

# Visual-Semantic Graphs: Using Queries to Reduce the Semantic Gap in Web Image Retrieval

Barbara Poblete†  
Marcelo Mendoza†

Benjamin Bustos‡  
Juan Manuel Barrios‡

†{bpoblete,mendozam}@yahoo-inc.com, Yahoo! Research, Santiago, Chile  
‡{bebustos,jbarrios}@dcc.uchile.cl, Department of Computer Science, University of Chile, Santiago, Chile

## ABSTRACT

We explore the application of a graph representation to model similarity relationships that exist among images found on the Web. The resulting similarity-induced graph allows us to model in a unified way different types of content-based similarities, as well as semantic relationships. Content-based similarities include different image descriptors, and semantic similarities can include relevance user feedback from search engines. The goal of our representation is to provide an experimental framework for combining apparently unrelated metrics into a unique graph structure, which allows us to enhance the results of Web image retrieval. We evaluate our approach by re-ranking Web image search results.

## Categories and Subject Descriptors

H.3.3 [Information Storage and Retrieval]: Information Search and Retrieval; H.2.8 [Database Applications]: Image Databases

## General Terms

Algorithms, Experimentation, Performance

## Keywords

Web Image Search, Web Image Re-ranking, Query Log Analysis, Content-based Image Features

## 1. INTRODUCTION

A key challenge in Web image retrieval is to efficiently combine the two most important types of image features: visual content-based features and semantic features. This problem poses additional questions such as: *which are the correct image descriptors to use in Web image retrieval?* and *which semantic features are the most appropriate for this task?* Moreover, *which combination of visual and semantic features works best?* For example, there are several image descriptors which reflect different image qualities based on their contents. Additionally, many semantic features associated to images are not always reliable.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.

CIKM'10, October 26–30, 2010, Toronto, Ontario, Canada.  
Copyright 2010 ACM 978-1-4503-0099-5/10/10 ...\$10.00.

In spite of the existence of general frameworks for different object similarity integration [11, 10, 3, 9, 8], these do not provide the recipe for successful combination in specific domains. Following this motivation, we present an experimental framework for relating similarity metrics of visual and semantic nature in Web image retrieval. The main contributions of our work are the experimental framework (the *visual-semantic graph*), a methodology for assessing unbiased Web image relevance based on click-through data from query logs, using time windows, and an evaluation over a large-scale real world data set.

## 2. THE VISUAL-SEMANTIC GRAPH CONCEPT

We present two types of similarity graphs that can be used to model relationships among images on the Web: the *visual similarity graph* and the *semantic similarity graph*. We discuss how to aggregate in a unified manner the information contained by them.

### 2.1 Visual Similarity Graph

We define a *visual similarity graph* as an undirected graph that represents content-based similarity relationships in a collection of images. The nodes of this graph correspond to images, and the edges of this graph connect images that are similar (given an image descriptor and a similarity measure  $\delta$ ). Each edge has an associated weight: the larger this weight, the more similar the connected images are.

The visual similarity graph is composed of a collection of images and an image descriptor with an associated distance measure (which is used to compute the weights of the edges). While our approach is not restricted to any particular descriptor, we implemented three different image descriptors for computing visual similarity graphs, listed below.

- *Edge Histogram Descriptor (EHD)* [6]. It captures the spatial distribution of edges. The EHD converts the input image to gray scale, partitions it into  $N_x \times N_y$  sub-images, divides each sub-image into  $M_x \times M_y$  blocks, further partitions each block into  $2 \times 2$  sub-blocks, computes the average intensity of the pixels for each sub-block, and applies an edge detector to each block. The edge detector is comprised of  $k$  different filters of  $2 \times 2$  pixels. If the filter with maximum strength exceeds a certain threshold  $t$ , the block is marked as an edge block. Finally, EHD computes an edge histogram for each sub-image. For this work, we set  $N_x = N_y = 4$ ,  $M_x = M_y = 8$ ,  $t = 5$ , and  $k = 5$ . This descriptor requires a minimum width and height of 64 pixels for the input image. We use the  $L_1$  metric (Manhattan distance) as similarity function  $\delta$ .

