1

Intelligent Search for Image Information on the Web through Text and Link Structure Analysis

Euripides G.M. Petrakis

Department of Electronic and Computer Engineering Technical University of Crete (TUC) Chania, Crete, GR-73100 Greece
petrakis@intelligence.tuc.gr

Searching for effective methods to retrieve information from the World Wide Web (WWW) has been in the center of many research efforts during the last few years. The relevant technology evolved rapidly thanks to advances in Web systems technology [1] and information retrieval research [15]. Image retrieval on the Web, in particular, is a very important problem in itself [8]. The relevant technology has also evolved significantly propelled by advances in image database research [20].

Several approaches to the problem of content-based image retrieval on the Web have been proposed and some have been implemented on research prototypes (e.g., ImageRover [23],WebSEEK [21]) and commercial systems. The last category of systems, includes general purpose image search engines (e.g., Google Image Search [1], Yahoo [2], Altavista [3]) as well as systems providing specific services to users such as detection of unauthorized use of images, Web and e-mail content filters, image authentication, licensing and advertising.

Image retrieval on the Web requires that content descriptions be extracted from Web pages and used to determine which Web pages contain images that satisfy the query selection criteria. The methods and systems referred to above differ in the type of content descriptions used and in the search methods applied. There are four main approaches to Web image search and retrieval.

Retrieval by text content: Typically images on the Web are described by text or attributes associated with images in `html` tags (e.g., filename, caption, alternate text etc.). These are automatically extracted from the Web pages and are used in retrievals. Google, Yahoo, and AltaVista are example systems of this category. The importance of the various text fields in retrieving images by text content depends also on their relative location with regard to the location of the images within the Web pages [19].

[1] http://www.google.com/imghp
[2] http://images.search.yahoo.com
[3] http://www.altavista.com/image

Retrieval by image annotations: The Web pages are indexed and retrieved by keywords or text descriptions which are manually assigned to images by human experts. This approach does not scale-up easily for the entire range of image types and the huge volumes of images on the Web. Its effectiveness for general purpose retrievals on the Web is questionable due to the specificity and subjectivity of image interpretations. This approach is common to corporate systems specializing in providing visual content to diverse range of image consumers (e.g., authentication, licensing and advertising of logos, trademarks, artistic photographs etc.).

Retrieval by image content: The emphasis is on extracting meaningful image content from Web pages and in using this content in the retrieval process. Image analysis techniques are applied to extract a variety of image features such as histograms, color, texture measurements, shape properties. This approach has been adopted mainly by research prototypes (e.g., [23, 21]).

Hybrid retrieval systems combining the above approaches such as systems using image analysis features in conjunction with text and attributes (e.g., [29, 25]).

Effective image retrieval on the Web requires integration of text and image content information into the retrieval process. A method is successful if it retrieves the images that the user expects to see in the answers with as few errors as possible. This is a highly subjective processes (i.e., the same results may be judged differently by different users). Query uncertainty and user subjectivity may have a disastrous impact on the quality of the results. Query uncertainty depends on users' level of expertise or familiarity with the system and system functions. Most commonly, users perceive image content in terms of high or semantic level concepts while, in the system, image content is represented in terms of low level image features (e.g., color, texture features). Consequently, users cannot express their information needs in queries or, even worst, there may exist a degree of uncertainty in queries as to what the users are really looking for. Relevance feedback [28, 30] is the state-of-the-art approach for adjusting query results to the needs of the users.

Queries on the Web are issued through the user interface by specifying keywords or free text. The system returns Web pages with similar keywords or text. The highest complexity of queries is encountered in the case of queries by example: The user specifies an example image along with a set of keywords (or annotation) expressing his or her information needs. Queries by example image require that that appropriate content representations be extracted from images in Web pages and matched with similar representations of the queries.

Focusing mainly on image and text content, the work referred to above does not show how to process queries by image example or how to select high quality web pages on the topic of the query. This is achieved by link analysis methods such as HITS [9] link analysis and PageRank [13]. Building upon the same idea, PicASHOW [10] retrieves high quality web images on the topic of the query. However, PicASHOW does not show how to handle image content

and queries by image example. In general, existing methods and systems suffer from one or more of the following drawbacks:

- Work only on annotated image collections without explicit use of image content. Image descriptions or annotations are either manually inserted or automatically computed from image file names, image captions and surrounding text.
- Support only keyword queries as opposed to the most general case of queries by example.
- Do not capture to notion of quality of Web pages. Text or image content are the only cues for achieving high quality results.
- Do not always capture the notion of topic relevance with the users query.
- They are capable of detecting text similarities between Web pages and queries containing lexicographically similar terms but not necessarily semantically (conceptual) similar terms.

In this chapter we show that it is possible to exploit text and image content characteristics of images in Web pages for enhancing the performance of retrievals on the Web. Searching for important (authoritative) Web pages and images is a desirable feature of many Web search engines and is also taken into account. Also, searching by semantic similarity for discovering information related to user's requests (but not explicitly specified in the queries) is a distinguishing feature of many retrieval methods and systems. An obvious enhancement for improving the effectiveness of retrieval methods on the Web is relevance feedback. This work shows how the existing framework of image retrieval with relevance feedback on the Web can be enhanced by incorporating text and image content into the search and feedback process.

As a case study and for demonstrating the efficiency of content-based image retrieval methods, this work deals with the problem of retrieval of logo and trademark images on the Web. Logos and trademarks in particular are important characteristic signs of corporate Web sites or of products presented there. A recent analysis of Web content [7] reports that logos and trademarks comprise 32,6% of the total number of images on the Web. Therefore, retrieval of logo and trademarks is of significant commercial interest (e.g., Patent Offices provide services on unauthorized uses of logos and trademarks).

