PicASHOW: Pictorial Authority Search by Hyperlinks on the Web

Ronny Lempel
Department of Computer Science
The Technion
Haifa 32000, Israel
rlempel@cs.technion.ac.il

Aya Soffer
IBM Research Lab in Haifa
Matam
Haifa 31905, Israel
ayas@il.ibm.com

ABSTRACT

We describe PicASHOW, a fully automated WWW image retrieval system that is based on several link-structure analyzing algorithms. Our basic premise is that a page p displays (or links to) an image when the author of p considers the image to be of value to the viewers of the page. We thus extend some well known link-based WWW *page retrieval* schemes to the context of image retrieval.

PicASHOW's analysis of the link structure enables it to retrieve relevant images even when those are stored in files with meaningless names. The same analysis also allows it to identify *image containers* and *image hubs*. We define these as Web pages that are rich in relevant images, or from which many images are readily accessible.

PicASHOW requires no image analysis whatsoever and no creation of taxonomies for pre-classification of the Web's images. It can be implemented by standard WWW search engines with reasonable overhead, in terms of both computations and storage, and with no change to user query formats. It can thus be used to easily add image retrieving capabilities to standard search engines.

Our results demonstrate that PicASHOW, while relying almost exclusively on link analysis, compares well with dedicated WWW image retrieval systems. We conclude that link analysis, a bona-fide effective technique for Web page search, can improve the performance of Web image retrieval, as well as extend its definition to include the retrieval of image hubs and containers.

Keywords

Image Retrieval; Link Structure Analysis; Hubs and Authorities; Image Hubs.

1. INTRODUCTION

The WWW is host to millions of images on almost every conceivable topic. Finding effective methods to retrieve these images has attracted many research efforts over the past few years. Such research has led to academic image retrieval systems ([10]), to general search engines with image retrieval capabilities ([7], [17]), and to search engines dedicated to WWW image retrieval ([9],[19]). There are three main approaches for WWW image search and retrieval:

1. Text-based retrieval. This approach annotates images with text derived from the HTML documents which contain (display) them, and then applies text-based retrieval algorithms to the annotated collection of images. The derived text can include the caption of the image, text surrounding the image, the entire text of the containing page, the filename of the containing HTML document and the filename of the image itself.

2. Content-based image retrieval (CBIR). This approach applies image analysis techniques in order to extract visual features (e.g., color, texture, orientation, shapes) from the images. The features are extracted in a preprocessing stage, and stored in the retrieval system's database. The extracted features (e.g., the color histogram of the image) are usually of high dimensionality, and in order to allow scalability of these systems (in terms of storage space and query processing times), some sort of dimension reduction is usually performed on the data (e.g., [21]).

3. Manually annotated image collections. There are several companies that specialize in providing visual content to a diverse range of image consumers. The two largest companies are Getty Images ([15]) and Corbis ([14]). The images are indexed and retrieved by keywords, which are manually assigned to each image. While end users may use these services, they are especially geared toward companies and professionals which require high volumes of diverse images.

One implication of the difference between the retrieval approaches is their support of different types of queries. Text-based retrieval systems, as well as the commercial image providers, support natural, topic-descriptive queries. These queries are friendly and familiar to the typical surfer of the Web. On the other hand, CBIR supports either queries which are formulated in terms of the extracted visual features, or similarity queries, in which a sample image is presented and the system is required to retrieve images with similar visual features.

WWW10, May 1-5, 2001, Hong Kong.
ACM 1-58113-348-0/01/0005.

Web Image Search

PicToSeek ([12]) is an example of a pure content-based image retrieval system. It classifies images into portraits, photographs of indoor/outdoor scenes, and synthetic images. It extracts many visual features from the images, and supports queries by image-examples and by image features.

Many WWW image search engines combine text-based retrieval and CBIR into an integrated system. In *webSeek* ([27]), for example, each image is processed and its visual features are extracted. Each image is also associated with the text in its containing page. The images are then classified into topics from a taxonomy that was developed for this purpose. The *WebSeer* system ([11]) uses associated text and feature extraction to support complex queries, which state both the search topic and some visual properties of the desired images. The system depicted in [3] unifies the textual representations (derived from the containing pages) and visual representations of images into a single representative vector. This enables the utilization of possible statistical couplings between the textual contents of the containing pages and the visual properties of the images.

Harmandas et al. in [13] suggest a text-based image retrieval system in which connectivity information is used to induce textual annotations of images. Each image i in their approach is assigned a weighted vector of representing terms, which is some function of the combined text of all of the pages that contain i and of the pages that link to pages containing i. While this scheme does consider hyperlinks, it is essentially a text-based retrieval scheme in which hyperlinks are used to induce textual annotations, without any analysis being done on the link-structure per se.

