

# Image Retrieval Using Multiple Evidence Ranking

Tatiana Almeida Souza Coelho      Pável Pereira Calado\*  
Lamarque Vieira Souza      Berthier Ribeiro-Neto†  
Richard Muntz‡

Computer Science Department  
Federal University of Minas Gerais  
30.123-970, Belo Horizonte MG  
Brazil

tati@inf.puc-rio.br  
{pavel,berthier,lamarque}@dcc.ufmg.br

Computer Science Department  
University of California at L.A.  
Los Angeles, CA 90095-1596  
USA

muntz@cs.ucla.edu

---

\*The author was supported by MCT/FCT scholarship SFRH/BD/4662/2001.

†The author was supported in part by CNPq Individual Grant 300.188/95-1 and Finep/MCT/CNPq Grant 76.97.1016.00, project SIAM under program Pronex.

‡The author was supported by NSF grants IIS-0086116, ANI-0085773, and EAR-9817773.

## Abstract

The World Wide Web is the largest publicly available image repository and a natural source of attention. An immediate consequence is that searching for images in the Web has become a current and important task. To search for images of interest, the most direct approach is keyword-based searching. However, since images in the Web are poorly labeled, direct application of standard keyword-based image searching techniques frequently yields poor results. In this work, we propose a comprehensive solution to this problem. In our approach, multiple sources of evidence related to the images are considered. To allow combining these distinct sources of evidence, we introduce an image retrieval model based on Bayesian belief networks. To evaluate our approach, we perform experiments on a reference collection composed of 54 thousand Web images. Our results indicate that retrieval using image surrounding text passages is as effective as standard retrieval based on HTML tags. This is an interesting result because current image search engines in the Web usually do not take text passages into consideration. Most important, according to our results, the combination of information derived from text passages with information derived from HTML tags leads to improved retrieval, with relative gains in average precision figures of roughly 50%, when compared to the results obtained by the use of each source of evidence in isolation.

**Keywords:** image retrieval, text-based, Bayesian networks, evidence combination,

World Wide Web

# 1 Introduction

## Motivation

In early times, some 30,000 years ago, images were the only form of written communication [3]. With the invention of printing, 500 years ago, writing took over as the preferential means of formal communication. One key reason is that written text is simpler and cheaper to produce and reproduce in a large scale.

In the last half century, however, the advent of digital devices, such as digital photo cameras, digital palm tops, lap tops, and personal computers have considerably simplified the tasks of producing, editing, and publishing good quality pictures and drawings. Further, the enormous success of the Web has granted low cost, large scale accessibility to this material.

Compounding this scenario, favorable to the diffusion of information encoded in the form of images, there is the dynamic nature of the communication needs of people living in modern societies. Nowadays, a variety of profit-driven activities of modern life demand image-based information transfer. Typical examples are architecture designs, engineering drawings, fashion designs, perfumes, new cars, marketing campaigns, Internet home pages, etc. All these needs have led to an increased interest in the organization, indexing, and retrieval of digital images in the last decade.

Some authors even argue that we may be “on the verge of an important historical swing back towards what may be called the primacy of the image” [17]. While we have no particular opinion on the chances of crystallization of such a tendency, we do believe that availability of digitized image data will continue to grow and that the need for searching such databases will become widespread. In this scenario, indexing and searching of image data are crucially important.

Since low cost high availability is almost always only attainable through the Web, where HTML formatting is dominant, we can assume that images will frequently appear accompanied by some form of HTML tagged text. These tags might be alternative optional fields, image titles, file names, and surrounding text. In this work, we discuss how to use such text-based information as labels (or annotations) that can be used to partially describe the contents of the images. Since there are well known problems regarding the quality of Web documents, selecting which parts of a document are better suited to describe an image is an important and valid problem, not yet approached by other works in image-retrieval.

## The Problem

The retrieval of images is usually based on one of two basic approaches: *content-based* retrieval and *text-based* retrieval [24]. In content-based retrieval, image characteristics, such as color, shape, or texture, are used for indexing and searching. For querying, the user provides a target image, or a description of image features, which is then compared to the images in the database. The images retrieved are ranked according to a distance metric from the user query. In text-based retrieval, some form of textual description of the image contents is assumed to be stored with the image itself. Such descriptions are usually based on annotations made by human beings. To query the database, the user provides a keyword description of his information need. This description is then compared to the descriptions of the stored images, using text retrieval techniques.

In the Web, the adoption of content-based indexing has some disadvantages. First, the computational cost of extracting image characteristics (such as color or shape) for a large collection might be prohibitive. Second, the user query must be provided in the form of a draft of the desired image, which is not simple to do [1]. For text-based

indexing, on the other hand, images can be described using keywords to refer to the image contents [13]. Also, queries can be formulated in standard keyword form. This makes text-based indexing an important approach for image retrieval in the Web. In fact, even when content-based techniques are applied, the textual content of the Web pages should not be disregarded, since it often includes some form of human generated descriptions of the images.

In text-based retrieval, the keywords in the user query are compared to labels associated with the image. In the Web, the natural labels are the name of the file containing the image and the terms in the ALT tag. However, this is frequently insufficient to provide good retrieval capability and additional evidence must be considered. Thus, determining how additional textual information appearing in a Web page can be used to improve the ranking of the images is a relevant problem, which we address in this work.

## Our Solution

Text in Web documents is often uninformative or misleading as to the contents of its images. While the text might include useful information related to an image, selecting which part of the text better describes the image is not a trivial task. In this work, we demonstrate that directly applying Information Retrieval techniques to this problem does not yield the best results. The process of finding relevant images does not depend on the full text of the page nor exclusively on textual information directly associated with the images (e.g. the terms in the ALT tag). Instead, other sources of information within a Web document can be used to improve the ranking of the images. We evaluate which parts of a Web document can be used to complement an effective description of the images and propose an image retrieval model, based on Bayesian belief networks [21], for combining such evidential information into a single image ranking. Bayesian belief networks are

used because they provide a flexible, efficient and formally sound method of combining different sources of evidence in a single information retrieval model [6, 22, 23, 28, 30].

To validate our approach, we perform a series of experiments with a reference collection of about 54,000 images extracted from the World Wide Web. The results indicate that the quality of the search results is highly improved when different sources of textual evidence are properly combined in a ranking function. Average precision figures for image retrieval went up to roughly 60%—an improvement of 50% over the best precision figures when a single source of evidence is used in isolation. The results also indicate that text passages surrounding the image and text contained in the image HTML tags are the best sources of information for image ranking.

Our research provides a useful insight for future work on Web image retrieval, since we now know which parts of an HTML document should be used for image description. Also, the probabilistic combination formula here proposed is flexible enough to allow the inclusion of any other image ranking method, so that, for instance, new solutions that combine text-based and content-based retrieval can use our results as a basis for the development of new ranking formulas.

