationships between Web images and their textual annotations for image retrieval. By considering Web images as one type of objects, their surrounding texts as another type, and constructing the links structure between them via webpage analysis, we can iteratively reinforce the similarities between images. The basic idea is that if two objects of the same type are both related to one object of another type, these two objects are similar; likewise, if two objects of the same type are related to two different, but similar objects of another type, then to some extent, these two objects are also similar. The goal of our method is to fully exploit the mutual reinforcement between images and their textual annotations. Our experiments based on 10,628 images crawled from the Web show that our proposed approach can significantly improve Web image retrieval performance.

Categories and Subject Descriptors

H3.3 [**Information Search and Retrieval**]: Retrieval models

General Terms

Algorithms, Design, Performance

Keywords

Multimedia Retrieval, Mixture Model, Mutual Reinforcement, Iterative Similarity Propagation

1. INTRODUCTION

Multimodal Image Retrieval attempts to leverage simultaneously several data types (e.g. image contents, surrounding texts, and links) to improve retrieval performance [2][4][7][8][9][17][20][22]. A major technical challenge in Multimodal Image Retrieval is how to combine different retrieval models in order to achieve best performance. As the data are heterogeneous and inter-related, it is difficult to evaluate the contribution of each individual data type, and therefore, the optimal combination of different models is unclear. Current approaches for combing different models include simple linear combination [8][17], resorting to human interaction [9][22], and probabilistic models [2][4][20].

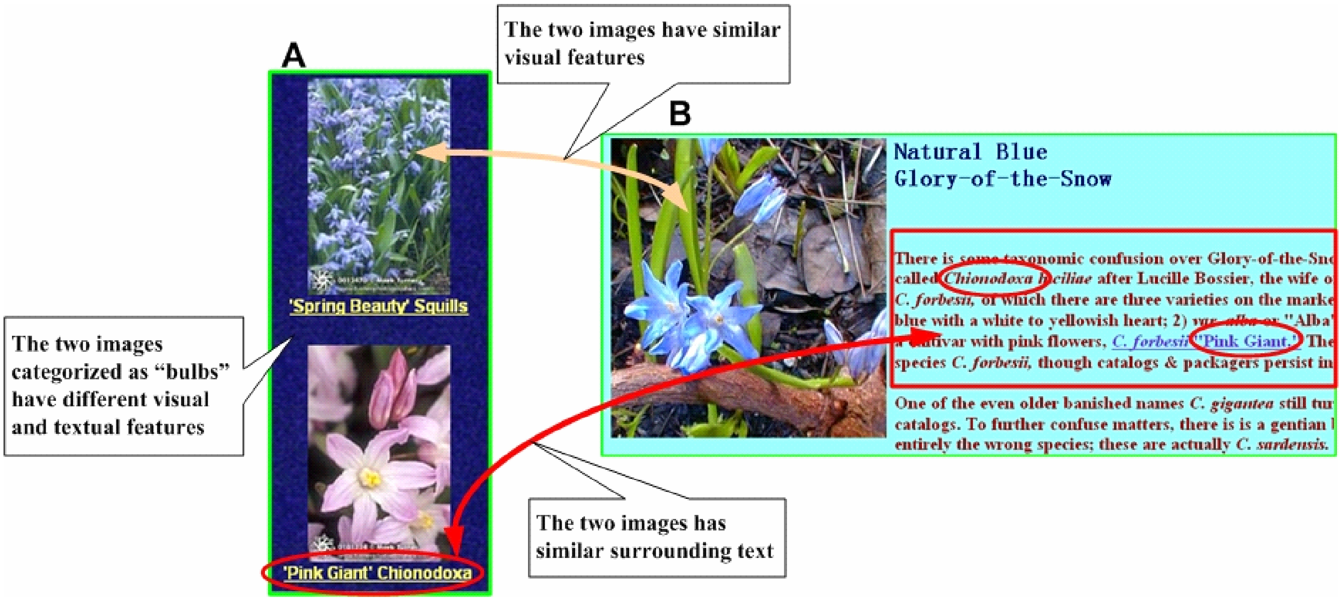
There are two big drawbacks in these existing approaches. First, they are greatly affected by the features of data [8][17][20]. Two semantically similar images may have entirely different visual features. Although [9][22] proposed methods that partly solve this problem by discovering the semantically similar terms/images through user interaction, the computational cost of these methods is high. Second, in these approaches, the relationships among different data types are treated as additional features, and these features remain unchanged during the learning process. The mutual reinforcement across sets of related data types is not fully explored.

Figure 1 shows an example when the former approaches may possibly fail. A and B are two web-pages. The left two images in Figure 1 are categorized by the author of web-page A as relevant images (i.e. "bulbs"). However, their visual and textual features (i.e. surrounding text) are both quite different. It is obvious that using the linear combination or probabilistic combination methods, these two images will most probably be regarded as dissimilar.

Recently, many applications in text retrieval and Web mining have indicated that relational links between objects provide a useful source of information [10][12][13][14][16][19][21]. In [19] the authors use relationships among different types of objects to improve the cluster quality of interrelated data. In [21], a method is proposed for spreading the similarities inside and across sets of interrelated objects to discover implicitly similar objects. Both of these methods leverage the inter-type relations by iteratively projecting the clustering results or similarities of one data type to another.