Link Structure Analysis in Web Page Search

Recent work in Web search has demonstrated that link structure analysis is very effective in finding authoritative Web pages. Information such as which pages are linked to others is commonly used to augment search algorithms, and has significantly improved the ability of search engines to rank quality pages at the top of their search results. Link-structure analysis is based on the notion that a link from page p to page q can be viewed as an endorsement of q by p, and as some form of positive judgment by p of q's content.

Two important types of techniques in link-structure analysis are co-citation based schemes, and random-walk based schemes. The main idea behind co-citation based schemes is the notion that when two pages p_1 and p_2 both point to some page q , it is reasonable to assume that p_1 and p_2 share a mutual topic of interest. Likewise, when p links to both q_1 and q_2, it is probable that q_1 and q_2 share some mutual topic. An important work in the context of co-citation based schemes was Jon Kleinberg's introduction of the notions of *hubs* and *authorities* ([23]) as two distinct types of Web pages. Authorities, or authoritative pages, are Web pages that contain high-quality information regarding some topic. Hubs, on the other hand, may not directly contain information but are rather resource lists, linking to authorities on a topic without necessarily displaying the information itself. Kleinberg devised an algorithm aimed at finding authoritative pages, and researchers from IBM's Almaden Research Center have implemented Kleinberg's algorithm in various projects, most notably CLEVER ([4]).

Random walk based schemes model the Web (or part of it) as a graph (where pages are nodes and links are edges), and apply some random walk model to the graph. Pages are then ranked by the probability of visiting them in the modeled random walk. The most notable algorithm of this type is PageRank ([2]), which is an important part of the ranking function and of the success of the Google search engine ([16]). Both Kleinberg's algorithm and PageRank are described in detail in Section 2.

Co-citation reasoning was combined with random walk theory in SALSA ([24]), to separate the random walk based rankings of hubs and authorities. Rafiei and Mendelzon ([25]) have also integrated co-citation and random walks in their work on computing page reputations.

Our Approach: Link Structure Analysis in Web Image Search

In this paper we present PicASHOW, a pictorial retrieval system that searches for images (pictures) on the Web using hyperlink-structure analysis. PicASHOW applies co-citation based approaches and PageRank influenced methods. Our basic premise is that a page p displays (or links to) an image when the author of p considers the image to be of value to the viewers of the page. We further contend that the standard reasoning behind the co-citation measure applies to images just as it does to HTML pages:

- Images which are co-contained in pages are likely to be related to the same topic.
- Images which are contained in pages that are co-cited by a certain page are likely related to the same topic.

In addition, in the spirit of PageRank, we assume that images which are contained in authoritative pages on topic t are good candidates to be quality images on that topic.

In the next sections, we describe several link-structure based WWW image retrieval schemes. Following are the highlights of the PicASHOW approach:

- Our method can be implemented, with reasonable overhead, by standard WWW search engines. It can thus be used to add image retrieving capabilities to these engines. We elaborate on this in Section 3.2.
- Our schemes require no image analysis whatsoever. This eliminates the need to deal with high-dimension image descriptions, and with the complexity which such representations introduce in terms of memory requirements, preprocessing overhead, query processing and retrieval operations.
- No change to the query format is required. The same queries which are used to retrieve pages, will be used to retrieve images. In particular, users do not need to present the system with sample images, nor do they need to formulate queries in terms of image properties.
- There is no need to create taxonomies for pre-classification of the wealth of images on the Web.
- We do not rely solely on file names and image captions assigned by content creators. Thus, we are are able to find images related to a query with meaningless file names such as "myimages/image1" (most text-based image search engines will miss these). We can also find images with titles that are only semantically related to the query. For example images labeled "Bridal Veil Falls", when searching for images of Yosemite.

- In addition to finding authoritative images, we are also able to locate image containers and image hubs. We define these as Web *pages* that are rich in relevant images, or from which many images are readily accessible. See Section 4 for more details.

- A natural modification of our methods allows for the support of similarity queries [1], where users present PicASHOW with URLs of images on the topic in question. The system will then find other authoritative images on the same topic. We believe this is a very useful feature. Section 5 elaborates on the details.

The remainder of this paper is organized as follows. In Section 2 we provide some background on link-structure analysis when searching for Web pages. In Section 3 we formally define the image collections which are to be analyzed, explain how such collections are assembled from a given query, and present our proposed image ranking schemes. Section 4 introduces the concept of image hubs and image containers and describes how we identify such hubs and containers. In Section 5 we discuss the pros and cons of our method, and suggest interesting extensions of this research direction. Appendix A lists the URLs of the images which are displayed in this paper.

We do not provide any formal evaluation section in this paper since there are no benchmarks for testing such systems and thus most evaluations are qualitative. Rather, we include results of sample queries throughout the paper. For comparison purposes, we also show the results of some of these queries on commercial Web search engines. The purpose of this comparison is to show that different search techniques yield different images and to highlight the benefits of link analysis in the context of image search.