- *Color Histogram (HSV)*. According with the Scalable Color Descriptor (SCD) [6], we used a uniform quantization of the HSV (hue-saturation-value) color space into 256 bins (16 levels for H, four levels for S, and four levels for V). We use  $L_1$  as similarity function  $\delta$ .
- *Ordinal Measurement Descriptor (OMD)* [1]. It partitions the input image into  $B_x \times B_y$  blocks, and for each block it computes the average intensity. Then, it sorts the set of average intensities in ascending order. The final descriptor corresponds to the set of ranks assigned to each block. For the present work, we set  $B_x = B_y = 9$ . We use the Hamming distance as similarity function  $\delta$ .

For each descriptor, we compute its associated maximum distance  $M$  (the largest distance between two images using that descriptor). We are only interested in connecting very similar images in the visual similarity graph. Thus, we define a threshold value  $\tau$  that indicates us which images must be connected. If the distance between images is smaller or equal than  $\tau$ , the nodes associated to those images are linked.

Therefore, let  $I$  be the set of images. We define the visual similarity graph  $G_\nu = (I, E)$ , where  $E$  is the set of edges of  $G_\nu$ . An edge  $(i, j) \in E$  is defined if  $\delta(i, j) \leq \tau$ . To each edge  $(u, v)$ , we associate a weight  $w(i, j) = \delta(i, j)$  that represents the content-based similarity between both images.

## 2.2 Semantic Similarity Graph

We define the *semantic similarity graph* as an undirected bipartite graph that represents semantic-based similarity relationships between a collection of *term-sets* and a collection of images. The edges in this graph connect term-sets with images that have a semantic relationship with them. Each edge has a weight associated to it which is a measure of the relevance of the term-set to the connected image. Formally, let  $T = t_1, \dots, t_m$  be a set of unique term-sets defined by users. Let  $I$  be the set of relevant images for term-sets in  $T$ . We define the semantic similarity graph as  $G_S = (T \cup I, E)$ , where  $E$  is the set of edges which connect nodes in  $G_S$ . An edge  $(t, i) \in E$  exists if at least one image  $i \in I$  is considered relevant to the term set  $t \in T$ . Each edge has an associated weight  $w(t, i)$  which corresponds to the *importance* of  $i$  to  $t$ .

In particular, we consider the *click graph* as our semantic similarity graph. The click graph is a bipartite graph of queries and images which denotes user searching behavior extracted from a search engine query log. Edges in this graph connect queries to the images which users selected in their searches. To apply the definition of the semantic graph to the click graph we must make the following considerations:

- $T = Q$ , where  $Q$  is the set of unique queries submitted to the search engine during the period which expands the query log.
- $I$  corresponds to the set of Web images which have been clicked by users after formulating a query in  $Q$ .
- An edge  $(t, i) \in E$  exists if at least one user clicked on an image  $i$  after submitting the query represented by  $t$ .
- The weight  $w(t, i)$  is defined as the number of unique session clicks registered in the query log from  $t$  to  $i$ .

Our selection of the click graph in this case is related to two characteristics which make it appropriate: 1) It gives a measure of relevance of term-sets to images, which is the click frequency, 2) it conveys user-relevance feedback, i.e. users select (click) on images which match their information need.

## 2.3 The visual-semantic graph

We define the visual-semantic graph  $G_{\nu S}$  as the union of the visual similarity graph and the semantic graph. There is an undirected weighted edge between two images  $i_1$  and  $i_2$  of weight  $w(i_1, i_2)$  if both images are similar according to the visual similarity graph. There is an undirected weighted edge between a term-set  $t$  and an image  $i$  if there is a user defined semantic relationship between  $q$  and  $i$ . The weight of this edge is given by  $w(t, i)$ .