In this paper, we attempt to learn the semantic similarities of images using the relations between Web images and their textual annotations. We propose a method called *Iterative Similarity Propagation* which iteratively reinforces the similarities between images by their textual annotations, and vice versa. Our goal is to explore the interrelation between multiple modalities, and by this way, to discover the intrinsic similarity of images and improve the

ACM Multimedia'04, October 10-16, 2004, New York, NY USA.


A
B
'Spring Beauty' Squills
'Pink Giant' Chionodoxa
Natural Blue
Glory-of-the-Snow
There is some taxonomic confusion over Glory-of-the-Sno
called Chionodoxa luciliae after Lucille Bossier, the wife o
C. forbesii, of which there are three varieties on the marke
blue with a white to yellowish heart; 2) var. alba or "Alba"
a cultivar with pink flowers, C. forbesii "Pink Giant." The
species C. forbesii, though catalogs & packagers persist in
One of the even older banished names C. gigantea still tur
catalogs. To further confuse matters, there is is a gentian
entirely the wrong species; these are actually C. sardensis.

The two images have similar visual features

The two images categorized as "bulbs" have different visual and textual features

The two images has similar surrounding text

Figure 1. The two modalities, image content and textual information, can together help group similar Web images

performance of image retrieval. The assumption is that if two objects of the same type are both related to one of another type, these two objects are similar; likewise, if two objects of the same type are related to two different, but similar objects of another type, then to some extent, these two objects are also similar. Note that the meaning of "similarity propagation" is two-fold: enhancing or reducing the similarity of two objects (image or text). By enhancing, we mean that the similarity of two objects will be increased if they relate to similar objects. Quite the reverse, by reducing we mean that the similarity of two objects is decreased if they relate to dissimilar objects. Figure 1 shows a case of similarity enhancement. Consider a segment of the web-page B which contains an image and its surrounding texts. The image in B has similar visual features to the upper image in A and has similar surrounding text to the bottom image. Thus using B as a bridge, the similarity between two images in A should be increased. However, using the former approaches [8][17][20] will not be able to cluster these two images together.

It is valuable to highlight some key-points of our method here:

1. Instead of treating the image textual annotations as an additional feature for image retrieval, we use an iterative approach to exploring the mutual reinforcement between images and their textual annotations. This approach avoids the bias of features mentioned above, and provides a better combination of the image and text retrieval modality.

2. In our proposed approach, it is the similarity that is propagated between different objects (i.e. images and texts). It can deal with data sparseness problem [8][17][20] and reduce space complexity since the visual and textual features are often of high dimensional. The intra- and inter-object similarities are refined during the process, which can reduce both false positives and false negatives and reveal the intrinsic similarities in the semantic level.

3. Our method is an iterative process. The effect of each retrieval modality is propagated to its related modalities in each iteration, by which the interactions inside and across the sets of relational data are explored during the mutual reinforcement.

4. Fundamentally, our approach can be seen as a non-linear combination of different retrieval modalities, which better exploits the relationships among different data types and use these relationships to discover the implicit but semantic object similarities.

This paper is organized as follows. We discuss some related works in Section 2. In Section 3, we present the algorithm of *iterative similarity propagation* and detail its usage in image retrieval in Section 4. Section 5 gives the experiment evaluation of our method. We conclude our work in Section 6 with discussions and possible future directions.

2. RELATED WORKS

A number of researchers have introduced systems for searching image databases/Web with combined text and image data. Chen et al. [8] linearly combine the dot product similarities on textual features and Euclidean distances on visual features and set the two models equal weight. Srihari et al. [17] intend to find the optimal weight set for the multi-modalities by involving a training phase to learn a group of optimal weight set for the selected set of representative queries. And in the retrieval phase, they linearly combine the individual models using the weight set of the representative query which is the most similar to the current user submitted query. In their work, not only image and text, but face detection and recognition etc. are adopted. Cascia et al [7] represent each image by a composite of visual and textual features. They use PCA to reduce the dimension of visual features and use LSI to address problems with synonyms, word sense, lexical matching and term omission. [2][4][20] propose probabilistic models to integrate information provided by associated text and image features. [22] assumes that images in the database have precise keyword annotations and resorting to users' relevance feedback to discover the common keywords from the keyword vectors of those “positive” images. The query concept is then inferred from these common keywords. [9] implements the image-text interaction by generating a thesaurus which establishes the relationships between keywords and visual features. In all these approaches, the relationships or interactions between objects are considered as additional features and these features remain unchanged during the process.

Recently, many works in text retrieval and Web mining have indicated the effectiveness of iterative reinforcement among different data types for various applications. [14] proposes an iterative classification procedure which exploits the characteristic of relational data. [12] leverages the relationship between text and document to exploit the semantic similarity between terms. Wang et al. [19] use relationships among data objects to improve the cluster quality of interrelated data objects through an iterative reinforcement clustering process. Though effective, these methods suffer from the data sparseness problem. And [19] partly solves this problem by propagating the cluster centroid instead of the data points enclosed in the clusters of one data type. Xue et al. [21] improves retrieval effectiveness by iteratively “spread” the inter- and intra-object similarities through hyperlinks and click-through logs between queries and web-pages.