2. LINK ANALYSIS FOR FINDING AUTHORITATIVE WEB PAGES

This section provides some technical background on applications of WWW link-structure analysis when searching for Web pages. Specifically, we provide a brief overview of two link-structure analyzing approaches: PageRank ([2]) and Kleinberg's Mutual Reinforcement approach ([23]). This background is required in order to describe our image ranking schemes which are inspired by these approaches. Indeed, for each of the two approaches described, we also describe the main points which we adapted and evolved in our image ranking schemes.

2.1 PageRank

PageRank ([2]) is an important part of the ranking function of the Google search engine ([16]). The PageRank of a page p is the probability of visiting p in a random walk of the entire Web, where the set of states of the random walk is the set of pages, and each random step is of one of the following two types:

1. From the given state s, choose at random an outgoing link of s and follow that link to the destination page.
2. Choose a Web page uniformly at random, and jump to it.

PageRank chooses a parameter d, $0 < d < 1$, and each state transition is of the first transition type with probability d and of the second type with probability $1 - d$. The PageRanks obey the following formula (where page p has incoming links from pages $q_1, \ldots, q_k$):

$$\text{PageRank}(p) = (1-d) + d\left(\sum_{i=1}^{k} \frac{\text{PageRank}(q_i)}{\text{out degree of } q_i}\right)$$

Thus, the PageRank of a page grows with the importance (=PageRanks) of the pages which point to it. An endorsement (=link) from a prominent (high ranking) site, like Yahoo! ([20]), contributes to a page's PageRank much more than an incoming link from some obscure personal homepage. Our image ranking schemes will imitate this property. In particular, the rankings of images will grow with the importance of the pages which contain them.

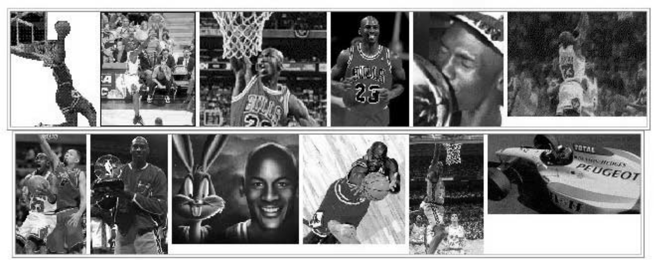


Figure 1: PicASHOW/Scour images for the query *"Michael Jordan"*

2.2 The Mutual Reinforcement Approach

Kleinberg's Mutual Reinforcement approach, introduced in [23], aims to find hubs and authorities which pertain to a given topic t. The key observation behind the approach is that hubs and authorities which pertain to t display a *mutually reinforcing* relationship: For a page to be considered a good t-hub, it must point to many t-authorities, while a page is considered to be an authority on topic t only if many hubs deem it as such (and point to it). Because of this relationship, prominent t-hubs and t-authorities tend to form *communities*, which can be seen as densely inter-connected bipartite portions of the Web-graph.

The algorithm starts by assembling a collection $\mathcal{C}$ of Web pages, which should contain many high quality Web pages which pertain to a given topic t. It then analyzes the link structure induced by that collection, in order to find the authoritative pages (and the hubs) on topic t.

Denote by q a term-based search query which describes the topic of interest t. The collection $\mathcal{C}$ is assembled as follows:

- A *root set* S of pages is obtained by applying a term based search engine, such as AltaVista ([7]), to the query q. This is the only step in which the lexical content of the Web sites is examined.

- From S, a *base set* $\mathcal{C}$ is derived, that consists of (a) pages in the root set S, (b) pages that point to a page in S and (c) pages that are pointed to by a page in S.

[1]Note that this is different than the similarity queries mentioned in the context of CBIR systems - our method will find images on the same topic as the sample images rather than images with similar visual properties.

The collection $\mathcal{C}$ and its link structure induce a $|\mathcal{C}| \times |\mathcal{C}|$ adjacency matrix, which is denoted by W.

Each page $s \in \mathcal{C}$ is then assigned a pair of weights, a hub-weight $h(s)$ and an authority weight $a(s)$, based on the following two principles:

- The quality of a hub is determined by the quality of the authorities it points at.
- A page is only as authoritative as the quality of the hubs which deem it as such.

The top ranking pages, according to both types of weights, form the Mutually Reinforcing communities of hubs and authorities.

Kleinberg uses the following iterative algorithm to assign the weights:

1. Initialize $a(s) \leftarrow 1,\ h(s) \leftarrow 1$ for all pages $s \in \mathcal{C}$.
2. Repeat the following operations until convergence:
   - Update the authority weight of each page s (the $\mathcal{I}$ operation): $a(s) \leftarrow \sum_{\{x|x\ points\ to\ s\}} h(x)$
   - Update the hub weight of each page s (the $\mathcal{O}$ operation): $h(s) \leftarrow \sum_{\{x|s\ points\ to\ x\}} a(x)$
   - Normalize both sets of hub and authority weights.