## 3. RANDOM-WALK PROCESS

In this section we describe the random-walk process for the visual-semantic graph. According to the definitions introduced in our prior work [7], a graph of  $N$  nodes is described by an  $N \times N$  matrix  $P$  of transition probabilities, where an entry  $P_{i,j}$  represents the transition probability between the nodes  $i$  and  $j$ . A row vector  $\pi^T$  represents the stationary distribution over the graph after performing a random-walk process. After  $k$  iterations of the process, the equation  $\pi^{(k)T} = \pi^{(k-1)T} \cdot P$  is satisfied, where  $\pi^{(k)T}$  and  $\pi^{(k-1)T}$  represents the vector  $\pi^T$  calculated at iterations  $k$  and  $k - 1$ , respectively. Under certain conditions of the process (irreducibility, finiteness and aperiodicity) the vector  $\pi^{(k)T}$  converges to  $\pi^T$ . Then, the  $i$ -th coordinate of  $\pi^T$  corresponds to the frequency with which a random surfer visits the  $i$ -th node of the graph during the process.

**Random-walk on a visual similarity graph.** In this process a user begins its image viewing process by selecting a random image from the collection. After viewing this image the user uses it to select a second image to view, selection which is biased by the degree of similarity between the first image and the second ( $w(i, j) = \delta(i, j)$ ). This process is repeated iteratively until the user begins a new search from a different image with probability  $(1 - \alpha)$ .

Let  $A_\nu$  be the adjacency matrix of the visual graph  $G_\nu$ , in which the entry  $(i, j)$  has the value of  $w(i, j) = \delta(i, j)$ . Let  $N_\nu$  be the row-normalized version of  $A_\nu$ . The transition-probability  $P_\nu$  of the similarity graph is given by:

$$P_\nu = \alpha N_\nu + (1 - \alpha)\mathbf{1},$$

where  $\alpha$  is the dumping factor of the process and  $\mathbf{1}$  is a matrix that has the value  $\frac{1}{N}$  in all its entries (teleportation matrix).

**Random-walk on the semantic graph.** This process corresponds to a random-walk on an undirected bipartite graph  $T \times I$ , where  $T$  represents the query terms and  $I$  the set of clicked images. Let  $A_S$  be the  $M \times N$  adjacency matrix of the semantic graph, whose  $M$  rows correspond to the terms of  $T$  and the  $N$  columns corresponds to the images of  $I$ . Each entry  $(t, i)$  in this matrix has a value  $w(t, i)$ , which corresponds to the query-to-image click frequency found in the query log. The transpose, denoted by  $A_S^T$ , models the fact that it is possible to go back from an image to a query.  $A'_S$  represents the  $(N+M) \times (N+M)$  adjacency matrix that considers both situations:

$$A'_S = \begin{pmatrix} A_S & 0 \\ 0 & A_S^T \end{pmatrix}.$$

Let  $N_S$  be the row normalized version of  $A'_S$ . Then the random-walk process on the semantic graph is given by:

$$P_S = \alpha N_S + (1 - \alpha)\mathbf{1}.$$

**Random-walk on the visual-semantic graph.** We combine both graphs performing a convex union:

$$\beta N_S + (1 - \beta)N_\nu,$$

where  $\beta$  is the probability of the user choosing a text-based image retrieval system, as opposed to the content-based image system. Then, the random-walk process over the visual-semantic graph is defined as follows:

$$P_{vS} = \alpha(\beta N_S + (1 - \beta)N_v) + (1 - \alpha)\mathbf{1}.$$

Haveliwala and Kamvar [4] showed that the convergence of the random-walk process depends on the second eigenvalue of the transition matrix which corresponds to the dumping factor. When  $\alpha$  increases, the convergence rate decreases. They showed also that a good balance between the convergence rate and the forced behavior introduced by the teleportation is achieved when  $\alpha = 0.85$ . The effect of the choice of  $\beta$  will be evaluated in the following section.