The paper is organized as follows. Section 2 briefly describes some work related to our research. Section 3 describes the textual sources of evidence we considered within a Web document. Section 4 shows how individual pieces of evidence are combined using a belief network model. Section 6 presents the results of our experiments. Finally, in Section 7 some conclusions on the results are drawn.

## 2 Related Work

Early approaches for image retrieval were text-based approaches. Usually, human annotations were manually added to the pictures and the search was performed using standard database management systems [7, 8]. With the growth of image databases, however, manually describing the images became unfeasible. Interest in image retrieval shifted towards content-based approaches [24].

Content-based image retrieval is based on extracting image features, such as color, texture, or shape, from the objects in the database. The extracted features are then stored and compared to an example image (the query) in a feature description language [1]. The various proposals based on content-based retrieval differ mainly on the techniques used for extracting and storing features [11, 20, 33] and on the use of such features for image searching [10, 12, 26].

With the growing popularity of the Internet, the focus is changing towards image retrieval in the World Wide Web [4, 5, 27]. In the Web, the images are usually integrated into HTML documents. As an immediate consequence, the documents can be used as a source for textual descriptions of the images [13]. The combination of textual information with image feature information has therefore been suggested to improve image search results [9, 16, 19, 29, 34].

In [29], a system to retrieve images and videos on the World Wide Web is presented. Words from the image URL and ALT attributes are used to describe the image and to associate it with a predefined category. This category can then be used to query the database. Lu and William [19] propose the extraction of terms from the HTML document, weighting them according to their location in the document. A normalized sum of these weights is then combined with a color comparison measure by adding both values. A

similar approach is taken in [9] and [16], where some refinements are introduced to allow for user relevance feedback and the use of other image features. In [34], image features and textual information are combined through a self-organizing neural network. In all these works, however, there is no formal study on which parts of the HTML document should be used, and how they should be combined for image retrieval. Either the combination of keywords is purely heuristic (e.g., a linear combination with an arbitrary selection of weights), or there is no selection of which parts to combine.

Our work differs from previous proposals in the following directions. First, we propose a formally sound image retrieval model, based on Bayesian belief networks, to combine several sources of text-based evidence into one image ranking formula. Second, we consider each part of the HTML document containing the images (e.g., URL, ALT, page title, image title, surrounding text) as an individual source of evidence, which we then combine through a belief network. Third, we achieve high precision results through the exclusive use of textual information, without requiring the use of costly image processing algorithms. Fourth, we evaluate empirically, on a real Web collection, which parts of the HTML document are better suited for describing the images and discuss our results. Finally, while we do not focus on content-based retrieval, our image network model provides a flexible and formally sound framework that allows naturally adding content-based evidence, or any other type of information, to improve image ranking.

### **3 Textual Sources of Evidence in Web Documents**

World Wide Web documents include a variety of textual data that can be used for image retrieval. However, due to the uncontrolled nature of the Web, the textual contents of a page does not necessarily include a proper description of the images in the page. For

instance, text surrounding the images can be provided for navigation purposes only, like “click here” messages. Also, image filenames might have been automatically generated, like “image001.gif”. These and other problems compromise the use of only one particular piece of text in the page to describe the images.

A possible solution is to consider each part of the text as an independent source of evidential information. Our proposal is to combine these sources of evidence using the Bayesian network model we devised. We consider four possible sources of evidence contained in Web documents, namely:

- 1) Description tags: composed of the terms (i.e., words) found in the filename of the image, in the ALT attribute of the IMG tag, and between the anchor tags `<A>` and `</A>`. These terms are usually used to describe the image with which they are associated. Absolute filenames, which include the host and directory of the images, were not used since they are often irrelevant.
- 2) Meta tags: composed of the terms located between the tags `<TITLE>` and `</TITLE>` and in the tag META of the HTML document. These terms are used to describe the document contents. When we use them to describe an image, we assume that the image topic is related to that of the document, i.e, words that describe the contents of the document also describe the contents of its images. In our approach, only the meta tags related to author, keywords, and description are used.
- 3) Full text: composed of all the words on the Web page. Here, we again assume that the topic of the document is related to the topic of the images contained in the document. This differs from the use of meta tags, since it provides a richer set of words for describing the document topic.

4) Text passages: composed of the words located close to the images. Text surrounding an image is expected to be related to the contents of the image. These short pieces of text are usually called *passages*[18]. In our experiments we found that a passage consisting of the 20 terms before and after the image produces the best results. It is important to note that the full text approach can be seen as a special case of the passage approach, where we consider passages large enough to contain the whole document. For this reason, text passages and full text were not combined together in our experiments.

Even though the use of any of these sources of evidence in isolation can lead to poor retrieval, we show, through experimentation, that they constitute complementary sources of evidence and that their combination in our Bayesian network model effectively improves the quality of the set of images retrieved.

## 4 The Belief Network Model

The image network model that we propose is based on the belief network model, which we now review.

### 4.1 Basic Concepts

Bayesian networks provide a graphical formalism for explicitly representing independencies among the variables of a joint probability distribution. The probability distribution is represented through a directed acyclic graph whose nodes represent the random variables of the distribution. The relationships among these variables are modeled as directed edges which represent causal dependencies among the linked variables. The strengths of

these dependencies are expressed by conditional probabilities. The fundamental principle in a Bayesian network is that the known independencies among the random variables of a domain are declared explicitly and that a joint probability distribution is synthesized from this set of declared independencies. For more details, we refer the reader to [21].

Two random variables,  $X$  and  $Y$ , are represented in a Bayesian network as two nodes in a directed graph, also referred to as  $X$  and  $Y$ . An edge directed from  $Y$  to  $X$  represents the influence of the node  $Y$ , the *parent* node, on the node  $X$ , the *child* node. Let  $x$  be a value taken by variable  $X$  and  $y$  a value taken by variable  $Y$ . The intensity of the influence of the variable  $Y$  on the variable  $X$  is quantified by the conditional probability  $P(x|y)$ , for every possible set of values  $(x, y)$ .

In general, let  $\mathbf{P}$  be the set of all parent nodes of a node  $X$ ,  $\mathbf{p}$  be a set of values for all the variables in  $\mathbf{P}$ , and  $x$  be a value of  $X$ . The influence of  $\mathbf{P}$  on  $X$  can be modeled by any function  $\mathcal{F}$  that satisfies the following conditions:

$$\sum_{x \in D_X} \mathcal{F}(x, \mathbf{p}) = 1$$

$$0 \leq \mathcal{F}(x, \mathbf{p}) \leq 1.$$

The function  $\mathcal{F}(x, \mathbf{p})$  provides a numerical quantification for the conditional probability  $P(x|\mathbf{p})$ .