Our approach is motivated by [12][19][21] on similarity propagation. And we extend the idea of iterative reinforcement to multi-model image retrieval area. To make it feasible, we convert the two kinds of features (i.e. the low-level visual and textual features of images) to two types of objects (i.e. images and web-blocks) and use the block-image containerships to represent their relationships. In this way, the mutual reinforcement approach can be implemented on a single type of objects (i.e. images) other than multi-types of objects (such as web-pages, users, queries) required by the former approaches [19][21]. In contrast with the former linear combination approaches [2][4][7][8][17][20], our approach provides a non-linear combination of different modalities.

3. MULTI-MODEL ITERATIVE SIMILARITY PROPAGATION

We discuss our *iterative similarity propagation* approach in this section. First we give an overview of the procedure in Section 3.1, and formularize it in Section 3.2. The convergence of this approach is proved in Section 3.3.

3.1 Overview of the Approach

The basic idea of *iterative similarity propagation* is that the similarities among heterogeneous object types can mutually influence each other, by enhancing or reducing the similarity between two objects. Figure 2 shows an example. Let *T* and *S* denote two heterogeneous object spaces. Let t_i and s_j represent two specific objects in these spaces respectively. The dotted lines represent links among objects (i.e. inter-object relation) and the real lines represent similarities (i.e. intra-object relation). The length of the real line represents the degree of similarities. The left part shows the original object relationships, i.e. in space *S*, s_1, s_2, s_3 and s_4 are similar to each other, but are dissimilar to s_5.

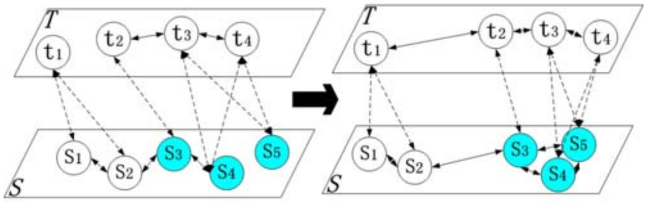


Figure 2. Sketch Map of Similarity Propagation

Assume that semantically s_5 is similar to s_3 and s_4 but s_1 and s_2 are dissimilar to s_3 and s_4 (as shown in the right part). Consider space *T*, s_4 and s_5 are both linked to the two objects t_3 and t_4 in *T* which are very similar. Although originally s_4 and s_5 are dissimilar based on their content features (i.e. the initial intra-object similarity), they obtain a certain similarity propagated from the similarity of t_3 and t_4. Likewise, because s_3 and s_4 are originally similar, so do their linked objects t_2, t_3 and t_4, their intra-object correlations are enhanced respectively after the similarity propagation.

On the other hand, although originally s_2 is similar to s_3 , but their linked objects, i.e. t_1 and t_2 are dissimilar, thus the similarity between s_2 and s_3 is reduced. Although t_1 and t_2 get similarity propagated from s_2 and s_3, it is weak because their similarity is totally depend on the decayed similarity propagated from s_2 and s_3.

When this process performs iteratively, the similarities of s_3 , s_4 and s_5 become stronger and stronger, while the similarity between s_2 and s_3 becomes weaker and weaker until converge. Hence the resulted object relationships turn to those in the right part of Figure 2, which better reflects the intrinsic similarities between objects.

3.2 The Algorithm

Without loss of generality, we describe the algorithm of *iterative similarity propagation* using two types of objects.

Assume the dimension of space *S* is *M* while the dimension of space *T* is *N*. Let $K_{M\times M}$ and $G_{N\times N}$ denote the intra-object similarity matrices based on the content features in space *S* and *T* respectively. Let $\hat{K}_{M\times M}$ and $\hat{G}_{N\times N}$ denote the intra-object similarity matrices after similarity propagation and both of which are normalized at the end of each iteration. Let $Z_{M\times N}$ be the link matrix from S to T (its transpose, i.e. Z' , is the link matrix from *T* to *S*) whose elements satisfy

$$Z_{ij} = \begin{cases} 1/\theta_i & \exists\, link\ s_i \to t_j \\ 0 & otherwise \end{cases} \qquad (1)$$

s_i and t_j denote the i^{th} and j^{th} data in *S* and *T* respectively. θ_i is the number of non-zero elements (i.e. the out-links of s_i) in the i^{th} row of Z .

The *iterative similarity propagation* process can be described as follows

$$\begin{cases} \hat{K} = \alpha K + (1-\alpha)\lambda Z\hat{G}Z' \\ \hat{G} = \beta G + (1-\beta)\lambda Z'\hat{K}Z \end{cases} \qquad (2)$$

where α and β are the weights. λ is a decay factor to ensure that the propagated similarities are weaker than the original similarities. $0 < \alpha, \beta, \lambda < 1$.

The physical meaning of equation (1) is obvious: $K, \hat{K}$ ($G, \hat{G}$) are the intra-object similarity matrices where K (G) is fully determined by the content feature of objects in S (T). $Z\hat{G}Z'$ ($Z'\hat{K}Z$) is inter-object similarity matrix, i.e. the part of intra-object similarities G (K) which are propagated from T (S) to S (T) through the links Z (Z'). And the similarities are decayed during this propagation of λ times.