## 4. EVALUATION

We present an experimental evaluation of our approach over a large-scale dataset. We evaluate by re-ranking search results at query level. We re-rank using the stationary distribution scores obtained for random-walks on the visual-semantic graph. The goal of our evaluation is to find a combination of visual similarity and semantic graphs that provide additional information than either graph on its own.

**Dataset.** The query log used in the evaluation was obtained from Yahoo! image search. Experiments were performed over a two-weeks period, from March 1st 2010 to March 13 2010. We consider the first week to build the dataset and the second week for evaluation purposes. Each week contains approximately 7 million unique images, with a 4.4 million images repeated in both weeks. Overall, each week registered around 11.2 million unique-session clicks on images. Additionally, 2.7 million queries were repeated in the first and the second week of data. We used these queries for the re-ranking experiment.

### 4.1 Graph Generation

Originally, the image collection was intended to include all of the clicked images in the query log. Nevertheless, due to the complexity involved in the generation of the visual similarity graph, for our evaluation we select a random sample of these images (1/3 of the original collection). Therefore, we generate the visual similarity graph over this reduced image set. This optimization still allows us to validate our proposal.

On the other hand, we keep all of the click-through information, considering all of the clicked images and queries in the log to build the semantic similarity graph (in this case, the click graph).

**Visual similarity graph generation.** We create the visual similarity graph using an incremental algorithm that uses a *pivot-based approach* [12] to find similar images. We use *Sparse Spatial Selection* (SSS) index structure [2] to compute efficiently the pairs of similar images. The algorithm is as follows:

- i. Add a new image  $x$  of the collection to the SSS index. Create its corresponding node  $u$  in the similarity graph.
- ii. Compute a range query using the index. This finds all indexed images at distance smaller or equal than  $\tau$  to  $u$ .
- iii. Add the correspondent links to the graph (from  $u$  to the nodes of the similar images to  $x$  computed at step (ii)), storing their associated distances. If all images were processed proceed to step (iv), else repeat from step (i).
- iv. After all images have been processed, the last step is the computation of the weights of the edges. Let  $M$  be the

maximum distance in the descriptor’s space. For each node  $u$ , the weight  $C(u, v) = M - \delta(u, v)$  is associated to the edge  $(u, v)$ . After computing all the weights, normalize the weights such that the sum of all edges connecting  $u$  is 1.

Although in the worst case the computation time for computing the similarity graph using the proposed algorithm is  $O(n^2)$  (as a brute force algorithm), in practice we obtained large reductions in the processing time (except for OMD), as Table 1 shows.

	EHD	HSV	OMD
Naive	323.2	543.0	77.0
Pivot-based	8.1	11.1	166.3
$\alpha$	0.54	0.76	1.00

**Table 1: Time (in hours) needed to compute the similarity graph for each image descriptor.**

**Semantic graph generation.** The nodes of the click graph are all of the unique queries and all of the images recorded in the query log. We consider only queries which register at least one click on an image. The weight in an edge  $(t, i)$  is the number of clicks from different sessions from the query  $t$  to the image  $i$ .

### 4.2 Re-Rank Evaluation

In this section we evaluate if different descriptors provide different amounts of information for the Web image retrieval task. Furthermore, we also evaluate how different values of  $\beta$  affect the results of the visual-semantic graph. To do this, we use the stationary distribution scores of each combination (using the different image descriptors and  $\beta$  values) to re-rank search results. In this stage, it is important to note that we use two datasets, one for computing the visual-similarity graphs (1st week in the query log) and the next time window for computing our *gold standard* or *ideal rank*. Therefore we are evaluating how well our approach performs with new data (which was not used for “training”). Equally important is the fact that we generate “global” stationary distribution scores, which are then used to generate “local” re-ranking (at query level). We use the re-ranking induced by the click graph as our baseline ( $\beta = 1$ ).