To illustrate, Figure 1 shows a Bayesian network for a joint probability distribution  $P(x_1, x_2, x_3, x_4, x_5)$ , where  $x_1, x_2, x_3, x_4$ , and  $x_5$  refer to values of the random variables  $X_1, X_2, X_3, X_4$ , and  $X_5$ , respectively. The node  $X_1$  is a node without parents and is called a *root node*. The probability  $P(x_1)$  associated with a value  $x_1$  of the root node  $X_1$  is called a *prior probability* and can be used to represent a previous knowledge of the modeled domain. Due to the independencies declared in Figure 1, the joint probability

distribution can be computed as

$$P(x_1, x_2, x_3, x_4, x_5) = P(x_1)P(x_2|x_1)P(x_3|x_1)P(x_4|x_2, x_3)P(x_5|x_3)$$

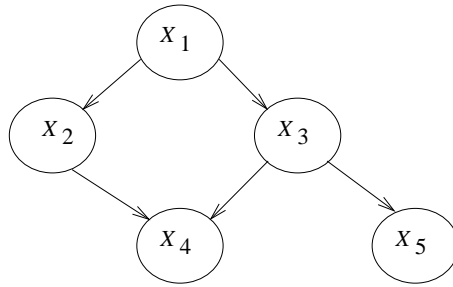


Figure 1: Example of a Bayesian network.

## 4.2 Belief Network Model for IR

In this section, we briefly review the main features of the *belief network model* as discussed in [22]. As in [30, 32], the belief network model takes an epistemological view of the IR problem and interprets probabilities as degrees of belief devoid of experimentation. This is the reason for calling it a *belief network model*.

### Defining the Sample Space

In a traditional information retrieval system, the documents in a collection are usually referred to through the use of *keywords*. Each keyword  $k_i$  can be seen as an event. Thus, the set of all keywords in the collection defines a sample space, which we call  $\mathcal{U}$ , and a document can be seen as the occurrence of a set of events, a subset of  $\mathcal{U}$ .

We associate with each keyword a binary random variable, denoted by  $k_i$ . This variable is 1 to indicate that the keyword was observed (i.e., is on the state *on*). A document  $d_j$  is modeled as a set composed of selected keywords that occur in its text. If

all the variables associated with the keywords in the document are in the *on* state, we say that the document has been observed. A query  $q$  is modeled analogously.

### The Network for IR

Given the sample space  $\mathcal{U}$ , it is natural to interpret queries and documents as subsets of  $\mathcal{U}$ . As a result of this interpretation, queries and documents are treated analogously. This symmetry induces the belief network of Figure 2.

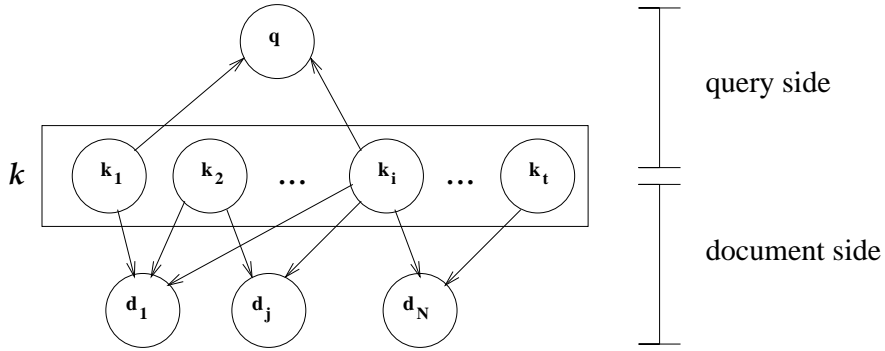


Figure 2: Belief network for a query  $q$  composed of the keywords  $k_1$  and  $k_i$ .

In this network, each node  $d_j$  models a document, the node  $q$  models the user query, and the  $k_i$  nodes model the keywords in the collection. The set  $\mathbf{k}$  is used to refer to any of the  $2^t$  possible states of the  $k_i$  root nodes. Instantiation of the root nodes *separates* the document nodes from the query node, making them mutually independent. Thus, in the belief network of Figure 2, we say that the query is on the *query side* of the network, while the documents are on the *document side* of the network.

With the node  $q$  is associated a binary random variable which is also denoted by  $q$ . This variable is 1 (also said to be *on*) to indicate that query  $q$  was observed. A document  $d_j$  is modeled analogously i.e., there is also a binary random variable associated with  $d_j$ . This variable is 1 (also said to be *on*) to indicate that document  $d_j$  was observed. For

computing a ranking, we use Bayes' law and the rule of total probabilities, as follows:

$$P(d_j|q) = \eta \sum_{\mathbf{k}} P(d_j|\mathbf{k}) P(q|\mathbf{k}) P(\mathbf{k}) \quad (1)$$

which is the generic expression for the rank of a document  $d_j$  with regard to a query  $q$ , in the belief network model. The conditional probability  $P(d_j|q)$  computes the probability of observing document  $d_j$  given that the query  $q$  was observed and can be used to compute a vector ranking, as we now discuss.

### The Vector Model

The vector space model represents queries and documents as vectors in a  $t$ -dimensional space. Each dimension of this space is represented by a orthonormal vector  $\vec{k}_i$  associated with the keyword  $k_i$ . The number  $t$  of dimensions is the number of distinct keywords in the collection.

Let  $N$  be the total number of documents in a collection,  $n_i$  be the number of documents in which the keyword  $k_i$  appears, and  $f_{ij}$  be the raw frequency of the keyword  $k_i$  in the document  $d_j$ . With each keyword-document pair  $(k_i, d_j)$  is associated a weight  $w_{ij}$  given by  $f_{ij} \times \log N/n_i$ . For the term-query weight  $w_{iq}$ , one can adopt  $f_{iq} \times \log N/n_i$ , where  $f_{iq}$  is the raw frequency of the keyword  $k_i$  in the query  $q$  (it common that  $f_{iq} = 1$ ). This type of weighting is usually called a *tf-idf* (term frequency and inverse document frequency) weighting scheme [2, 25].