Equation (2) combines both the intra- and inter-object similarities and addresses such mutual reinforcement in an iterative way. It points out that the similarities of one type of objects are affected by other types of objects related to them. It is, fundamentally, a non-linear combination method for the effects of different modalities on relational data.

Interestingly, the traditional single-modality image retrieval method and linear combination method can be seen as two special cases of equation (2):

1. If $\alpha\ (or\ \beta) = 1$, $\hat{K} = K\ (or\ \hat{G} = G)$, then equation (1) is reduced to the traditional single-modality retrieval method.
2. In the initial phase, if we set $\hat{K}^{(0)} = K\ \ (\hat{G}^{(0)} = G)$ and $\lambda = 1$, then (1) becomes the traditional linear combination method.

The superiority of our method to the traditional single-modality retrieval method and linear combination method is obvious. It is already proved by recent research that relational links between objects are helpful since they provide a unique source of information [10][13][16]. Hence it is almost definitely true that our method will surpass the traditional content-based image retrieval methods. Moreover, the interactions among heterogeneous objects are most probably non-linear, which can not be well approached by a simple linear combination method, no matter how optimal the weight set is selected [8][17][20].

3.3 Convergence of the Algorithm

In this section, we prove that the equation (2) will converge at the end. We denote the $\hat{K}$ and $\hat{G}$ in the *n*-th iteration as $\hat{K}^{(n)}$ and $\hat{G}^{(n)}$.

Proof:

Assume the process begins with the propagation from S to ***T***.

From equation (2), we have:

$$\begin{aligned} \hat{K}^{(n)} - \hat{K}^{(n-1)} &= (\alpha K + (1-\alpha)\lambda Z\hat{G}^{(n)}Z') - (\alpha K + (1-\alpha)\lambda Z\hat{G}^{(n-1)}Z') \\ &= (1-\alpha)\lambda Z(\hat{G}^{(n)} - \hat{G}^{(n-1)})Z' \end{aligned}$$

Because likewise,

$$\begin{aligned} \hat{G}^{(n)} - \hat{G}^{(n-1)} &= (\beta G + (1-\beta)\lambda Z'\hat{K}^{(n-1)}Z) - (\beta G + (1-\beta)\lambda Z'\hat{K}^{(n-2)}Z) \\ &= (1-\beta)\lambda Z'(\hat{K}^{(n-1)} - \hat{K}^{(n-2)})Z \end{aligned}$$

Replace $\hat{G}^{(n)} - \hat{G}^{(n-1)}$ in the equation of $\hat{K}^{(n)} - \hat{K}^{(n-1)}$, we obtain that

$$\begin{aligned} \hat{K}^{(n)} - \hat{K}^{(n-1)} &= (1-\alpha)(1-\beta)\lambda^2 ZZ'(\hat{K}^{(n-1)} - \hat{K}^{(n-2)})ZZ' \\ &\overset{\omega=(1-\alpha)(1-\beta)\lambda^2,\ A=ZZ'}{=} \omega A(\hat{K}^{(n-1)} - \hat{K}^{(n-2)})A \\ &= \cdots \\ &= \omega^{n-1}A^{n-1}(\hat{K}^{(1)} - \hat{K}^{(0)})A^{n-1} \\ &\overset{\hat{K}^{(0)}=K}{=} \omega^{n-1}A^{n-1}(\hat{K}^{(1)} - K)A^{n-1} \end{aligned}$$

Denote $A = \left[a_{ij}\right]_{M\times M}$, because of $A = ZZ'$, according to the definition of Z given in equation (1), we have

$$a_{ij} = \begin{cases} \dfrac{\min(\theta_i,\theta_j)}{\theta_i\theta_j} = \dfrac{1}{\max(\theta_i,\theta_j)} \le 1 & \theta_i, \theta_j > 0 \\ 0 & \theta_i = 0\ or\ \theta_j = 0 \end{cases}$$

Hence we have $A^{n-1} \xrightarrow{n\to\infty} 0$.

On the other hand, because $\hat{K}^{(1)} - K$ is a constant matrix and $\omega < 1$, we have

$$\hat{K}^{(n)} - \hat{K}^{(n-1)} \xrightarrow{n\to\infty} 0,$$

which proves the convergence of equation (1).