Note that applying the $\mathcal{I}$ operation is equivalent to assigning authority weights according to the result of multiplying the vector of all hub weights by the matrix W^T. The $\mathcal{O}$ operation is equivalent to assigning hub weights according to the result of multiplying the vector of all authority weights by the matrix W.

Kleinberg showed that this algorithm converges, and that the resulting authority weights [hub weights] are the coordinates of the normalized principal eigenvector[2] of W^TW [of WW^T]. The pages which correspond to the largest coordinates of these eigenvectors are returned by the algorithm as the principal community of authorities[hubs].

The two matrices W^TW and WW^T are well known in the field of bibliometrics:

1. W^TW is the *co-citation matrix* ([26]) of the collection. $[W^TW]_{i,j}$ is the number of pages which jointly point at (cite) pages i and j.
2. WW^T is the *bibliographic coupling matrix* ([22]) of the collection. $[WW^T]_{i,j}$ is the number of pages jointly referred to (pointed at) by pages i and j.

It is important to note that the outcome of the algorithm, namely the communities of hubs and authorities which the algorithm will identify, are determined solely by the adjacency matrix W of the collection $\mathcal{C}$. The adjacency matrix implies the co-citation and bibliographic coupling matrices, and it is the eigenvectors of these matrices, in turn, which uniquely determine the principal communities of hubs and authorities.

Our co-citation based image retrieval schemes basically imitate this algorithm, albeit with different adjacency matrices. Defining the adjacency matrices to be used will suffice to uniquely define our schemes.

[2]The eigenvector which corresponds to the eigenvalue of highest magnitude of the matrix.

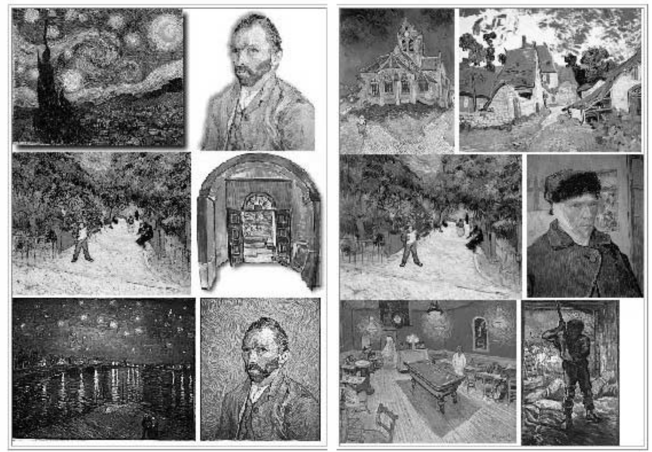


Figure 2: PicASHOW (left)/AltaVista images for the query *Vincent Van Gogh*

3. LINK STRUCTURE ANALYSIS FOR FINDING AUTHORITATIVE IMAGES

We now describe our method for finding authoritative images given a query. First, we formally define the image collections which are analyzed. Next, we explain how such collections are assembled from a given query. Finally, we present our image ranking schemes.

3.1 Formal Definition of the Model

A page p is said to contain an image i (denoted by $p \rightsquigarrow i$) in either of the following two cases:

1. p displays i: When page p is loaded in a Web browser, i is displayed. For example, either of the following html directives:
   `<IMG source="foo.gif" alt="nice image">` or
   `<A HREF="foo.html"><IMG source="foo.gif"></A>`.
2. p points to i's image file (in some image file format such as .gif or .jpeg). For example,
   `<A HREF="foo.jpeg">nice image</A>`. Note that p does not contain i when p points to an HTML file which contains i, even if i is the only visible object in the HTML file.

We define a topical *WWW image collection* as a quadruple $\mathcal{IC} = (\mathcal{P}, \mathcal{I}, \mathcal{L}, \mathcal{E})$, where $\mathcal{P}$ is a set of Web pages (many of which deal with a certain topic t), $\mathcal{I}$ is the set of images which are contained in $\mathcal{P}$, $\mathcal{L} \subseteq \mathcal{P} \times \mathcal{P}$ is the set of (directed) links which exist on the Web between the pages of $\mathcal{P}$, and $\mathcal{E} \subseteq \mathcal{P} \times \mathcal{I}$ is the relation *page p contains image i*.

We denote by W the adjacency matrix of the page-to-page relation $\mathcal{L}$, and by $M = [m_{ij}]$ the $|\mathcal{P}| \times |\mathcal{I}|$ adjacency matrix of the page-to-image relation $\mathcal{E}$.

Page-Image Adjacency

Two of the most important observations of link-structure analysis are the following:

1. The notion of authority being conferred through links from a pointing resource to a pointed resource. In our context, the pointing resources are Web pages, while the pointed resources are the images.