Given the sets of weights  $w_{ij}$  and  $w_{iq}$ , the document  $d_j$  and the query  $q$  are represented as  $t$ -dimensional vectors  $\vec{d}_j = (w_{1j}, w_{2j}, \dots, w_{tj})$  and  $\vec{q} = (w_{1q}, w_{2q}, \dots, w_{tq})$ , respectively. The similarity between a document  $d_j$  and a query  $q$  is computed as the cosine of the angle between the document and query vectors, given by

$$\text{sim}(d_j, q) = \frac{\vec{d}_j \cdot \vec{q}}{|\vec{d}_j| \times |\vec{q}|} = \frac{\sum_{i=1}^t w_{ij} \cdot w_{iq}}{\sqrt{\sum_{i=1}^t w_{ij}^2} \sqrt{\sum_{i=1}^t w_{iq}^2}} \quad (2)$$

## The Network for the Vector Model

A belief network can be used to compute a vectorial ranking by making Eq. (1) equivalent to Eq. (2). This is accomplished through proper specification of the probabilities  $P(d_j|\mathbf{k})$  and  $P(q|\mathbf{k})$ , as follows.

$$P(q|\mathbf{k}) = \begin{cases} 1 & \text{if } \forall_i, I_q(k_i) = I_{\mathbf{k}}(k_i) \\ 0 & \text{otherwise} \end{cases} \quad (3)$$

$$P(d_j|\mathbf{k}) = \frac{\sum_{i=1}^t w_{ij} \cdot w_{i\mathbf{k}}}{\sqrt{\sum_{i=1}^t w_{ij}^2} \sqrt{\sum_{i=1}^t w_{i\mathbf{k}}^2}} \quad (4)$$

where  $I_q(k_i)$  is an indicator function (1 if  $k_i \in q$ , 0 otherwise),  $w_{i\mathbf{k}}$  and  $w_{ij}$  are the weights used in the vector model. This specification is valid and consistent because  $P(d_j|\mathbf{k})$  measures the cosine of the angle between two vectors, which is a number between 0 and 1. As Eq. (3) restricts the computation only for those states  $\mathbf{k}$ , such that  $I_q(k_i) = I_{\mathbf{k}}(k_i)$  for all  $i$ , we can replace  $w_{i\mathbf{k}}$  by  $w_{iq}$  in Eq. (4). As a result, the ordering of documents (i.e., ranking) defined by  $P(d_j|q)$  coincides with the ordering of documents defined by  $\text{sim}(d_j, q)$  in Eq. (2).

We observe that the belief network is used here as a modeling framework and not as an inference engine. While more complex designs are possible, our simple representation is powerful enough to allow modeling important relationships between documents, queries, and user needs.

## 5 Image Network Model

To combine the distinct sources of information associated with Web images into a single ranking function, we extend the belief network model for IR. This is accomplished by adding new edges, nodes, and probabilities to the original network presented in Figure 2.

We say that this expansion is modular in the sense that it preserves all the properties of the previous network. the extended network is referred to as an *image network model*.

Figure 3 shows this image network model. Each new node represents a binary variable that models a piece of information. The  $K_i$  nodes model the terms extracted from the Web documents using all sources of evidence described in Section 3. The  $ID_j$  nodes model the evidence extracted using the *description tags*. The  $IM_j$  nodes model the evidence extracted using the *meta tags*. The  $IP_j$  nodes model the evidence extracted using the *full text* or the *passages*. Finally, the  $I_j$  nodes model the retrieved images. As in [22, 28], the distinct pieces of evidence are combined through a disjunctive operator, to yield a rank for each retrieved image.

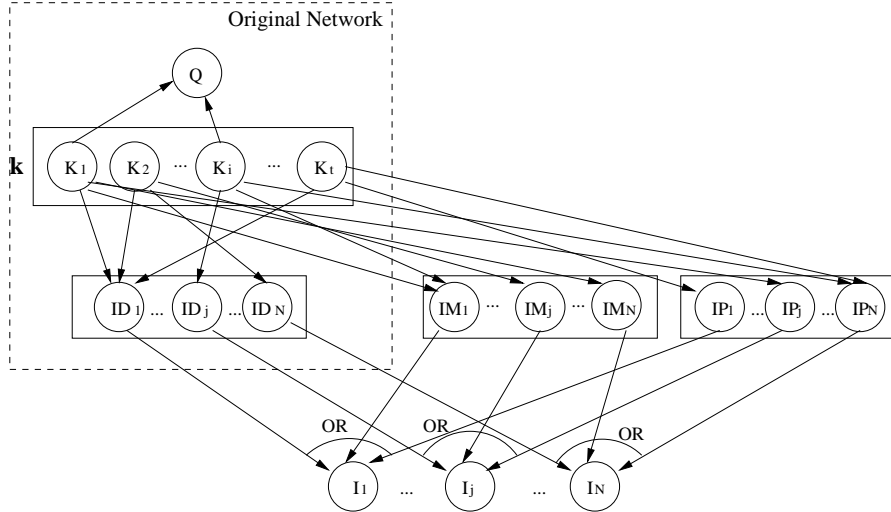


Figure 3: Image network model for combining the multiple evidences extracted from the text of a Web document.

To rank the images in the collection, we are interested in the degree of belief associated with an image  $I_j$ , given that the user query has been declared, i.e.,  $P(I_j = 1|Q = 1)$ , or simply  $P(i_j|q)$ . By the rule of total probabilities and the independencies modeled in the

network we can write:

$$P(i_j|q) = \eta \sum_{\mathbf{k}} P(i_j|\mathbf{k}) \times P(q|\mathbf{k}) \times P(\mathbf{k}) \quad (5)$$

We consider  $P(i_j|\mathbf{k})$  as an accumulator of evidential knowledge, which is modeled by an *OR* operator. We can thus write:

$$P(i_j|q) = \eta \sum_{\mathbf{k}} [1 - (1 - P(id_j|\mathbf{k})) \times (1 - P(im_j|\mathbf{k})) \times (1 - P(ip_j|\mathbf{k}))] \times P(q|\mathbf{k}) \times P(\mathbf{k}) \quad (6)$$

where  $\eta$  is a normalizing constant [21], introduced to make the sum of all probabilities equal 1, and  $\mathbf{k}$  is the state of all the  $K_i$  variables. In our belief network model, Eq. (6) represents the generic expression for computing the rank of an image  $I_j$  with regard to the user query. To define the final ranking equation, we must now specify each of the probabilities  $P(id_j|\mathbf{k})$ ,  $P(im_j|\mathbf{k})$ ,  $P(ip_j|\mathbf{k})$ ,  $P(q|\mathbf{k})$ , and  $P(\mathbf{k})$ .

## 5.1 Image Ranking

The rank attributed to an image will depend on the conditional probabilities stated in Eq. (6). We therefore proceed by defining them, as follows:

$$P(q|\mathbf{k}) = \begin{cases} 1 & \text{if } \forall_i I_q(k_i) = I_{\mathbf{k}}(k_i) \\ 0 & \text{otherwise} \end{cases} \quad (7)$$

As for the network in Figure 2, this means that only state  $\mathbf{k}$  (of the set of the  $K_i$  nodes) contributes positively to  $P(i_j|q)$ , where  $\mathbf{k}$  is the state for which the active keywords are exactly those in the user query. The prior probabilities  $P(\mathbf{k})$  take a constant value, because we have no preference to any set of keywords (before evidence, such as the query, is revealed).