1.1 Web Content Representation

Typically, images are retrieved by addressing text associated with them (e.g., captions) in Web pages [25]. This is the stat-of-the-art approach for achieving consistency of representation and high accuracy results. Image analysis approaches for extracting meaningful and reliable descriptions for all image types are not yet available. The adaptation of image descriptions to the different image types coexisting on the Web or to the search criteria or different interpretations of image content by different users is also very difficult.

1.1.1 Text Representation

Typically, images are described by the text surrounding them in the Web pages [25]. The following types of image descriptive text are derived based on the analysis of `html` formatting instructions:

Image filename: The `URL` entry (with leading directory names removed) in the `src` field of the `img` formatting instruction.

Alternate text: The text entry of the `alt` field in the `img` formatting instruction. This text is displayed on the browser (in place of the image), if the image fails to load. This attribute is optional (i.e., is not always present).

Page title: The title of the Web page in which the image is displayed. It is contained between the `TITLE` formatting instructions in the beginning of the document. It is optional.

Image caption: A sentence that describes the image. It usually follows or precedes the image when it is displayed on the browser. Because it does not correspond to any `html` formatting instruction it is derived either as the text within the same table cell as the image (i.e., between `td` formatting instructions) or within the same paragraph as the image (i.e., between `p` formatting instructions). If neither case applies, the caption is considered to be empty. In either case, the caption is limited to 30 words before or after the reference to the image file.

All descriptions are lexically analyzed and reduced into term (noun) vectors. First, all terms are reduced into their morphological roots, a stemming algorithm. Similarly, text queries are also transformed to term vectors and matched against image term vectors [15]. More specifically, the similarity between the query Q and the image I is computed as a weighted sum of similarities between their corresponding term vectors

$$\begin{aligned} &S_{text}(Q,T) = \\ &S_{file\_name}(Q,I) + S_{alternate\_text}(Q,I) \ + \\ &S_{page\_title}(Q,I) \ + S_{image\_caption}(Q,I). \end{aligned} \tag{1.1}$$

Each S term is computed as a weighted sum of $tf \cdot idf$ terms without normalizing by query term frequencies (it is not required for short queries). All measures above are normalized on [0,1].

1.1.2 Image Content Representation

Logo and trademark images are easier (than natural images) to describe by low level features computed from raw images. For logo and trademark images the following features are computed [22]:

Intensity histogram: Shows the distribution of intensities over the whole range of intensity values (e.g., [0..255]).

Energy spectrum: Describes the image by its frequency content. It is computed as a histogram showing the distribution of average energy over 256 co-centric rings (with the largest ring fitting the largest inscribed circle of the DFT spectrum).

Moment invariants: Describes the image by its spatial arrangement of intensities. It is a vector of 7 moment coefficients.

The above representations are used to solve the following two problems:

Logo-Trademark detection: Because images on the Web are not properly categorized, filters based on machine learning by decision trees for distinguishing logo and trademark images from images of other categories (e.g., graphics, photographs, diagrams, landscapes) are designed and implemented. In our case, a five-dimensional vector is formed from each image: Each image is specified by the mean and variance of its Intensity and Energy spectrums plus a count of the number of distinct intensities per image. A set of 1,000 image examples is formed consisting of 500 logo-trademark images and 500 images of other types. Images of other types can belong to more than one class: non-logo graphics, photographs, diagrams etc. Their feature vectors are fed into a decision-tree [26] which is trained to detect logo and trademark images. The estimated classification accuracy by the algorithm is 85%. For each image the decision computes an estimate of its likelihood of being logo or trademark or "Logo-Trademark Probability".

Logo-Trademark similarity: The similarity between two images Q, I (e.g., query and a Web image) is computed as

$$\begin{aligned} &S_{image}(Q, I) = \\ &S_{intensity\_spectrum}(Q, I) + S_{energy\_spectrum}(Q, I) + \\ &S_{moment\_invariants}(Q, I). \end{aligned} \tag{1.2}$$

The similarity between histograms is computed by their intersection whereas the similarity between their moment invariant is computed as *1 - Euclidean_vector_distance*.

All measures above are normalized to lie in the interval [0, 1]. To answer queries consisting of both text and example image, the similarity between a query Q and an image I is computed as

$$w = S_{image}(Q, I) + S_{text}(Q, I), \tag{1.3}$$

1.2 Image Information Retrieval on the Web

Image retrieval search engines for the Web supports queries by free text and keywords (the most frequent type of image queries in Web image retrieval systems) addressing text or images in Web pages. Methods for computing the text similarity between queries and Web page or image descriptions are reviewed below.

1.2.1 Vector Space Model (VSM)

Queries and texts are syntactically analyzed and reduced into term (noun) vectors. A term is usually defined as a stemmed non stop-word. Very infrequent or very frequent terms are eliminated. Each term in this vector is represented by its weight. Typically, the weight d_i of a term i in a document is computed as $d_i = tf_i \cdot idf_i$, where tf_i is the frequency of term i in the document and idf_i is the inverse frequency of i in the whole text collection. The formula is modified for queries to give more emphasis to query terms.

Traditionally, the similarity between two documents (e.g., a query Q and a document D) is computed according to the Vector Space Model (VSM) [15] as the cosine of the inner product between their vector representations

$$S(Q, D) = \frac{\sum_i q_i d_i}{\sqrt{\sum_i q_i^2}\sqrt{\sum_i d_i^2}}, \tag{1.4}$$

where q_i and d_i are the weights in the two vector representations. Given a query, all documents (Web pages or images) are ranked according to their similarity with the query.

1.2.2 Semantic Similarity Retrieval Model (SSRM)

For queries by keywords or text, existing methods and systems (e.g., Google, Yahoo) are capable of locating Web pages that contain terms that the users specify in queries. However, the lack of common terms in Web pages and queries does not necessarily mean that they are not related. Two terms can be semantically similar (e.g., can have similar meaning) although they are lexicographically different.