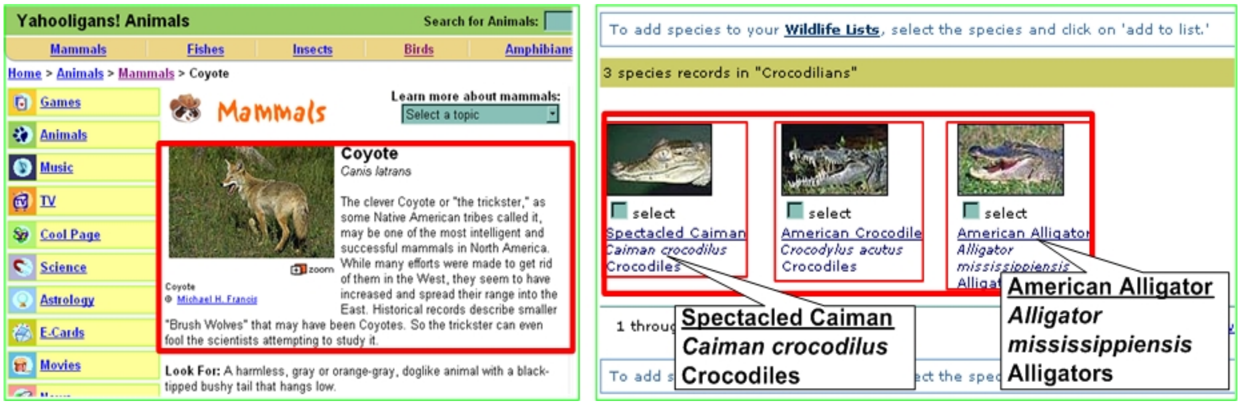
4. IMAGE RETRIEVAL USING ITERATIVE SIMILARITY PROPAGATION

We have detailed in Section 3 the *iterative similarity propagation* algorithm. In this section, we present how we improve the performance of image retrieval using this mutual reinforcement approach.

4.1 Link Graph Construction

We crawled 10,628 images mainly from the websites listed in Table 1, and use an effective page segmentation technique called VIPS (VIsion-based Page Segmentation) [5][6] to obtain the "blocks", i.e. the web-page segments that contain these images as well as their surrounding texts. VIPS 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. The red rectangles in Figure 3 show some examples of the "blocks" obtained by [6].

Figure 3 shows three examples of image-block relationship. The left web-page shows a one-to-one projection between an image and a block. The right web-page shows a more-to-one and a one-to-more projections between images and blocks. That is, a block can contain multiple images (see the thick red rectangle) and an image can belong to more than one block (see the



Coyote
Canis latrans

The clever Coyote or "the trickster," as some Native American tribes called it, may be one of the most intelligent and successful mammals in North America. While many efforts were made to get rid of them in the West, they seem to have increased and spread their range into the East. Historical records describe smaller "Brush Wolves" that may have been Coyotes. So the trickster can even fool the scientists attempting to study it.

Look For: A harmless, gray or orange-gray, doglike animal with a black-tipped bushy tail that hangs low.

To add species to your **Wildlife Lists**, select the species and click on 'add to list.'

3 species records in "Crocodilians"

Figure 3. Examples of Image-Block Relationship

crocodile images. All of them are inside both the thick red Rectangle and a thin red rectangle).

We refer the block with its image being removed as t-block and treat the images and their t-blocks as two types of objects. The containerships between blocks and their images are considered as links between the t-blocks and the images. The images are represented by a set of low-level color and texture features. The content features of t-blocks are extracted in such a way: first, the image surrounding texts, image captions and hyperlinks are parsed from the HTML documents. Then stop-words are filtered out and the rest of the terms construct term vectors which are weighted using TF*IDF [15]. The resulted term vectors are used as the t-block features. When a block contains more than one image, the content feature of its t-block is obtained from the collection of the textual annotations of all images inside it. In this way, the textual annotations of images are transmitted to another type of object other than an additional feature vector.

Each of the three image-block relationships mentioned above has simultaneously its advantages and disadvantages. For example, although the content features of the large block in the right web-page are less precise than those of the left one, this block points out that two images are similar although they have different textual annotations (see the “alligator” and “caiman” in Figure 3, both of them are crocodiles).

Based on these three kinds of relationships (i.e. one-to-one, one-to-more and more-to-one), a similar link graph as in Figure 2 between the images and the blocks can be established: the nodes in *S* and *T* are the images and blocks respectively. If an image is contained in a certain block, there will be a link (i.e. the dotted line) between them. Based on this graph, the intra- and inter-object similarities can be propagated inside and across the images and blocks.

4.2 Content Similarity Matrix Formulation

Let X be the visual feature matrix with rows as the images and columns as their visual features. Let X_i denote the i^{th} row of X . Let Y be the block feature matrix with rows the blocks and the columns the terms (weighted by TF*IDF). Y_j represents the j^{th} row of Y .

The initial image similarity matrix $K=\left[K_{ij}\right]_{M\times M}$, which is totally based on the low-level visual features, is given by converting the Euclidean distances between images into similarities which monotonically increase as the distances decrease. K is given by

$$K_{ij}=1-\frac{Eud(X_i,X_j)}{\max\limits_{i,j} Eud(X_i,X_j)}=1-\frac{\sqrt{(X_i-X_j)(X_i-X_j)^T}}{\max\limits_{i,j}\sqrt{(X_i-X_j)(X_i-X_j)^T}} \qquad (3)$$

The initial block similarity matrix $G=\left[G_{ij}\right]_{N\times N}$ is calculated using the traditional cosine similarity measure in text retrieval which is given by:

$$G_{ij}=\frac{Y_i\bullet Y_j}{\|Y_i\|\cdot\|Y_j\|} \qquad (4)$$

We set the initial intra-object similarities to be their content similarities, i.e. $\hat{K}^{(0)}=K$ and $\hat{G}^{(0)}=G$. And we perform the iterative reinforcement using equation (2).

5. EXPERIMENTS

10,628 images associated with 16,720 blocks are crawled mainly from the websites listed in Table 1. These images cover from nature to artificial objects, human beings and Web logos.