2. The topical similarity between resources (images, in our case) which is inferred through co-citation.

Both principles are reflected in the adjacency relation which exists in the data. The adjacency matrix conveys the flow of authority, while the entries of the co-citation matrix, which the adjacency matrix implies, define the strength of the topical affinities between the resources. We therefore aim to define an adjacency relation between Web pages and images, in a manner which best reflects both the flow of authority from pages to images, and the topical affinities between the various images.

There are several reasonable definitions for such adjacency relations. The intuition behind these definitions is perhaps best explained through an example. Consider the scenario depicted in Figure 3, which consists of five Web pages $P_1, \ldots, P_5$ and four images, one of them replicated.

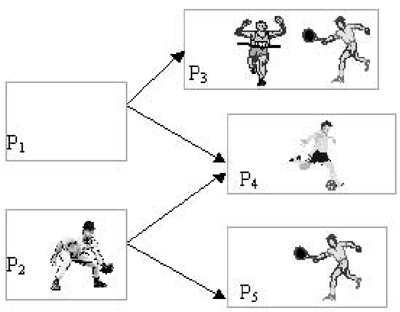
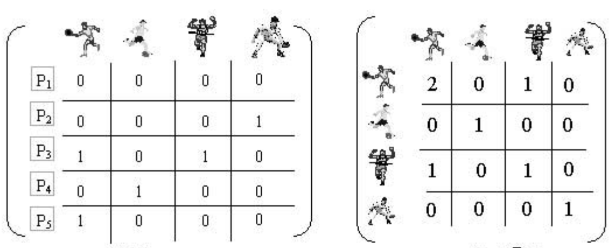


Figure 3: An example page/image case

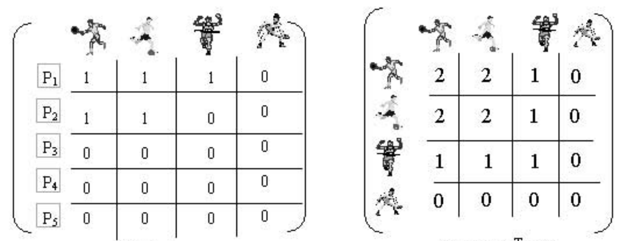
The most basic adjacency relation which comes to mind is to adopt the page-to-image relation $\mathcal{E}$, defined above, as the adjacency relation (and M as the adjacency matrix). M represents the outright manner for a page to endorse an image, which is simply to display it or point to it. This approach also reflects some topical affinities between images, through the corresponding co-citation matrix $M^T M$. Pairs of images which are co-contained in the same page are considered topically related. For example, note the entry which corresponds to the runner and tennis player in Figure 4.



(a) M (b) $M^T M$

Figure 4: The $k = 0$ adjacency and co-citation matrices for the example case

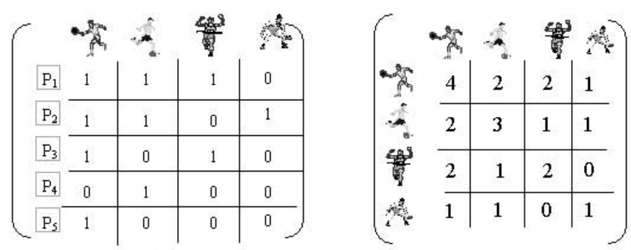
However, this approach fails to convey some other fairly intuitive topical relations, such as between the soccer player and the tennis player. These images appear in pages which are co-cited, and assuming we agree that co-cited pages are topically related, then perhaps so are the images which are contained in them. To reflect such a connection, we need to consider the adjacency matrix WM, which associates each page p with the images which are displayed *in pages to which* p *links*. Using WM as the adjacency matrix, the co-citation matrix $M^T W^T W M$ (Figure 5) now reflects some topical affinity between the soccer image and the tennis image.



(a) W M (b) $(WM)^T WM$

Figure 5: The $k = 1$ adjacency and co-citation matrices for the example case

This approach also suffers from some obvious setbacks. The affinity between the soccer image and the runner image is considered as strong as the affinity between the tennis image and the runner, although it seems logical that the tennis and runner images are more tightly coupled, since they appear in the same page. A greater problem exists with the connection between the image of the baseball player and the images of the soccer and tennis players. Why is a linkage between the soccer image and the tennis image inferred by P_2 co-citing pages P_4 and P_5, while no linkage is inferred between those two images and the image of the baseball image, contained in P_2 itself? Perhaps the answer to these issues lies in using the matrix $(W + I_{|\mathcal{P}|})M$ as the adjacency matrix (where $I_{|\mathcal{P}|}$ is the $|\mathcal{P}| \times |\mathcal{P}|$ identity matrix).