SSRM [5] (Semantic Similarity Retrieval Model) works by discovering semantically similar terms using WordNet [4] to estimate the similarity between different terms. The similarity between an expanded and re-weighted query q and a text d is computed as

$$S(Q, D) = \frac{\sum_i \sum_j q_i d_j \text{sim}(i, j)}{\sum_i \sum_j q_i d_j}, \tag{1.5}$$

where i and j are terms in the query and the query Q and document D respectively and $sim(i, j)$ denotes the semantic similarity between terms i and j [11, 4]. Query terms are expanded with synonyms and semantically similar terms (i.e., hyponyms and hypernyms) while document terms d_j are computed as $tf \cdot idf$ terms (they are neither expanded nor re-weighted).

SSRM outperforms VSM, the classic information retrieval method and demonstrates promising performance improvements over other semantic information retrieval methods in Web image retrieval based on text image descriptions extracted automatically [5]. SSRM can work in conjunction with

[4] http://wordnet.princeton.edu

any taxonomic ontology and any associated document corpus. Current research is directed towards extending SSRM to work with compound terms (phrases), and more term relationships (in addition to the Is-A relationships).

1.3 Image Link Analysis Methods

Effective content-based image retrieval on the Web often requires that important (authoritative) images satisfying the query selection criteria are assigned higher ranking over other relevant images. This is achieved by exploiting the results of link analysis for re-ranking the results of retrieval. Classical link analysis methods such as HITS [9], and PageRank [13] estimate the quality of Web pages and the topic relevance between the Web pages and the query. These methods estimate the importance of Web pages as a whole. PicASHOW [10], estimates the importance of images contained within Web pages. However, PicASHOW does not show how to handle image content and queries by image example. This is solved by WPicASHOW [25] (Weighted PicASHOW) a weighted scheme for co-citation analysis that incorporates, within the link analysis method of PicASHOW, the text and image content of the queries and of the Web pages.

1.3.1 PicASHOW

Co-citation analysis is proposed as a tool for assigning importance to pages or for estimating the similarity between a query and a Web page. A link from page a to page b may be regarded as a reference from the author of a to b. The number and quality of references to a page provide an estimate of the quality of the page and also a suggestion of relevance of its contents with the contents of the pages pointing to it.

HITS [9] exploits co-citation information between pages to estimate the relevance between a query and a Web page and ranking of this page among other relevant pages. The analysis results into pages on the topic of the query referred to as "authorities" and directory-like pages pointing to pages on the topic, referred to as "hubs" . HITS computes authority and hub values by link analysis on the *query focused graph* $\mathcal{F}$ (i.e., a set of pages formed by initial query results obtained by VSM expanded by backward and forward links). The page-to-page adjacency matrix W relates each page in $\mathcal{F}$ with the pages it points to. The rows and the columns in $\mathcal{W}$ are indices to pages in $\mathcal{F}$. Then, $w_{ij} = 1$ if page i points to page j; 0 otherwise. The authority and hub values of pages are computed as the principal eigenvectors of the page co-citation $W^T \cdot W$ and bibliographic matrices $W \cdot W^T$ respectively. The higher the authority value of an image the higher its likelihood of being relevant to the query.

Building upon HITS, PicASHOW [10] handles pages that link to images and to pages that contain images. PicASHOW demonstrates how to retrieve

high quality Web images on the topic of a keyword-based query. It relies on the idea that images co-contained or co-cited by Web pages are likely to be related to the same topic. Fig. 1.1 illustrates examples of co-contained and co-cited images. PicAHOW computes authority and hub values by link analysis on the query focused graph $\mathcal{F}$ as in HITS. PicASHOW filters out from $\mathcal{F}$ non-informative images such as banners, logo, trademarks and "stop images" (bars, buttons, mail-boxes etc.) from the query focused graph utilizing simple heuristics such as small file size.

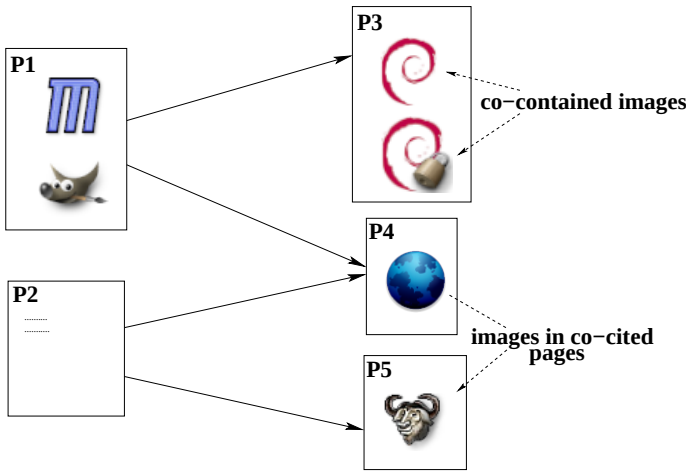

Fig. 1.1. The focused graph corresponding to query "Debian logo".

PicASHOW introduces the following adjacency matrices defined on the set of pages in the query focused graph:

$\mathcal{W}$: The page to page adjacency matrix (as in HITS) relating each page in $\mathcal{F}$ with the pages it points to. The rows and the columns in $\mathcal{W}$ are indices to pages in $\mathcal{F}$. Then, $w_{ij} = 1$ if page i points to page j; 0 otherwise.

$\mathcal{M}$: The page to image adjacency matrix relating each page in $\mathcal{F}$ with the images it contains. The rows and the columns in $\mathcal{M}$ ar indices to pages and images in $\mathcal{F}$ respectively. Then, $m_{ij} = 1$ if page i points to (or contains) image j.