Figure 1 shows that the original ranking induces a click-bias. Positions farther down the list of responses consistently concentrate fewer clicks than the first positions. Therefore, an interesting observation is that the click-bias is the same as would be expected if image results were displayed in a vertical listing, instead of a matrix-like interface.

To avoid the effect of the click-bias for our evaluation purposes, we calculate the fraction of clicks that each image concentrates over the total number of clicks related to a given query. Figure 2 shows that using this measure (instead of clicks), we significantly decrease the image click-bias that existed for each query.

To evaluate the perform of each descriptor we compute the Normalized Discounted Cumulative Gain measure (NDCG for short) [5] which consider an explicit position discount factor in its definition. We calculate this measure is calculated at query level. Then, the values obtained from the NDCG measure are are averaged across the queries at each rank value.

We combine each visual graphs with the semantic graph (click graph) using  $\beta = 0, 0.25, 0.50, 0.75, 1$ . It should be noted that a value of  $\beta = 0$  creates a graph with only the visual similarity graph, and that  $\beta = 1$  corresponds to using only the click graph. We obtain 15 possible visual-semantic graphs for evaluation. The results are displayed in Table 2.

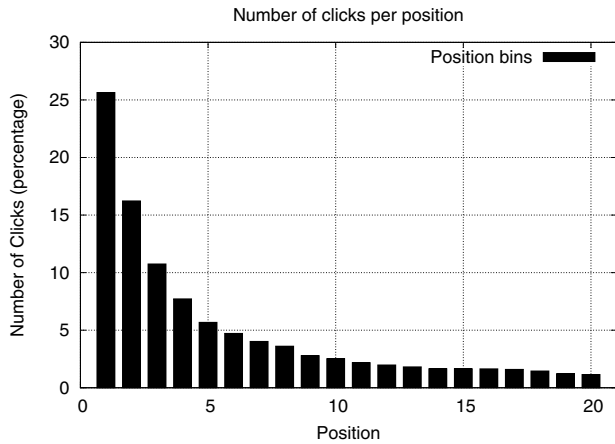


Figure 1: Number of clicks per position (top-20 results).

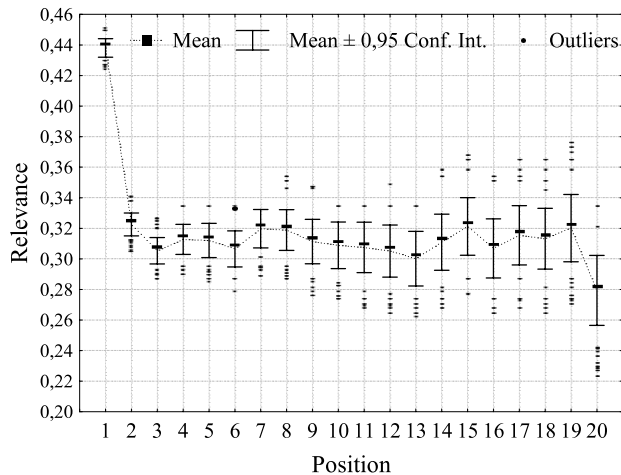


Figure 2: Relevance distribution per position (top-20 results).

Table 2 shows that the best combination with the click graph is reached when we use  $\beta = 0.5$ . This is an important result because it allows us to claim that the combination of the visual similarity graph and the semantic graph works better than either graph on its own. In fact, we can observe that the combination of the click graph (shown for  $\beta = 1$ ) with any visual similarity graph improves its results. The best results are obtained when the click graph is combined with the OMD(0) graph, showing that the unification of the best graphs produces also the best combination. Moreover, the combination of the click graph and the OMD(0) graph improves the re-ranking induced by the click graph by more than 5%.