Table 1. Website List of Our Image Database

www.yahooligans.com/content/animals/mammals/	
www.naherpetology.org	www.enature.com
www.homeearth.com	www.pbs.org
www.visualsunlimited.com	www.ups.edu
www.tomvezo.com	www.bbc.co.uk
www.turnerphotographics.com	www.birdsasart.com
www.kevinschafer.com	users.1st.net
www.briansmallphoto.com	spaceflightnow.com
amazing-space.stsci.edu	www.space.com
www.suziophoto.com	www.sunfarm.com
www.thebackyardbirdwatcher.com	quest.arc.nasa.gov
www.worldwildlife.org	www.nwf.org

Ten volunteers are asked to manually label the ground truth. The animals and plants images were labeled in accordance with the category information in www.enature.com and yahooligan.yahoo.com. For the other images that have no category information on web-pages, we let the volunteers select the most representative keywords as the labels.

The image visual features are 36-bin color correlogram [11], three-level color moment [18] and three-level wavelet textures [1].

Two performance measures: precision-scope and recall-scope, are applied. Scope specifies the number of images returned to the user. Precision is defined as the number of retrieved relevant objects over the value of scope. Recall is defined as the number of retrieved relevant objects over the total number of relevant objects.

We use the alike linear combination method proposed in [8] as our baseline method but tune an optimal weight set for it rather than fix each weight to 0.5 as proposed in [8]. Although the low-level visual features we used are different from [8], these differences will not bias the final evaluation since it is the method itself rather than features used that determine the performances.

The reason that we choose the method proposed in [8] as our baseline method is as follows. First, the approach in [8] represents a traditional way of combining multi-modalities for image retrieval. Second, we do not involve a training phase to select representative query set and learn the optimal weight set for them, nor do we apply face detection and face recognition approaches in our method, hence it is impossible to compare our approach with [17]. Third, our method is based on global images, while the approaches in [2][4] are based on segmented images. The works in [2][4] are more like image auto-annotation and recognition.

We randomly selected 2,500 images to form the query set. And the final performances are the average precision and recall on these 2,500 images.

5.1 Performance Evaluation

Figure 4 shows the cooperation of retrieval performance. The red diamond lines represent the retrieval performance of our method. The blue square lines correspond to the baseline method, i.e. the linear combination method. The yellow and green lines show the performance of single-modality method. The yellow triangle lines are based on only textual features, and the green dot lines are corresponding to the method using only image low-level features.

The parameters selected are $\alpha = 0.3, \beta = 0.8, \lambda = 0.8$ in our method, where α and β are the weights of image similarity matrix and t-block similarity matrix respectively. λ is the decay factor. In the baseline method, the weight of similarity matrix based on visual features is 0.2. The parameters are determined based on an extensive experiment which will be discussed in Section 5.3.

It can be seen from this figure that our method significantly outperforms the baseline method and the single-modality method. The average precision@10 for these four methods, i.e. the *iterative similarity propagation* method, the linear combination method, the textual feature based retrieval method and the visual-feature based retrieval, are 46.3%, 37.1%, 32% and 23.2% respectively.

From this figure, we can see that retrieval using only textual features surpasses greatly the method using only image content feature. It on the one hand proves the effectiveness of VIPS [6] web-page segmentation algorithm, and on the other hand, confirms that the existence of semantic gap greatly affects the performance of content-based image retrieval.

However, although the textual features are better than image content features, still many web images will have noisy annotations. Also, textual annotations can be ambiguous, e.g. "apple" can both indicate "apple tree" and "apple computer". Hence, intuitively, combining the two kinds of features will do a better job, just as indicated in [3] that "while text and images are separately ambiguous, jointly they tend not to be". And our experiment doubly confirmed this.

However, separately combining different features can also be biased by the features themselves, as mentioned in Section 1. The iterative propagation approach better explores the mutual reinforcement among different data types which in some sense correct such biases. It can also be regarded as a non-linear combination method on different feature types.