The probability of each of the possible sources of evidence being observed, given  $\mathbf{k}$ , is computed using the vector space model [25]. For the *description tags*, we define:

$$P(id_j|\mathbf{k}) = \frac{\sum_{i=1}^t w_{ij} \times w_{iq}}{\sqrt{\sum_{i=1}^t w_{ij}^2} \times \sqrt{\sum_{i=1}^t w_{iq}^2}} \quad (8)$$

where  $w_{ij}$  is the weight of the term  $k_i$  in the description tags of image  $I_j$ , and  $w_{iq}$  is the weight of the term  $i$  in the user query. As in [31], we define these weights as:

$$w_{ij} = (1 + \ln f_{ij}) \quad w_{iq} = \ln \left( 1 + \frac{N}{n_i} \right) \quad (9)$$

where  $f_{ij}$  is the frequency of term  $k_i$  in the description tags of image  $I_j$ ,  $N$  is the total number of images, and  $n_i$  is the number of all description tags in the collection that contain term  $k_i$ . The factor  $\ln(1+N/n_i)$  is the *inverse document frequency* [2] and provides for proper accounting of specific terms (i.e., terms that are discriminating because they are not very frequent).

Similar equations can be written for  $P(im_j|\mathbf{k})$ , the probability that evidence from the meta tags is observed, and  $P(ip_j|\mathbf{k})$ , the probability that evidence from the full text, or related passage, is observed. In each case, the weights of the terms are calculated according to their occurrence in the respective source of evidence.

To simplify Eq. (6), we adopt the following notation:

$$P(id_j|\mathbf{k}) = RD_{jq}$$

$$P(im_j|\mathbf{k}) = RM_{jq}$$

$$P(ip_j|\mathbf{k}) = RP_{jq}$$

where  $RD_{jq}$ ,  $RM_{jq}$ , and  $RP_{jq}$  are the partial ranking scores (with regard to query  $q$ ) contributed by description tags, meta tags, and passages, respectively, and  $\mathbf{k}$  is the state of the variables  $K_i$  for which the only variables that are *on* are those corresponding to

the query terms. We can then rewrite Eq. (6) as:

$$P(i_j|q) = \eta \times [1 - (1 - RD_{jq}) \times (1 - RM_{jq}) \times (1 - RP_{jq})] \quad (10)$$

To test the distinct evidence combinations, we can set to zero the factor associated with the evidence that we want to disregard. For instance, to evaluate the *description tags* approach in isolation, we set  $RM_{jq} = 0$  and  $RP_{jq} = 0$ . Substituting these values in Eq. (10), we obtain:

$$P(i_j|q) = \eta \times RD_{jq} \quad (11)$$

Analogously,  $P(i_j|q)$  can be calculated for the other proposed evidences or combination of evidences. Table 1 shows the various image ranking equations that we can generate. These seven different rankings were the subject of our experiments.

Ranking approach	$P(i_j q)$
Description tags	$\eta \times RD_{jq}$
Meta tags	$\eta \times RM_{jq}$
Passage/Full text	$\eta \times RP_{jq}$
Description+Meta tags	$\eta \times [1 - (1 - RD_{jq}) \times (1 - RM_{jq})]$
Description+Passage/FT	$\eta \times [1 - (1 - RD_{jq}) \times (1 - RP_{jq})]$
Passage/FT+Meta tags	$\eta \times [1 - (1 - RP_{jq}) \times (1 - RM_{jq})]$
Description+Meta tags+Passage/FT	$\eta \times [1 - (1 - RD_{jq}) \times (1 - RM_{jq}) \times (1 - RP_{jq})]$

Table 1: Evidence combinations modeled in the belief network model.

## 6 Evaluation Strategy

### 6.1 Reference Collection

Tests were carried out to verify the effectiveness of the seven proposed ranking variations. Table 2 shows some statistics on the reference collection used to perform the experiments. The collection consisted of a set of 128,712 pages with their respective images, extracted from the Brazilian Web (under the domain “.br”). We considered as distinct images those that presented distinct absolute URLs. Therefore, images that appeared on distinct pages were considered distinct images.

<b>Size of the Collection (GB)</b>	1.8
<b>Number of Pages</b>	128,712
<b>Number of Distinct Images</b>	54,571
<b>Number of Queries</b>	25
<b>Number of Images per Query Pool</b>	28.2
<b>Relevant Images per Query Pool</b>	7.3

Table 2: Statistics of the collection used in the tests.

We used 25 keyword-based queries in the tests, which is a number considered appropriate in the IR community. The queries used in the experiments were: *sunset, football ball, Marisa Monte, Snoopy, church, coca cola, mangalarga horse, map of Brazil, Corcovado, basset, Canastra sierra, Edson Arantes do Nascimento, federal university of Minas Gerais, Mônica’s gang, linux, Jesus, Pirenópolis, Fernando de Noronha, flower vase, Skol bier, Glória hotel, carnival photos, Carrefour, shark, and Rio de Janeiro beach.*

In our experiments, only images of size larger than 45 pixels in width or height were

considered. Smaller images have no informational content and are used for decoration or navigation purposes mainly, like back arrows or separator lines. For the same reason, background images and images included through the HTML INPUT tag were also excluded. As a result, we were left with a total of 54,571 images.

## 6.2 Precision and Recall

To evaluate our results we use *precision* and *recall* figures[2]. These metrics consider that, for each test query, a set of relevant documents has been defined. In the case of specific domain document collections, this is usually done by specialists in the domain of knowledge. In the case of generic domain collections, relevance judgments are accomplished through human evaluations.

Given a query  $Q$  and a set  $R$  of relevant documents for a query  $Q$ , precision and recall figures can be used to evaluate the quality of a retrieval method, as follows. Using the retrieval method, we obtain a set  $A$  of documents as answers to the query  $Q$ . This set is then compared to the set  $R$  of relevant documents. The higher is the overlap between them, the better the results considered. Precision and recall are defined as a means to characterize this overlap, as follows. Precision is the fraction of all answers in  $A$  that are correct, i.e:

$$\text{precision} = \frac{|A \cap R|}{|A|}$$

Recall is the fraction of answers in  $R$  that were retrieved, i.e.:

$$\text{recall} = \frac{|A \cap R|}{|R|}$$

Frequently, we want to evaluate average precision at given recall levels. The standard 11-point average precision measure returns precision at 0%, 10%, 20%, ..., 100% of recall.

Precision at 0% recall is the precision when the first relevant document is seen in the ranking (starting from the top). Precision at 10% recall is the precision when 10% of the relevant documents in the set  $R$  have been seen in the ranking (starting from the top). Average precision at 10% recall is the average precision for all test queries, taken at 10% recall.

To determine the set  $R$  of relevant images to each of our 25 test queries, we proceeded as follows. For each test query we ran our 7 ranking variants in Table 1. The 25 highest ranked images retrieved by each of our 7 rankings were pooled into a set of unique images. This way, it is not possible to tell which ranking retrieved which image. The images in each pool were then classified (by volunteers) as relevant or non-relevant with regard to the respective query topic. As a result, we have a set of images, labeled as relevant or non-relevant, independently of how they were retrieved. By matching this set against each of the ranking alternatives, we can evaluate their effectiveness. It is important to note that this method does not guarantee that every relevant image in the collection is found, a clearly infeasible task. For this reason, when a query returns all the images in set  $R$ , we say that the query has achieved 100% of *relative recall*.