$(\mathcal{W} + \mathcal{I})\mathcal{M}$: The page to image adjacency matrix ($\mathcal{I}$ is the identity matrix) relating each page in $\mathcal{F}$ both, with the images it contains and with the images contained in pages it points to.

Figure 1.2 illustrates these matrices for the pages $(P_1, P_2, \ldots P_5)$ and images of Figure 1.1. Notice that, in PicASHOW all non-zero values in $\mathcal{M}$, $\mathcal{W}$ and $(\mathcal{W} + \mathcal{I})\mathcal{M}$ matrices are 1 (non normalized weights). Figure 1.3.1 illustrates authority and hub values computed by PicASHOW in response to

	P_1	P_2	P_3	P_4	P_5
P_1	0	0	1	1	0
P_2	0	0	0	1	1
P_3	0	0	0	0	0
P_4	0	0	0	0	0
P_5	0	0	0	0	0

						
P_1	0	0	1	1	0	0
P_2	0	0	0	0	0	0
P_3	1	1	0	0	0	0
P_4	0	0	0	0	1	0
P_5	0	0	0	0	0	1

P_1	1	1	1	1	1	0
P_2	0	0	0	0	1	1
P_3	1	1	0	0	0	0
P_4	0	0	0	0	1	0
P_5	0	0	0	0	0	1

Fig. 1.2. Adjacency matrices $\mathcal{W}$, $\mathcal{M}$ and $(\mathcal{W}+\mathcal{I})\mathcal{M}$ for the focused graph of Figure 1.1.

query "Debian logo". Notice the high authority scores of pages showing logo or trademark images of "Debian Linux". Notice that Mozila trademark has higher authority value than Debian trademark.

Image						
Authorities	0.492	0.492	0.339	0.339	0.519	0.117

Page	P_1	P_2	P_3	P_4	P_5
Hubs	0.519	0.0001	0.854	0.001	0

Fig. 1.3. Image authority (top) and hub values (bottom) computed by PicASHOW in response to query "Debian trademark".

Hub and Authority values of images are computed as the principal eigenvectors of the image-co-citation $[(\mathcal{W}+\mathcal{I})\mathcal{M}]^{\mathcal{T}} \cdot (\mathcal{W}+\mathcal{I})\mathcal{M}$ and bibliographic matrices $(\mathcal{W}+\mathcal{I})\mathcal{M}) \cdot [(\mathcal{W}+\mathcal{I})\mathcal{M}]^{T}$ respectively. The higher the authority value of an image the higher its likelihood of being relevant to the query.

PicASHOW can answer queries on a given topic but, similarly to HITS, it suffers from the following problems [2]:

Mutual reinforcement between hosts: Encountered when a single page on a host points to multiple pages on another host or the reverse (when multiple pages on a host point to a single page on another host).

Topic drift: Encountered when the query focused graph contains pages not relevant to the query (due to the expansion with forward and backward links). Then, the highest authority and hub pages tend not to be related to the topic of the query.

1.3.2 Weighted PicASHOW (WPicASHOW)

PicASHOW cannot handle image content or image text context. This problem is addressed by WPicASHOW [25] (or Weighted PicASHOW), a weighted

scheme for co-citation analysis. WPicASHOW relies on the combination of text and visual content and on its resemblance with the query for regulating the influence of links between pages. Co-citation analysis then takes this information into account. WPicASHOW has been shown to achieve better quality answers and higher accuracy results (in terms of precision and recall) than PicASHOW using co-citation information alone [25].

WPicASHOW handles topic drift and mutual reinforcement as follows: Mutual reinforcement is handled by normalizing the weights of nodes pointing to k other nodes by $1/k$. Similarly, the weights of all l pages pointing to the same page are normalized by $1/l$. An additional improvement is to purge all intra-domain links except links from pages to their contained images. Topic drift is handled by regulating the influence of nodes by setting weights on links between pages. The links of the page-to-page relation $\mathcal{W}$ are assigned a relevance value computed by VSM and Eq. 1.6 as the similarity between the term vector of the query and the term vector of the anchor text on the link between the two pages. The weights of the page-to-image relation matrix $\mathcal{M}$ are computed by VSM and Eq. 1.7 (as the similarity between the query and the descriptive text of an image).

WPicASHOW starts by formulating the query focused graph as follows:

- An initial set R of images is retrieved. These are images contained or pointed-to by pages matching the query keywords according to Eq. 1.1.
- Stop images (banners, buttons, etc.) and images with logo-trademark probability less than 0.5 are ignored. At most T images are retained and this limits the size of the query focused graph ($T = 10,000$ in *IntelliSearch*).
- The set R is expanded to include pages pointing to images in R.
- The set R is further expanded to include pages and images that point to pages or images already in R. To limit the influence of very popular sites, for each page in R, at most t (e.g., $t = 100$) new pages are included.
- The last two steps are repeated until R contains T pages and images.

WPicASHOW then builds $\mathcal{M}$, $\mathcal{W}$ and $(\mathcal{W} + \mathcal{I})\mathcal{M}$ matrices for information in R. Fig. 1.4 illustrates these matrices for the example set R of Fig. 1.1 with weights corresponding to query "Debian logo". Notice that, in PicASHOW all non-zero values in $\mathcal{M}$ and $\mathcal{W}$ are 1 (non normalized weights).

Figure 1.3.2 illustrates authority and hub values computed by WPicASHOW in response to query "Debian logo". Notice the trademark images of "Debian Linux" are assigned the highest authority values followed by the images of "Mozilla Firefox".