(a) $(W+I_{|P|})$ M (c) $[(W+I_{|P|})M]^T(W+I_{|P|})M$

Figure 6: Two times the $k = 0.5$ adjacency matrix (and the corresponding co-citation matrix) for the example case

The matrix $(W + I_{|\mathcal{P}|})M$ defines each page to be adjacent both to the images which are contained in it, and to the images which are contained in pages to which it points. The co-citation measure which this adjacency matrix implies (the corresponding co-citation matrix is $[(W + I_{|\mathcal{P}|})M]^T (W + I_{|\mathcal{P}|})M$), now addresses the concerns raised in the previous case, as shown in Figure 6.

A closer look at our three proposed adjacency matrices (M, WM and $(W + I_{|\mathcal{P}|})M$) reveals that they are all members of the following parametric family of adjacency matrices (up to, perhaps, a constant factor):

$$A_{\mathcal{IC}} = \{[kW + (1-k)I_{|\mathcal{P}|}]M \quad : \quad 0 \le k \le 1\}$$

By denoting $A_{\mathcal{IC}}(k) \stackrel{\triangle}{=} [kW + (1-k)I_{|\mathcal{P}|}]M$, we have $M = A_{\mathcal{IC}}(0)$, $WM = A_{\mathcal{IC}}(1)$, and $(W + I_{|\mathcal{P}|})M = 2 \cdot A_{\mathcal{IC}}(\frac{1}{2})$. In general, choosing large values of k will introduce bias towards relations between pages and images contained in pages linked from them, while small values of k will boost the relationship between pages and the images which they themselves contain.

Our experiments and sample results were derived using the three adjacency matrices defined above, although we do not claim that any of these choices is in any way optimal.

3.1.1 Weighted Relations

The definitions of the previous subsection are easily extended to the case where both $\mathcal{L}$ and $\mathcal{E}$ are weighted relations, that is $\mathcal{L} \subseteq \mathcal{P} \times \mathcal{P} \times \mathrm{R}^+$ [3] is the set of weighted page to page links and $\mathcal{E} \subseteq \mathcal{P} \times \mathcal{I} \times \mathrm{R}^+$ is the weighted page-image relation. Weighted relations are derived by assigning weights to the links (relations) which reflect the amount of authority that the pointing (containing) page confers to the pointed page (image). Possible factors which may contribute to the weight of a link or relation include the following (the first factor is considered in PicASHOW):

- Anchor text which is relevant to the query. Such text around a link raises our confidence that the pointed page or contained image is relevant to the topic at hand ([5]). Similarly, in the case of the page-image relation, when the content of the `ALT` field is relevant to the query, then the image is most likely related to the topic of interest.

- The position of the link[image] in the pointing [containing] page. Many search engines consider the text at the top of a page as more reflective of its contents than text further down the page. The same line of thought can be applied to the links/images which appear in a page, with those which are closer to the top of the page receiving more weight than those appearing at the bottom of the page.

3.2 Assembling a Topical Collection

Our assumption in applying link-structure analysis when searching for quality images on topic t is that t-relevant pages will contain quality images on t. Thus, by examining a large enough set of t-relevant pages, we should be able to identify high quality t-images. Therefore, the first step in assembling a topical collection of images is to assemble a large collection of t-relevant pages. This collection is assembled in the same manner as described in Section 2.2. That is, for a query q which describes the topic t, we assemble a q-induced collection of Web pages by submitting q first to traditional search engines, and adding pages that point to or are pointed by pages in the resultant set. This provides us with the page set $\mathcal{P}$ and the page-to-page link set $\mathcal{L}$. Note that we do not utilize any image search engine in this step – we are using only standard (HTML-page finding) search engines. Once we compile the page set $\mathcal{P}$, we define the set $\mathcal{I}$ as the set of images which are contained in $\mathcal{P}$ (this also implies the page-to-image relation $\mathcal{E}$).

This scheme for assembling the topical image collection can be implemented with reasonable overhead, in terms of both computations and storage, by many standard WWW search engines. All search engines continuously crawl the Web with robots, which collect the textual content of the pages to be indexed. Many engines (such as AltaVista[7], Google[16] and Lycos[18]) additionally collect connectivity information that captures the information regarding the links between the pages as they crawl the Web. Our scheme requires the engines to also catalog, for each page, which images are contained (or pointed to) in the page. However, the images themselves need not be stored. As explained below, each image requires only the storage of a 32 byte signature, plus a URL where it can be found.

When building the image collection $\mathcal{IC} = (\mathcal{P}, \mathcal{I}, \mathcal{L}, \mathcal{E})$ we must consider a common authoring technique of Web pages. Specifically, when a Web site creator encounters an image of his liking on a remote server, the usual course of action would be to copy the image file to the local server, thus replicating the image. The motivation behind this practice is to enable the author's page to load faster from within the author's domain/organization, as the displayed images are stored locally.

This behavior of authors with respect to images is different from the corresponding behavior with respect to HTML pages. In most cases, authors will not copy a remote page (or some portion of its contents) to the local servers, but rather provide links from their site to the remote page. There are exceptions to this rule (such as the replication of system manuals or software APIs), but for most types of content, HTML pages are not replicated. This authoring mode has two important implications for link-based image search techniques, which are in contrast to the corresponding link based techniques for searching Web pages. We expand on these now.