After these relevance judgments were concluded, we found that the average number of relevant images per query pool is 7.3. This pooling method was also used with the Web-based collection of TREC [14, 15]. It avoids the need to evaluate the whole collection and guarantees that the user classifying the images has no knowledge of the ranking strategy used to retrieve them, thus providing an impartial evaluation of the relevance of the images retrieved.

## 6.3 Results

### Full Text Versus Text Passages

Our first test was performed to determine the best size for the passages surrounding the images. We tested 10-term, 20-term, 40-term passages, and full text. A 10-term, 20-term, or 40-term passage is one in which the reference to the image appears at the middle point in the passage. Table 3 shows average precision figures for each passage size, for our 25 test queries, considering the 25 top ranked images for each query. Passages of 40 terms provided the best results. Larger passages led to poorer results. Further, the full text approach presented the lowest precision figures. The reason is that the whole body of text in a Web document is often very ambiguous, dealing with several topics frequently not related to the contents of the images in the document. In fact, a passage of text surrounding the image is much more informative about the image contents than the full text in the page. As a consequence of these results, in the remainder of our experiments, we consider only passages of 40 terms.

10 Terms	20 Terms	40 Terms	Full text
0.182951	0.235012	0.258544	0.159184

Table 3: Average precision figures for passages of distinct sizes.

### Single Sources of Evidence

To determine how each single source of evidence described in Section 3 (i.e., description tags, meta tags, and text passages) contributes to the relevance of the images, we evaluate each of them separately. Figure 4 shows the average precision/recall figures. We observe that the description tags and the text passage approaches presented similar results over

a large spectrum of recall values. On the other hand, the results indicate that the use of meta tags in a Web document is the least informative approach. This lack of useful information in the text of the meta tags is due to three main reasons: a) meta tags describe the topic of the page, which is not always coherent with the topic of its images, b) Web page authors often insert non-descriptive words in the meta tags, and also repeat them several times, in an attempt to increase the score of the page in commercial search engines (spamming), and c) many authors do not even use the meta tags.

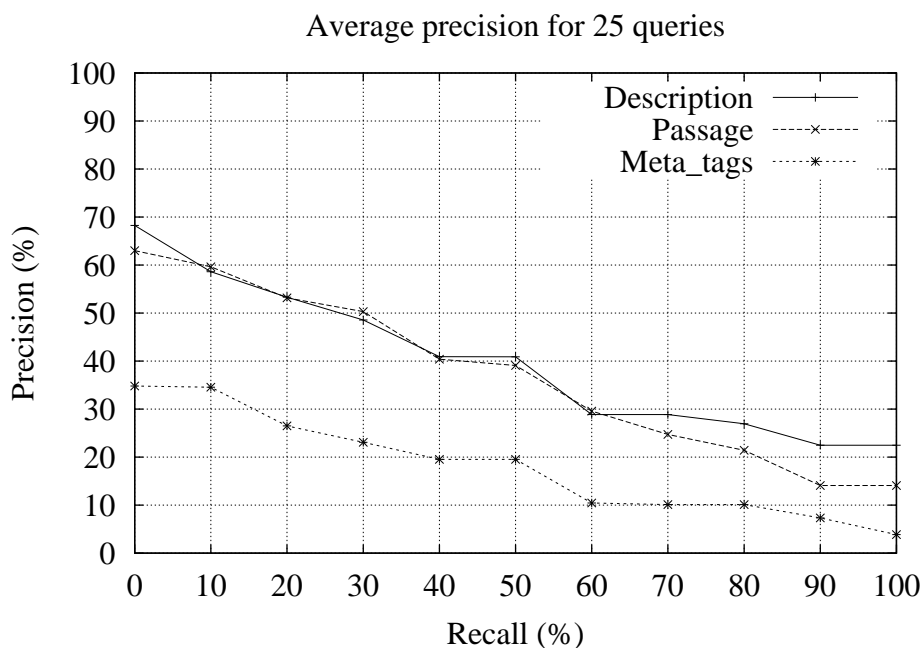


Figure 4: Average precision-recall figures obtained using each of the proposed sources of evidence in isolation, for our 25 test queries.

### Multiple Sources of Evidence

We now examine the results obtained when we combine multiple sources of evidence in our image network model. Figure 5 presents the results for our four possible combinations of evidence: *description+meta\_tags*, *description+passage*, *passage+meta\_tags*, and *descrip-*

*tion+meta\_tags+passage*. Due to the poor performance of the meta tags approach, the *description+meta\_tags* and the *passage+meta\_tags* combinations performed the worst. The *description+passage* approach yielded an overall average precision of 60.1% (obtained by averaging the precision at all recall levels). The *description+passage+meta\_tags* approach yielded an overall average precision of 59.0%. Although both performed similarly, the *description+passage* approach showed higher precision values at recall levels below 40%. This is an important result for Web search engines, where precision is most important among the first documents in the ranking.

For recall levels above 60% the *description+passage+meta\_tags* approach presents a slightly better retrieval performance. This shows that, even though meta tags showed poor results when used as a single source of evidence, they can still provide some useful information. However, for a real Web image search engine it is arguable whether these results justify the use of meta tags, since high precision is most important at low recall values.

To better appreciate the actual gain provided by the combination of multiple sources of evidence to rank images, we plot in Figure 6 the results for the *description+passage* combination of evidences and for each source of evidence in isolation. The results clearly show that the combination of evidences leads to a high increase in average precision figures. The combination of evidence from the text in the description tags with evidence from the text in the 40-term passages surrounding the images yielded a gain of 50.3% in overall average precision, with regard to the use of the description tags in isolation.

These results indicate that the image network model we proposed provides an adequate framework for combining several sources of textual evidence into a single improved ranking formula for Web images. Since all of our results are based on sources of evidence