1.4 Relevance Feedback

Relevance feedback [28, 30, 14] is the state-of-the-art approach for adjusting query results to the needs of the users. A common assumption is that there exists an ideal query (or matching method) that captures the information

	P_1	P_2	P_3	P_4	P_5
P_1	0	0	.6	.1	0
P_2	0	0	0	.1	.1
P_3	0	0	0	0	0
P_4	0	0	0	0	0
P_5	0	0	0	0	0

						
P_1	0	0	.1	.1	0	0
P_2	0	0	0	0	0	0
P_3	.8	.7	0	0	0	0
P_4	0	0	0	0	.2	0
P_5	0	0	0	0	0	.15

						
P_1	.48	.42	.1	.1	.02	0
P_2	0	0	0	0	.02	.015
P_3	.8	.7	0	0	0	0
P_4	0	0	0	0	.2	0
P_5	0	0	0	0	0	.15

Fig. 1.4. Adjacency matrices $\mathcal{M}$, $\mathcal{W}$ and $(\mathcal{M}+\mathcal{I})\mathcal{W}$ for the focused graph of Figure 1.1 corresponding to query "Debian logo".

Image						
Authorities	0.751	0.657	0.0418	0.0418	0.008	0

Page	P_1	P_2	P_3	P_4	P_5
Hubs	0.519	0.0001	0.854	0.001	0

Fig. 1.5. Image authority (top) and hub values (bottom) computed by WPicASHOW in response to query "Debian logo".

needs of the users. Relevance feedback attempts to guess the ideal query (or matching method) from answers that are initially obtained from the database. The users mark relevant (positive) or irrelevant (negative) examples among the retrieved answers, these examples are processed to form a new query which is combined with the original query and is resubmitted to the system. The process is repeated until convergence (i.e., the answers do not change). A categorization of methods includes:

Query point movement methods assuming that the ideal query is a point in a multi-dimensional space that the method approximates iteratively [16].

Term re-weighting methods that adjust the relative importance (weights) of terms in image representations [17, 6]. Terms that vary less in the set of positive examples are more important and should weigh more in retrievals. The inverse of the standard deviation is usually used for re-weighting the query terms.

Query expansion methods that attempt to guess an ideal query by adding new terms into the user's query [19, 3, 12].

Similarity adaptation methods that approximate the ideal matching method by substituting the system similarity (or distance) function with one that better captures the user's notion of similarity [27].

There are also approaches combining the above ideas. MindReader [6] combines query point movement and term re-weighting and handles correlations between attributes. Weight estimation is formulated as a minimization problem. MARS [18] is a prototype image retrieval system implementing

a variation of the standard term re-weighting method. iFind [12] supports keyword-based image search along with queries by image example. The main idea behind this approach is that images which are similar to the same query represent similar semantics. Images are linked to semantics by applying data mining on user's feedback log [3].

In the following, the existing framework of image retrieval with relevance feedback on the Web is extended to handle more sophisticated queries (e.g., queries by image example), by incorporating text and image content into the image retrieval and relevance feedback processes [14]. To do so, the concepts of text and image similarity of Sec. 1.1 are generalized as follows: The text similarity between a query Q and an image I is computed as

$$S^{text}(Q,I) = \sum_{i \in representation} w_i^{text} S_i^{text}(Q,I), \tag{1.6}$$

where w_i^{text} are weights (inner weights) denoting the relative significance of the above lists. Each S_i component is computed as list similarity: The more common terms (in the same order) two term lists have in common, the more similar they are. Similarly, The image similarity between a query image Q and an image I is computed as

$$S^{image}(Q,I) = \sum_{i \in representation} w_i^{image} S_i^{image}(Q,I), \tag{1.7}$$

where w_i^{image} are weights (inner weights) denoting the relative significance of the above types of image content representations. The computation of each S_i component depends on feature type: The similarity between histograms is computed by their intersection whereas the similarity between moment invariants is computed by subtracting the Euclidean vector distance from its maximum value.

To answer queries combining text and image example, the similarity between a query Q and a Web image I is computed as

$$S(Q,I) = W^{image} S^{image}(Q,I) + W^{text} S^{text}(Q,I), \tag{1.8}$$

where W^{text} and W^{image} are weights (outer weights) denoting the relative significance of image and text descriptions. All measures above are normalized to lie in the interval [0,1].

The inner and outer weights of Eq. 1.1, Eq. 1.2 and Eq. 1.8 place different emphasis on different features or representations respectively and can be used to adapt the query results to user's preferences. Typically, the weights are user defined. However, weight definition is beyond the understanding of most users. Relevance feedback is employed to estimate good weight values. Query expansion, term re-weighting and similarity adaptation methods are considered as representatives of most important categories of methods. Query point movement methods assume vector representations and cannot be applied. In

the following the basic steps of each method are discussed. The same steps are applied iteratively until convergence (i.e., the results of the retrieval method do not change). Initial results are obtained by applying either Eq. 1.1 (for text queries) or Eq. 1.8 (for queries combining text with image example). All weights are initialized to 1.

1.4.1 Query Expansion

The query is expanded with new terms obtained from positive examples. Two methods are evaluated. These methods work only with text.

Accumulation [19]: The most relevant image is selected from the answers and its text representation (i.e., a list of descriptive terms) is extracted. The query is matched with each term in this representation. A new query is formed by merging the query representation with the most similar terms of the most relevant image.

Integration and Differentiation [19]: Relevant and irrelevant images are selected from the answers. From each relevant image, its text representation (i.e., list of descriptive terms) is extracted and matched with the query. The most similar terms are combined to form a new "positive query". Similarly, the most dissimilar (to the query) terms are extracted from all irrelevant answers and combined to form a "negative query". The positive query is applied. Images which are more similar to the negative query rather than to the positive query are removed from the the answer.

1.4.2 Term Re-Weighting

Term re-weighting works by adjusting the relative importance of query terms [17]. The method is extended to accommodate for the definition of image similarity by text and image content as follows [14].