Identifying replicated images

We must identify that multiple pages contain a certain image, even when the pages contain different copies of the image. Thus, images cannot be identified by their URIs, but rather must be identified by their content. In contrast, when applying link-analysis in search of authoritative pages, identifying replications is less crucial. Satisfactory results can be obtained even when the issue of page-replications is ignored.

Fortunately, it is possible to decide whether two images are identical with a relatively high probability by examining a small portion of the image. In PicASHOW, we only download the first 1024 bytes of the image and apply a double hash function to these bytes, so that each image is represented by a signature consisting of 32 bytes. Two images with the same signature are considered identical. Our experience shows that very rarely do different images result in the same signature since the first 1024 bytes usually capture the header information as well as the first part of the image itself. The storage overhead which is associated with each image is thus quite minimal. Note that replications of the same image result in only one 32-byte signature (and one URL) in terms of storage requirements.

Filtering non-informative images

Link-analysis based page-search methods usually interpret a link from page p to page q as a measure of authority which p confers on q ([23]). However, there are many kinds of links which confer little or no authority ([6]), and we refer to these as *non-informative* links. Some examples for such links are intra-domain (inner) links (whose purpose is to provide

[3] R^+ is the set of non-negative real numbers.

navigational aid in complex Web sites), commercial/sponsor links, and links which result from link-exchange agreements. A crucial task which should be completed prior to analyzing the link structure of a given collection, is to filter out as many non-informative links as possible.

Similarly, filtering non-informative page-to-image links is crucial for successful link-based image retrieval. Site banners and logos can be thought of as the image equivalents of non-informative links. These images introduce a large amount of noise into image collections, which we would like to be able to filter. When building the set of page-to-page links $\mathcal{L}$, we identify and filter out intra-domain links, ruling them to be navigational links which do not confer authority. However, the practice (described above) of image replication implies that filtering out the intra-domain page-to-image links of $\mathcal{E}$ will be destructive as we may also lose quality images in this fashion. We thus introduce a few heuristics, that can mitigate the noise that is introduced by non-informative images:

- Banners and logos tend to be wide and short. We can thus filter out images with an aspect ratio greater than some threshold. Note that we only need to examine the image header for this information.
- Images which are stored in small files (less than 10 kilobytes, for example) tend to be banners. Even if they are not banners, they are usually not quality topical images. Therefore, they can be filtered from the collection.
- Images that are stored in files whose names contain the words *logo* or *banner* are probably logos and banners...
- In addition to banners and logos, people tend to include other non-informative images such as clipart in order to liven up their Web page. These include colorful buttons, bars, mail boxes, spinning globes, etc. Some of these are highly popular, and are replicated and used in large numbers on the Web. We consider these the equivalent of stop words in information retrieval ([28]) and thus term them *stop images*. Many of these stop images are filtered based on the aspect ratio and file size heuristics. In addition, we have constructed a list of stop images (common names of some of these images, and 32-byte signatures of others), and we filter out any image that appears in this list. Currently, this list is assembled manually. It is, however, feasible to compile such a list based on distribution statistics of images on the WWW.

These heuristics do not filter out all the noise caused by non-informative images. Some non-informative images survive this process, and introduce noise into our page-to-image adjacency matrices. We found that this noise affects the image rankings more than usually happens in the corresponding page ranking schemes (where page-to-page adjacency matrices are used). As a consequence, link-based image search seems to be more noisy than link-based page search, and specifically may be easier to spam. Devising more elaborate and effective filtering schemes is left for future work.

3.3 Image Ranking Schemes

After assembling an image collection $\mathcal{IC} = (\mathcal{P}, \mathcal{I}, \mathcal{L}, \mathcal{E})$ pertaining to a certain topic t, we need to rank the images of $\mathcal{I}$ with respect to t. We assume that every page $p \in \mathcal{P}$ is associated with a t-relevance score, denoted by $r_t(p)$. Note that this is not a limiting assumption, since we can always calculate the authority scores of the pages and use them as relevance scores. In particular, the collection $\mathcal{IC}$ contains the linkage information which is required to calculate authority scores by the Mutual Reinforcement approach.

Below is a list of the ranking schemes with which we have experimented. They are divided into three categories: A naive image in-degree approach, PageRank-influenced ranking schemes, and co-citation based analyses. These ranking schemes are based on the matrices which were defined in Section 3.1.

1. In-degree rank according to the matrix M. Here, the score of image i equals $\sum_{\{p \in P | p \text{ contains } i\}} m_{p,i}$, where $m_{p,i}$ is the weight associated with the page-image relation $p \rightsquigarrow i$. That is, the score of image i is the sum of the weights of all relations $p \rightsquigarrow i$ for all pages p which contain image i in the collection.