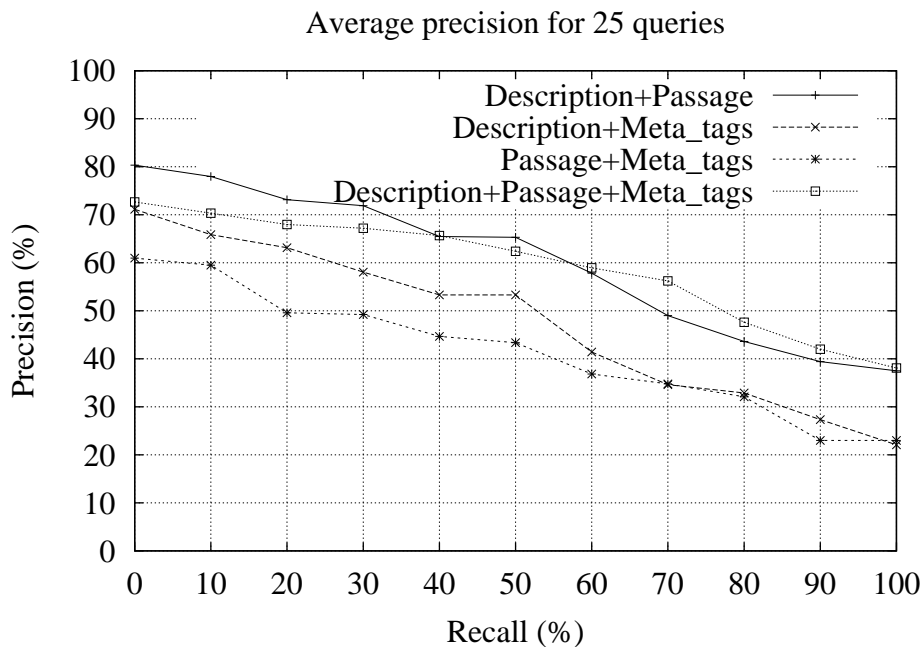


Figure 5: Average precision-recall figures for various combinations of evidence for our 25 test queries.

that can be obtained in fully automatic mode, our image network model provides a new and viable solution to the problem of ranking images in the Web. This is the main result of this work.

## 7 Conclusions

We proposed an image retrieval model, based on belief networks, to combine information from distinct sources of text-based evidence, occurring within an HTML document, to improve the ranking of images in the Web. Textual information was extracted from HTML tags and from the text in the Web document. This information was then combined into a single image ranking formula using the image retrieval model we proposed.

Through experiments, we compared the impact on retrieval of four distinct sources

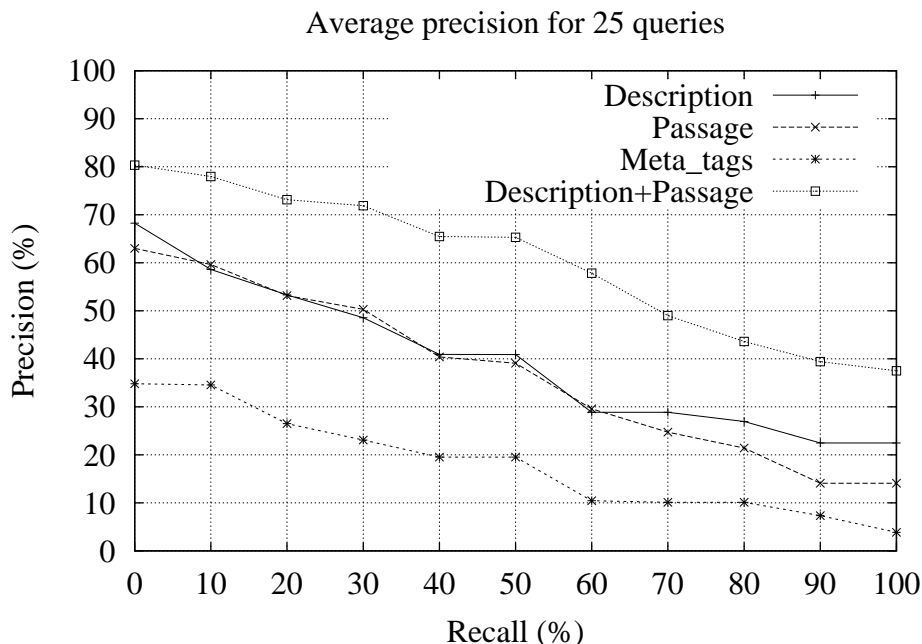


Figure 6: Average precision-recall figures for single sources of evidence and for the best combination, *description+passage*.

of evidence associated with the images in a Web document: the description tags (i.e., ALT, IMG, and anchor tags), the meta tags of the document containing the image, the whole text in the document containing the image, and passages of text surrounding the image. We concluded that the combination of description tags with 40-term passage information provides the best description of an image topic in general, thus being the most useful approach for image retrieval. We found that the whole text of a Web page is too ambiguous to be of use when ranking images.

We experimented with combining each of our sources of evidence in our image retrieval model. The combination of description tags with 40-term passages of text presented the best results and yielded overall average precision figures of 60.1%. This represents an improvement of 50.3% over the best results obtained when a single source of evidence is used in isolation. Combining evidence for image ranking is therefore an advantageous

and practical solution to the problem of image retrieval in the Web. Further, the image network model we introduced provides an effective and formally sound method for combining multiple evidences into a single image ranking.

Since the model we introduced is quite flexible, different alternatives for evidence combination might be tried. For instance, the model can be adapted to combine content-based information, encoded in the form of feature vectors, with the text-based information used here. Also, information derived from the hypertextual structure of the Web can also be combined with evidence from the above approaches.

## References

- [1] Y. A. Aslandogan and C. T. Yu, "Techniques and systems for image and video retrieval," *IEEE Transactions on Knowledge and Data Engineering*, vol. 11, pp. 56–63, January 1999.
- [2] R. Baeza-Yates and B. Ribeiro-Neto, *Modern Information Retrieval*. Addison-Wesley, 1st ed., 1999.
- [3] D. Bathurst, R. Bathurst, and D. Davies, *The Telling Image: The Changing Balance Between Pictures and Words in a Technological Age*. Claredon Press, 1st ed., 1990.
- [4] A. B. Benitez, M. Beigi, and S.-F. Chang, "Using relevance feedback in content-based image metasearch," *IEEE Internet Computing*, vol. 2, pp. 59–69, July 1998.
- [5] A. B. Benitez, M. Beigi, and S. F. Chang, "MetaSEEk: A content-based meta search engine for images," in *Proceedings of Storage and Retrieval for Image and Video Databases (SPIE)*, (Sao Jose, California), December 1997.

- [6] P. Calado, B. Ribeiro-Neto, N. Ziviani, E. Moura, and I. Silva, “Local versus global link information in the web.” January 2003.
- [7] N. S. Chang and K. S. Fu, “Query-by-pictorial-example,” *IEEE Transactions on Software Engineering*, vol. 6, pp. 519–524, November 1980.
- [8] S. K. Chang and T. L. Kunii, “Pictorial data-base systems,” *IEEE Computer*, vol. 14, pp. 13–21, November 1981.
- [9] Z. Chen, L. Wenyin, F. Zhang, M. Li, and H. Zang, “Web mining for Web image retrieval,” *Journal of the American Society for Information Science and Technology*, vol. 52, pp. 831–839, August 2001.
- [10] I. J. Cox, M. L. Miller, T. P. Minka, T. Papathomas, and P. N. Yianilos, “The Bayesian image retrieval system, PicHunter: Theory, implementation and psychophysical experiments,” *IEEE Transactions on Image Processing*, vol. 9, pp. 3–19, January 2000.
- [11] E. A. El Kwaie and M. R. Kabuka, “Efficient content-based indexing of large image databases,” *ACM Transactions on Information Systems*, vol. 18, pp. 171–210, April 2000.
- [12] M. Flickner, H. Sawhney, W. Niblack, J. Ashley, Q. Huang, B. Dom, M. Gorkani, J. Hafner, D. Lee, D. Petkovic, D. Steele, and P. Yanker, “Query by image and video content: The QBIC system,” *IEEE Computer*, vol. 28, pp. 23–32, September 1995.
- [13] V. Harmandas, M. Sanderson, and M. D. Dunlop, “Image retrieval by hypertext links,” in *Proceedings of the 20th Annual International ACM SIGIR Conference on*