Let R be the set of the N_R most similar images (e.g., $N_R = 30$). A relevance score taking values -3 (for highly non-relevant answers) through 3 (for highly relevant answers) is assigned to each answer in R (neutral or no-opinion answers take score 0). R also denotes the query results at the beginning of each feedback cycle.

The outer weights W^j ($j \in \{text, image\}$) are dynamically updated during each feedback cycle: The database is queried by each S^j separately (using either Eq. 1.1 or Eq. 1.2) and its answer set R^j is sorted by similarity. The weights are then updated according to the following formula

$$W^j = \begin{cases} W^j + score_I \text{ if } I \in R, \\ W^j + 0 \quad \text{otherwise;} \end{cases} \tag{1.9}$$

where $score_I$ is the score assigned to image I in R. Initially all $W^j = 0$. After iterating over the images in each R^j all weights W_i^j are normalized by $W_{total}^j = \sum_{I \in R^j} W^j$. Negative weights are set to 0.

The inner weights $w_i^j (j \in \{text, visual\})$ for each term i of the text or image representation are also dynamically updated using the set R' of relevant answers in R ($R' \subset R$): The smaller the variance of each S_i^j the larger the significance of the i-th term (and the reverse). Therefore, $w_i^j = 1/\sigma_i^j$, where σ_i^j is the variance of the i-th feature in the j-th representation. Each weight is normalized by $w_{total}^j = \sum_{I \in R'} w_i^j$.

1.4.3 Similarity Adaptation

Falcon [27] estimates an ideal distance function $\mathcal{D}_\mathcal{G}$ that retrieves the best results. Initially, Falcon searches the database using $d(Q, I) = 1 - S(Q, I)$ as distance function and the user adds positive examples to a set $\mathcal{G}$ (initially empty). During a feedback cycle, Falcon searches the database again using a new distance function $\mathcal{D}_\mathcal{G}$ while the user adds new positive examples to $\mathcal{G}$. The distance between the query Q and a Web image I is computed as the distance of I from the current members of $\mathcal{G}$. Falcon estimates $\mathcal{D}_\mathcal{G}$ iteratively as follows

$$\mathcal{D}_\mathcal{G}(I) = \begin{cases} 0 & \text{if } \exists i : d(g_i, I) = 0, \\ \left(\frac{1}{k} \sum_{i=1}^{k} d(g_i, I)^\alpha\right)^{1/\alpha} & \text{otherwise;} \end{cases} \tag{1.10}$$

where k is the number of positive examples in $\mathcal{G}$, g_i is a member of $\mathcal{G}$ and α is a user defined constant (e.g., $\alpha = -5$).

1.5 IntelliSearch

All methods previously stated are implemented and integrated into *IntelliSearch* [24], a complete and fully automated system for retrieving text pages and images on the Web. It provides an ideal test-bed for experimentation and training and serves as a framework for a realistic evaluation of retrieval methods for the Web. The system stores a crawl of the Web with 1,5 million Web pages with images. The system is implemented in Java and is accessible on the Web [5]. The architecture of *IntelliSearch* is illustrated in Fig. 1.6. It consists of several modules, the most important of them being the following:

Crawler module: Implemented based upon Larbin [6], the crawler assembled locally a collection of 1,5 million pages with images. The crawler started its recursive visit of the Web from a set of 14,000 pages which is assembled from the answers of Google image search to 20 queries on topics related to Linux and Linux products. The crawler worked recursively in breadth-first order and visited pages up to depth 5 links from each origin.

[5] http://www.intelligence.tuc.gr/intellisearch
[6] http://larbin.sourceforge.net

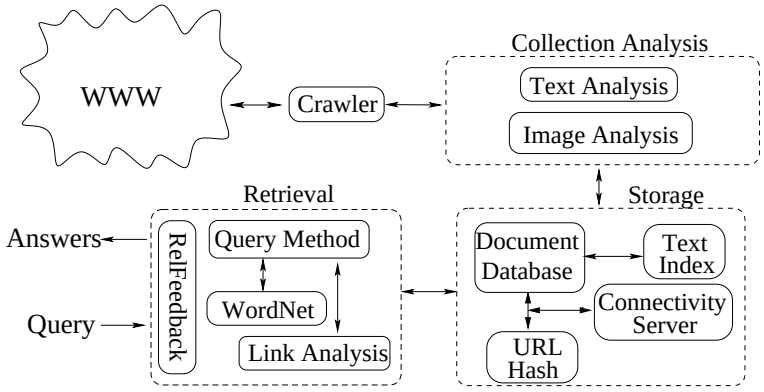

Fig. 1.6. *IntelliSearch* Architecture.

Collection analysis module: The content of crawled pages is analyzed. Text, images, link information (forward links) and information for pages that belong to the same site is extracted.

Storage module: Implements storage structures and indices providing fast access to Web pages and information extracted from Web pages (i.e., text, image descriptions and links). For each page, except from raw text and images, the following information is stored and indexed: Page URLs, image descriptive text (i.e., alternate text, caption, title, image file name), terms extracted from pages, term inter document frequencies (i.e., term frequencies in the whole collection), term intra document frequencies (i.e., term frequencies in image descriptive text parts), link structure information (i.e., backward and forward links). Image descriptions are also stored.

Retrieval module: Queries are issued by keywords or free text. The user is prompted at the user interface to select mode of operation (retrieval of text pages or image retrieval).

The Entity Relationship Diagram (ERD) of the database in Fig. 1.7 describes entities (i.e., Web pages) and relationships between entities. There are many-to-many (denoted as $N : M$) relationships between Web pages implied by the Web link structure (by forward and backward links), one-to-many (denoted as $1 : N$) relationships between Web pages and their constituent text and images and $N : M$ relationships between terms in image descriptive text parts and documents and. The ERD also illustrates properties of entities and relationships (i.e., page URLs for documents, titles for page text, image content descriptions for images, stemmed terms, inter and intra document frequencies for terms in image descriptive text parts.)