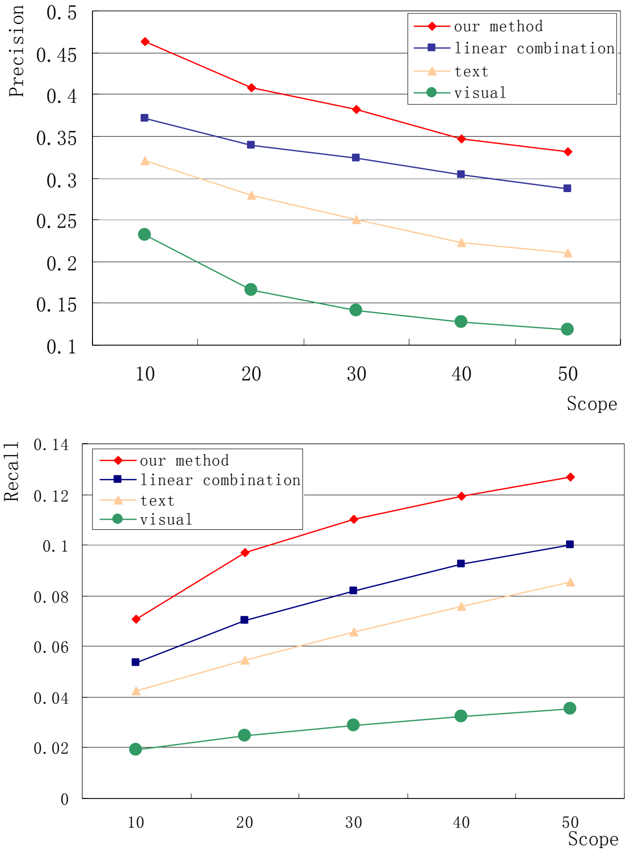


Figure 4. Performance Evaluation

5.2 The Convergence of Our Approach

It is proved in Section 3.3 that our approach will finally converge. In this section, we show the empirical result on the convergence test.

Note that the evaluation is independent with parameter selected. In our evaluation, we set $\alpha = 0.3, \beta = 0.8, \lambda = 0.8$.

The precision vs. number of iteration is shown in Figure 5. The method converges when the number of iteration is 3 (the precision@2 is a bit higher than precision@4). This figure proves the convergence of our approach as analyzed in Section 3.3.

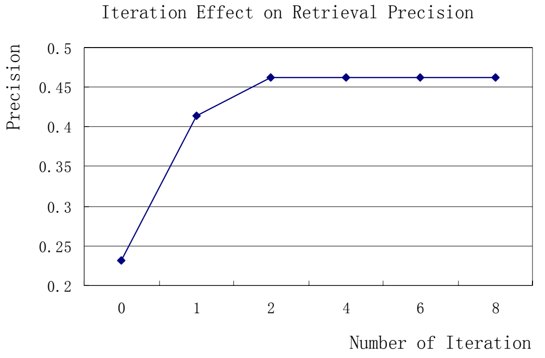


Figure 5. Convergence of Our Algorithm

5.3 The Effect of Decay Factor

As mentioned in Section 3.2, the decay factor λ ensures that the propagated similarities are weaker than the original similarities. In this section, we evaluate its effect on the performance of image retrieval.

The dataset and the value of α and β are the same as that in Section 5.1. Figure 6 shows the retrieval precision for different λ . It can be seen that the best performance is obtained when $\lambda = 0.8$. This is reasonable because too much or too less propagation will both degrade the retrieval performance. When $\lambda = 0$, this method is reduced to the single-modality image retrieval (i.e. CBIR), in which case the mutual reinforcement is not taken into consideration. When $\lambda = 1$, it becomes the linear combination method, in which case the mutual reinforcement is not fully explored, and the result lean to be biased by the features of data as discussed in Section 1.

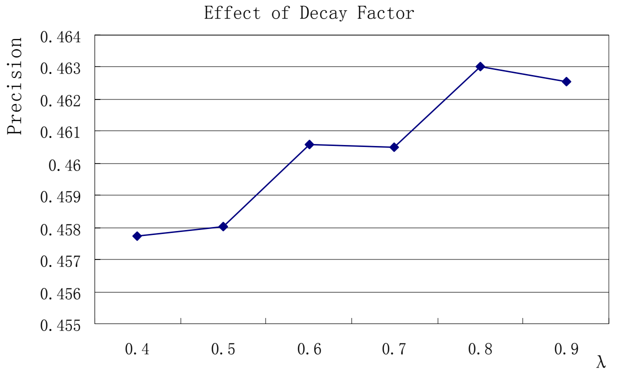


Figure 6. Effect of Decay Factor λ

5.4 Precision vs. Weighting Schema

Different weight α, β in equation (2) will affect the retrieval performance. Figure 7 shows the variation of retrieval precision vs. α with $\beta = 0.8$. It can be seen that the best performance is achieved at $\alpha = 0.3$.

The best performance is obtained when $\alpha < \beta$ shows that when calculating the similarities of images (i.e. $\hat{K}$ in equation (2)), the similarities based on image visual features (i.e. K) are less effective than the similarities propagated from the t-blocks (i.e. $\hat{G}$). This coincides with the retrieval performance given in Figure 4, where the retrieval performance based on text retrieval is far better than that based on image contents.

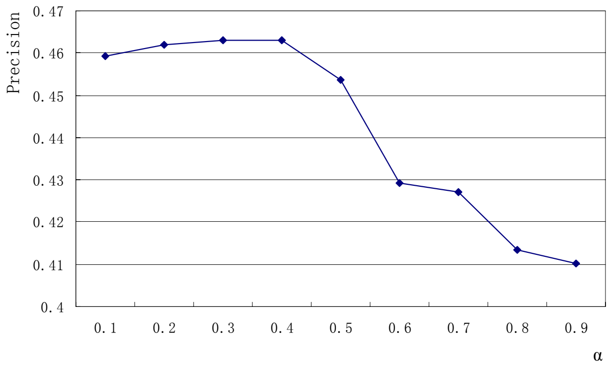


Figure 7. Precision vs. Visual Feature Based Image Similarity Matrix Weighting Schema

6. CONCLUSION

Multi-model Image Retrieval intends to deal with several data types which is heterogeneous and inter-active. How to seamlessly combine different retrieval models is still a research topic. In this paper, we proposed an *Iterative Similarity Propagation* model to solve this problem. It attempts to fully exploit the mutual reinforcement of relational data which result in a non-linear combination of different modalities. It uses the intra-object similarities of one data type to affect those of another data type which links to it, and perform this approach iteratively, by which the similarities of images in the semantic level are approached. The assumption is that, if two objects of the same type are both related to an object of another type, these two objects are similar; and if two objects of the same type are related to two different, but similar objects in another type, then to some extent, these two objects can also be considered similar.