## 5. CONCLUSIONS & FUTURE WORK

We have presented a new type of graph that combines visual and semantic characteristics that are useful for web image retrieval. Performing a random-walk process over this graph and using the steady-state probability distribution as scores for image re-ranking, our experiments show that it is possible to improve over 5% a baseline. We have also shown that not all combinations of visual fea-

$\beta$	0	0.25	0.5	0.75	1
rank	EHD				
1	0.235	0.539	0.545	0.541	0.540
2	0.310	0.616	0.623	0.617	0.616
3	0.388	0.687	0.692	0.687	0.687
4	0.481	0.760	0.766	0.761	0.760
5	0.595	0.842	0.847	0.842	0.841
rank	HSV				
1	0.326	0.544	0.549	0.542	0.540
2	0.362	0.619	0.624	0.618	0.616
3	0.430	0.689	0.694	0.688	0.687
4	0.518	0.762	0.767	0.761	0.760
5	0.630	0.843	0.848	0.842	0.841
rank	OMD				
1	0.398	0.548	<b>0.610</b>	0.546	0.540
2	0.409	0.622	<b>0.676</b>	0.621	0.616
3	0.467	0.691	<b>0.733</b>	0.691	0.687
4	0.551	0.763	<b>0.797</b>	0.763	0.760
5	0.662	0.844	<b>0.875</b>	0.844	0.841

Table 2: NDCG results for combination of the visual-semantic graphs. Bold fonts indicate best results.

tures are useful, illustrating that only one of them is recommendable for web image retrieval.

Currently we are working on new combinations of visual descriptors with semantic graphs, defining new strategies to optimize the combination. We are also exploring scalable-methods to increase the number of nodes used in our graphs.

## 6. REFERENCES

- [1] D. N. Bhat and S. K. Nayar. Ordinal measures for image correspondence. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 20(4):415–423, 1998.
- [2] N. Brisaboa, A. Farina, O. Pedreira, and N. Reyes. Similarity search using sparse pivots for efficient multimedia information retrieval. In *ISM '06 Proc.*, 2006.
- [3] B. Gao, T.-Y. Liu, T. Qin, X. Zheng, Q. Cheng, and W.-Y. Ma. Web image clustering by consistent utilization of visual features and surrounding texts. In *MM '05 Proc.*, 2005.
- [4] T. Haveliwala and S. Kamvar. The second eigenvalue of the google matrix. Technical report, Stanford University, 2003.
- [5] K. Järvelin and J. Kekäläinen. Cumulated gain-based evaluation of ir techniques. *ACM Trans. Inf. Syst.*, 20(4):422–446, 2002.
- [6] B. Manjunath, J. rainer Ohm, V. V. Vasudevan, and A. Yamada. Color and texture descriptors. *IEEE Trans. on Circuits and Systems for Video Technology*, 11(6):703–715, 2001.
- [7] B. Poblete, C. Castillo, and A. Gionis. Dr. searcher and mr. browser: a unified hyperlink-click graph. In *CIKM '08 Proc.*, 2008.
- [8] L. Wang, L. Yang, and X. Tian. Query aware visual similarity propagation for image search reranking. In *MM '09: Proc.*, 2009.
- [9] X.-J. Wang, W.-Y. Ma, G.-R. Xue, and X. Li. Multi-model similarity propagation and its application for web image retrieval. In *MM '04 Proc.*, 2004.
- [10] W. Xi, E. A. Fox, W. Fan, B. Zhang, Z. Chen, J. Yan, and D. Zhuang. Simfusion: measuring similarity using unified relationship matrix. In *SIGIR '05 Proc.*, 2005.
- [11] W. Xi, B. Zhang, Z. Chen, Y. Lu, S. Yan, W.-Y. Ma, and E. A. Fox. Link fusion: a unified link analysis framework for multi-type interrelated data objects. In *WWW '04 Proc.*, 2004.
- [12] P. Zezula, G. Amato, V. Dohnal, and M. Batko. *Similarity Search: The Metric Space Approach (Advances in Database Systems)*. Springer-Verlag, 2005.