The database schema is implemented in BerkeleyDB [7] Java Edition. BerkeleyDB is an embedded database engine providing a simple Application Pro-

[7] http://www.sleepycat.com

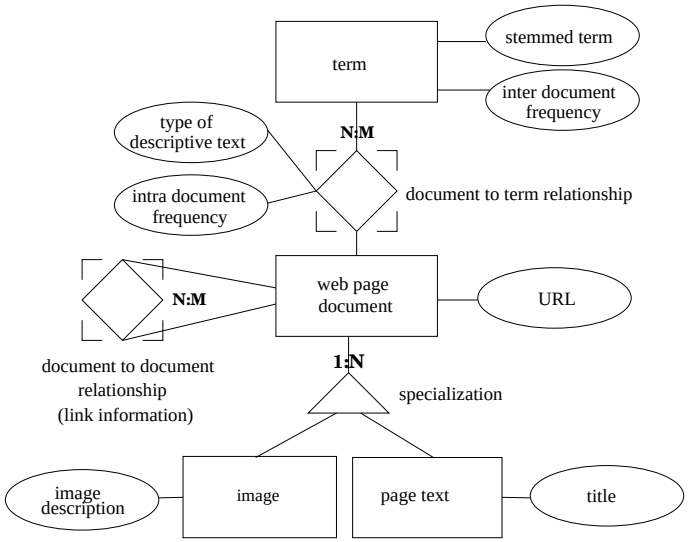

Fig. 1.7. The Entity Relational Diagram (ERd) of the database.

gramming Interface (API) supporting efficient storage and retrieval of Java objects. The mapping of the ERD of Fig. 1.7 to database files (Java objects) was implemented using the Java Collections-style interface. Apache Lucene[8] is providing mechanisms (i.e., inverted files) for indexing text and link information. There are Hash tables for URLs and inverted files for terms and link information. Two inverted files implement the connectivity server [2] and provide fast access to linkage information between pages (backward and forward links) and two inverted files associate terms with their intra and inter document frequencies and allow for fast computation of term vectors.

1.6 Conclusions

This Chapter presents comparative study of several retrieval methods for the Web with emphasis on methods for retrieving images by content. Several aspects of the problem of content-based image retrieval on the Web are examined including retrieval by text, text semantics, image content features and retrieval by authority (importance) characteristics. Relevance feedback is also discussed in this context is a tool for adjusting the retrieved results to the actual needs of the users.

The experimental results [25] demonstrate that Web search methods utilizing content information (or combination of content and link information)

[8] http://lucene.apache.org

perform significantly better than methods using link information alone. Link analysis improved the quality of the results but not necessarily their accuracy (at least for data sets smaller than the Web). The analysis revealed that content relevance and searching for authoritative answers can be traded-off against each other: Giving higher ranking to important pages seems to reduce the accuracy of the results (i.e., link analysis methods tend to assign higher ranking to higher quality but not necessary relevant pages. High quality pages, on the other hand, may be irrelevant to the content of the query. Weighted link analysis methods (WHITS, WPicASHOW) attempted to compromise between text and link analysis methods.

Text searching methods like Vector Space Model (VSM), the same as semantic retrieval methods are far more effective than link analysis methods implying that text is a very effective descriptor of Web content itself. Between the two, semantic retrieval methods demonstrated promising performance improvements over VSM[5]. However, text similarity methods tend to assign higher ranking even to Web pages and images pointed to by very low quality pages (e.g., pages created by individuals or small companies).

The size of the data set is also a problem. If the queries are very specific, the set of relevant answers is small and within it, the set of high quality and relevant answers are even smaller. The results may improve with the size of the data set, implying that it is plausible for the method to perform better when applied to the whole Web.

The evaluation of relevance feedback methods [14] demonstrated that term re-weighting based on text and image content is the most effective approach. The results demonstrate that term re-weighting is the most effective relevance feedback approach for all query types. Term re-weighting allows also for much smaller iteration cycles (and therefore for faster retrieval with less users effort) while maintaining good performance. All methods converge very fast (i.e., after two iteration cycles).

Future work includes experimentation with larger data sets and more elaborate detection and matching methods for more image types. Extracting semantic information from Web pages (image concepts and relationships) through automatic text analysis, combining text with image features as well as representing this information by image ontologies, is another aspect of future research. Image ontologies would not only serve as a means for bridging the semantic gap between image features and concepts, but also as a means for more effective image content representation and for supporting semantically rich query answering on the Web.

References

1. A. Arasu, J. Cho, H. Garcia-Molina, A. Paepke, and S. Raghavan. Searching the Web. *ACM Transactions on Internet Technology*, 1(1):2–43, Aug. 2001.