In this paper, this approach is used to learning the semantic similarities of images by leveraging the relationships between Web images and their textual annotations. The experimental results based on 10,628 images crawled from the Web showed the effectiveness of our proposed *Iterative Similarity Propagation* model for image retrieval.

In fact, the importance of each web page is different as well as their blocks. In the future, we will combine importance of the image and the text into our algorithm according to their blocks. Moreover, we did not discuss the integration of the proposed method to a relevance feedback system. A simple way could be weighting the content feature matrix K and G in equation (2) in each iteration according to users' feedbacks. We will research on this also in our future works.

7. ACKNOWLEDGEMENTS

Special thanks should be given to Deng Cai, Xuemei Jiang and Shen Huang for their sincerely helps.

8. REFERENCES

[1] Antonini, M., Barlaud, M., Mathieu, P., and Daubchies, I., Image Coding Using Wavelet Transform, *IEEE Trans. on Image Processing*. 1(2). (Apr. 1992). 205-220.

[2] Barnard, K., Forsyth, D. Learning the Semantic of Words and Pictures. *ICCV* (2001).

[3] Barnard, K., Duygulu, P., and Forsyth, D. Clustering Art. *CVPR* (2001), 434-439

[4] Blei, D.M., and Jordan, M.I. Modeling Annotated Data. *SIGIR* (2003).

[5] Cai, D., Yu, S.P., Wen, J.R., and W.-Y. Ma, Extracting Content Structure for Web Pages Based on Visual Representation, *APWeb* (2003), 406-417

[6] Cai, D., Yu, S.P., Wen, J.R., and W.-Y. Ma, VIPS: a Vision-Based Page Segmentation Algorithm, *Microsoft Technical Report* (2003), MSR-TR-2003-79

[7] Cascia, M.L., Sethi, S., and Sclaroff, S. Combining Textual and Visual Cues for Content-based Image Retrieval on the World Wide Web. *Proc. IEEE Workshop on Content-Based Access of Image and Video Libraries*, (1998)

[8] Chen, Z., Liu, W.Y., Zhang, F., Li, M.J. and Zhang, H.J. Web Mining for Web Image Retrieval, *Journal of the American Society for Information Science and Technology*, 52(10), (2001), 831--839.

[9] Duffing, G. Text-Image Interaction for Image Retrieval and Semi-Automatic Indexing. *20th Annual BCS-IRSG Colloquium on IR*, (1998)

[10] Gibson, D., Kleinberg, J., and Paghavan, P. Inferring Web Communities From Link Topology. In *Proc. 9th ACM Conference on Hypertext and Hypermedia*, (1998), 225-234

[11] Huang, J., Kumar, S.R., Mitra, M., Zhu, W.J. and Zabih R. Image Indexing Using Color Correlograms, *In Proc. IEEE Conference on CVPR.*, (1997) 762--768

[12] Kandola, J., Shawe-Taylor, J., Cristianini, N. Learning Semantic Similarity. *NIPS* (2002)

[13] Kleinberg, J. Authoritative Sources in a Hyperlinked Environment. In *Proc. of the 9th ACM-SIAM Symposium on Discrete Algorithms*, (1998)

[14] Neville, J., and Jensen, D. Iterative Classification in Relational Data. *Proceedings of the AAAI 2000 Workshop*, AAAI Press. (2000). 42-49.

[15] Salton, G., and Buckley, C. Term Weighting Approaches in Automatic Text Retrieval. *Information Processing and Management*, 24(5). (1988). 513--523.

[16] Slattery, S., and Craven, M. Combining Statistical and Relational Methods in Hypertext Domains, In *Proc. ILP.* (1998)

[17] Srihari, R.K., Rao, A.B., Han, B., Munirathnam, S., and Wu, X.Y. A Model For Multimodal Information Retrieval. *ICME* (2000)/.

[18] Stricker, M., and Orengo, M. Similarity of Color Images, *In Storage and Retrieval for Image and Video Databases III*, SPIE 2420, (Feb. 1995), 381--392

[19] Wang, J.D., Zeng, H.J., Chen Z., Lu, H.J., Tao, L., and Ma, W.Y. ReCoM: Reinforcement Clustering of Multi-Type Interrelated Data Objects. *SIGIR* (2003).

[20] Westerveld, T. Probabilistic Multimedia Retrieval. *SIGIR* (2002)

[21] Xue, G.R., Zeng, H.J., Chen Z., Ma, W.Y., and Yu Y. Similarity Spreading: A Unified Framework for Similarity Calculation of Interrelated Objects. *WWW* (2004)

[22] Zhou, X.S., and Huang T.S. Unifying Keywords and Visual Contents in Image Retrieval. *IEEE MultiMedia* 9(2) (2002), 23-33