- Research and Development in Information Retrieval*, (Philadelphia, Philadelphia), July 1997.
- [14] D. Hawking, N. Craswell, and P. B. Thistlewaite, "Overview of TREC-7 very large collection track," in *The Seventh Text REtrieval Conference (TREC-7)*, (Gaithersburg, Maryland), pp. 91–104, November 1998.
- [15] D. Hawking, N. Craswell, P. B. Thistlewaite, and D. Harman, "Results and challenges in web search evaluation," *Computer Networks*, vol. 31, pp. 1321–1330, May 1999. Also in Proceedings of the The 8th International World Wide Web Conference.
- [16] C. Hu, X. Zhu, H. Zhang, and Q. Yang, "A unified framework for semantics and feature based relevance feedback in image retrieval systems," in *Proceedings of the 8th ACM Internation Conference on Multimedia*, (Los Angeles, California), pp. 31–37, October 2000.
- [17] C. Jørgensen, "Image access: Bridging multiple needs and multiple perspectives — introduction and overview," *Journal of the American Society for Information Science and Technology*, vol. 52, pp. 906–910, September 2001.
- [18] M. Kaszkiel, J. Zobel, and R. Sacks-Davis, "Efficient passage ranking for document databases," *ACM Transactions on Information Systems*, vol. 17, pp. 406–439, October 1999.
- [19] G. Lu and B. William, "An integrated WWW image retrieval system," in *Fifth Australian World Wide Web Conference*, (Lismore, Australia), April 1999.
- [20] V. E. Ogle and M. Stonebraker, "Chabot: Retrieval from a relational database of images," *IEEE Computer*, vol. 28, pp. 40–48, September 1995.

- [21] J. Pearl, *Probabilistic Reasoning in Intelligent Systems: Networks of plausible inference*. Morgan Kaufmann Publishers, 2nd ed., 1988.
- [22] B. Ribeiro-Neto and R. Muntz, “A belief network model for IR,” in *Proceedings of the 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, (Zurich, Switzerland), pp. 253–260, August 1996.
- [23] B. Ribeiro-Neto, I. Silva, and R. Muntz, *Bayesian Network Models for IR*. In *Soft Computing in Information Retrieval: Techniques and Applications*, ch. 11, pp. 259–291. Springer Verlag, 1st ed., 2000.
- [24] Y. Rui, T. S. Huang, and S.-F. Chang, “Image retrieval: Current techniques, promising directions, and open issues,” *Journal of Visual Communication and Image Representation*, vol. 10, pp. 39–62, March 1999.
- [25] G. Salton and M. J. McGill, *Introduction to Modern Information Retrieval*. McGraw-Hill, 1st ed., 1983.
- [26] S. Santini and R. Jain, “Integrated browsing and querying of image databases,” *IEEE Multimedia*, vol. 7, pp. 26–39, July 2000.
- [27] S. Sclaroff, L. Taycher, and M. L. Cascia, “ImageRover: A content-based image browser for the World Wide Web,” in *Proceedings of IEEE Workshop on Content-based Access of Image and Video Libraries*, (Puerto Rico), June 1997.
- [28] I. Silva, B. Ribeiro-Neto, P. Calado, E. Moura, and N. Ziviani, “Link-based and content-based evidential information in a belief network model,” in *Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, (Athens, Greece), pp. 96–103, July 2000.

- [29] J. R. Smith and S.-F. Chang, “An image and video search engine for the World-Wide Web,” in *Symposium on Electronic Imaging: Science and Technology - Storage and Retrieval for Image and Video Databases V*, (San Jose, California), February 1997.
- [30] H. Turtle and W. B. Croft, “Evaluation of an inference network-based retrieval model,” *ACM Transactions on Information Systems*, vol. 9, pp. 187–222, July 1991.
- [31] I. H. Witten, A. Moffat, and T. C. Bell, *Managing Gigabytes: Compressing and Indexing Documents and Images*. Morgan Kaufmann Publishers, 2nd ed., 1999.
- [32] S. K. M. Wong and Y. Y. Yao, “On modeling information retrieval with probabilistic inference,” *ACM Transactions on Information Systems*, vol. 13, pp. 38–68, January 1995.
- [33] J.-K. Wu, “Content-based indexing of multimedia databases,” *IEEE Transactions on Knowledge and Data Engineering*, vol. 9, pp. 978–989, November 1997.
- [34] Q. Wu, S. S. Iyengar, and M. Zhu, “Web image retrieval using self-organizing feature map,” *Journal of the American Society for Information Science and Technology*, vol. 52, pp. 868–875, August 2001.

## Biographies

**Tatiana Almeida Souza Coelho** received a Bachelor degree in Computer Science from the Federal University of Goiás, and an MS degree in Computer Science from the Federal University of Minas Gerais, in Brazil. She is pursuing her PhD at the Informatics Department of the Pontifical Catholic University of Rio de Janeiro, also in Brazil. Her

main interests are information retrieval systems, database systems, workflow systems, mobile systems and interfaces.

**Pável Pereira Calado** received a degree in Computer Engineering from the Instituto Superior Técnico of the Technical University of Lisbon and an MS degree in Computer Science from the Federal University of Minas Gerais, where he is now a PhD student. His research interests include information retrieval, digital libraries, intelligent agents, and Web technologies.

**Lamarque Vieira Souza** received a Bachelor degree in Computer Science from the Federal University of Minas Gerais, Brazil, in 2000. He is currently a MS student in Computer Science at the same university (since 2001). His research interests include operating systems, information retrieval systems, and video on demand. He has been involved in research projects financed through Brazilian national agencies such as the Ministry of Science and Technology (MCT) and the National Research Council (CNPq). From the projects currently under way, the main ones deal with universal access to the Internet in Brazil and video on demand (joint project with UFRJ, UCLA, and UMass).

**Berthier Ribeiro-Neto** received a PhD degree in Computer Science from the University of California at Los Angeles, in 1995. He is an Associate Professor at the Computer Science Department at the Federal University of Minas Gerais. His main interests are information retrieval systems, digital libraries, interfaces for the Web, and video on demand. He is co-author of the book entitled *Modern Information Retrieval*, Addison Wesley, 1999.