2. K. Bharat and M. R. Henzinger. Improved Algorithms for Topic Distillation in a Hyperlinked Environment. In *Proc. of SIGIR-98*, pages 104–111, Melbourne, 1998.
3. Z. Chen, L. Wenyin, F. Zhang, M. Li, and H.-J Zhang. Web Mining for Web Image Retrieval. *Journal of the American Society of Information Science*, 52(10):831–839, 2001.
4. A. Hliaoutakis P. Raftopoulou E. G.M. Petrakis, G. Varelas. X-Similarity: Computing Semantic Similarity between Concepts from Different Ontologies. *Journal of Digital Information Management (JDIM)*, 4(4):233–238, December 2006.
5. A. Hliaoutakis, G. Varelas, E. Voutsakis, E. G.M. Petrakis, and E. Milios. Information Retrieval by Semantic Similarity. *Intern. Journal on Semantic Web and Information Systems (IJSWIS)*, 3(3):55–73, July/Sept. 2006. Special Issue of Multimedia Semantics.
6. Y. Ishikawa, R. Subramanya, and C. Faloutsos. Mindreader: Query Databases Through Multiple Examples. In *Proc. of the 24^{th} VLDB Conference*, pages 218–227, New York. USA, 1998.
7. J.-Hu and A. Bagga. Identifying Story and Preview Images in News Web Pages. In *7^{th} Intern. Conf. on Document Analysis and Recognition (ICDAR'2003)*, pages 640–644, Edinburgh, Scotland, Aug. 2003.
8. M.L. Kherfi, D. Ziou, and A. Bernardi. Image Retrieval from the World Wide Web: Issues, Techniques, and Systems. *ACM Computing Surveys*, 36(1):35–67, March 2004.
9. J. M. Kleinberg. Authoritative Sources in a Hyperlinked Environment. *Journal of the ACM*, 46(5):604–632, 1999.
10. R. Lempel and A. Soffer. PicASHOW: Pictorial Authority Search by Hyperlinks on the Web. *ACM Transactions on Information Systems*, 20(1):1–24, Jan. 2002.
11. Y. Li, Z. A. Bandar, and D. McLean. An Approach for Measuring Semantic Similarity between Words Using Multiple Information Sources. *IEEE Trans. on Knowledge and Data Engineering*, 15(4):871–882, July/Aug. 2003.
12. Y. Lu, C. Hu, X. Zhu, H.-J. Zhang, and Q. Yang. A Unified Framework for Semantic and Feature Based Relevance Feedback in Image Retrieval Systems. In *Proc. ACM Multimedia*, pages 31–37, Los Angeles CA, USA, 2000.
13. L. Page, S. Brin, R. Motwani, and T. Winograd. The PageRank Citation Ranking: Bringing Order to the Web. Technical report, Computer Systems Laboratory, Stanford Univ., CA, 1998.
14. E. G.M. Petrakis, K. Kontis, E. Voutsakis, and E. Milios. Relevance Feedback Methods for Logo and Trademark Image Retrieval on the Web. In *ACM Symposium on Applied Computing (ACM SAC'2006)*, pages 1084–1088, Dijon, France, April 23-27 2006. Special Track on Information Access and Retrieval (IAR).
15. Eds R. Baeza-Yates. *Modern Information Retrieval.* Addison Wesley, 1999.
16. J.J. Rocchio. Relevance Feedback in Information Retrieval. In G. Salton, editor, *The SMART Retrieval System - Experiments in Automatic Document Processing*, pages 313–323. Prentice Hall, Englewood Cliffs, 1971.
17. Y. Rui, T.-S. Huang, M. Ortega, and S. Mechrota. Relevance Feedback: A Power Tool for Interactive Content-Based Image Retrieval. *IEEE Trans. on Circ. and Syst. for Video Technology*, 8(5):644–655, Sept. 1998.
18. Y. Rui, T.S. Huang, and S. Mehrotra. Content-Based Image Retrieval with Relevance Feedback in MARS. In *Proc. IEEE Int. Conf. on Image Processing*, pages 515–518, Santa Barbara, CA, Oct. 1997.

19. H.-T. Shen, B.-Chin Ooi, and K.-Lee Tan. Giving Meanings to WWW Images. In 8^{th} *Intern. Conf. on Multimedia*, pages 39–47, Marina del Rey, CA, 2000.
20. A. W.M. Smeulders, M. Worring, S. Santini, A. Gupta, and R. Jain. Content-Based Image Retrieval at the End of the Early Years. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(11):1349–1380, Dec. 2000.
21. J.R. Smith and S.-Fu Chang. Visually Searching the Web for Content. *IEEE Multimedia*, 4(3):12–20, July-Sept. 1997.
22. M. Sonka, V. Hlavec, and R. Boyle. *Image Processing Analysis, and Machine Vision*, chapter 6 & 14. PWS Publishing, 1999.
23. L. Taycher, M.La Cascia, and S. Sclaroff. Image Digestion and Relevance Feedback in the ImageRover WWW Search Engine. In 2^{nd} *Intern. Conf. on Visual Information Systems*, pages 85–94, San Diego, Dec. 1997.
24. E. Voutsakis, E. G.M. Petrakis, and E. Milios. IntelliSearch: Intelligent Search for Images and Text on the Web. In *Image Analysis and Recognition (ICIAR 2006)*, pages 697–708, Povoa de Varzim, Portugal, Sept. 18-20 2006.
25. E. Voutsakis, E.G.M. Petrakis, and E. Milios. Weighted link analysis for logo and trademark image retrieval on the web. In *IEEE/WIC/ACM International Conference on Web Intelligence (WI2005)*, pages 581–585, Compiegne, France, Sept. 2005.
26. I. H. Witten and E. Frank. *Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations*, chapter 4. Morgan Kaufmann, Academic Press, 2000.
27. L. Wu, K. Sycara, T. Payne, and C. Faloutsos. FALCON: Feedback Adaptive Loop for Content-Based Retrieval. In *Proc. 26^{th} VLDB Conf.*, pages 297–306, Cairo, Egypt, Sept. 2000.
28. H.-J. Zhang, Z. Chen, W.-Y. Liu, and M. Li. Relevance Feedback in Content-Based Image Search. In *Proc. 12^{th} Intern. Conf. on New Information Technology (NIT)*, pages 29–31, Beijing, China, Aug. 2003. (invited keynote).
29. R. Zhao and W.I. Grosky. Narrowing the Semantic GapImproved Text-Based Web Document Retrieval Using Visual Features. *IEEE Transactions on Multimedia*, 4(2):189–200, June 2002.
30. X.-S. Zhou and T.S. Huang. Relevance Feedback in Image Retrieval: A Comparative Study. *Multimedia Systems*, 8(6):536–544, April 2003.

Index