2. PageRank ([2]) influenced ranking schemes. Under the hypothesis that images which are contained in t-relevant pages should be of higher quality (with respect to t) than images contained in t-irrelevant pages, we factor the relevance score of page p ($r_t(p)$) into the score of image i. In particular, we set the score of image i to equal $\sum_{\{p \in P | p \text{ contains } i\}} r_t(p)$. In the case that M is a weighted matrix (and not simply a binary matrix), the straightforward variant of this score is $\sum_{\{p \in P | p \text{ contains } i\}} r_t(p) m_{p,i}$.

3. Co-citation based analyses, with each of the matrices M, WM and $(W + I_{|\mathcal{P}|})M$ serving as the citation (adjacency) matrix, and with the co-citation matrices $M^T M$, $(WM)^T (WM)$ and $[(W + I_{|\mathcal{P}|})M]^T [(W + I_{|\mathcal{P}|})M]$ used for the purpose of the analysis. Specifically, we tested the rankings that are produced by the Mutual Reinforcement approach and by SALSA [4].

   We can also use the t-relevance score $\{r_t(p), p \in \mathcal{P}\}$ to enhance our co-citation analysis and boost the rankings of images that are cited by highly t-relevant pages. As an example, consider applying co-citation analysis to the $|\mathcal{P}| \times |\mathcal{P}|$ page-to-image matrix M_R that is defined as follows:

   $$[M_R]_{i,j} \triangleq m_{i,j} \sqrt{r_t(i)}$$

   By examining the *t-relevance image co-citation* matrix $M_R^T M_R$, we note that when M is unweighted (a binary adjacency matrix), $[M_R^T M_R]_{i,j}$ sums the relevance weight of all pages which co-display images i and j:

   $$[M_R^T M_R]_{i,j} = \sum_{\{k: k \rightsquigarrow i, k \rightsquigarrow j\}} [M_R]_{k,i} [M_R]_{k,j} =$$

   $$= \sum_{\{k: k \rightsquigarrow i, k \rightsquigarrow j\}} (\sqrt{r_t(k)})^2 = \sum_{\{k: k \rightsquigarrow i, k \rightsquigarrow j\}} r_t(k)$$

[4]An in-depth description of SALSA ([24]) is out of this paper's scope. However, SALSA is based on co-citation, and its analysis, just like Kleinberg's Mutual Reinforcement approach, is completely governed by the adjacency matrix which is used.

Thus, not all image co-citations are considered equal - highly relevant pages endorse their co-contained images to a larger extent than do less relevant pages.

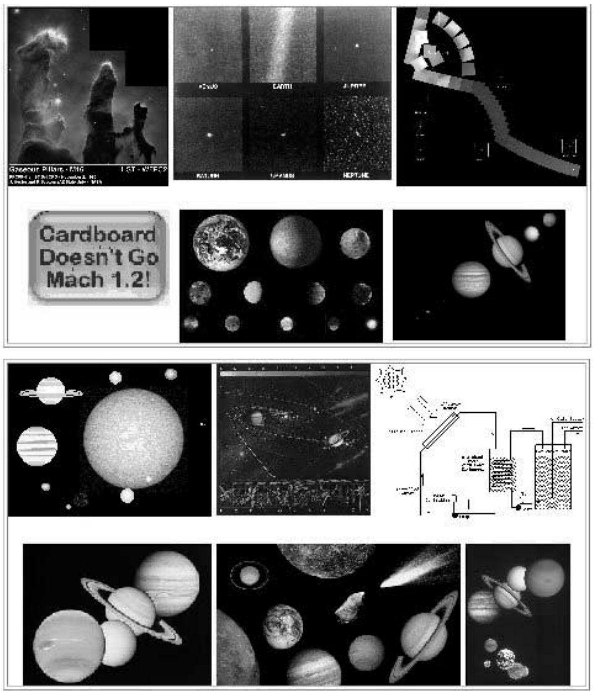


Figure 7: PicASHOW/Ditto images for the query *Solar System*

4. IMAGE CONTAINERS / IMAGE HUBS

One of the major benefits of hyperlink-based image search is that in addition to finding good images which pertain to a certain query, it can also identify Web pages that are rich in relevant images, or from which many images are readily accessible. Our proposed co-citation based ranking schemes naturally allow for such Web pages to be found.

While we have concentrated, in the previous section, on how to rank the authoritative images in $\mathcal{IC}$, we can similarly find Web pages whose role corresponds to that of hubs. Just as hubs were defined as pages which link to many authoritative pages, in our context *image hubs* should be pages which are, in some sense, linked to many authoritative images.

However, the notion of an *image hub* is somewhat ambiguous. Do we, by calling p an "image hub", mean that many authoritative images are displayed in p, or do we mean that p points to many pages which contain quality images? We claim that both possible interpretations are of value, and so we define them separately as follows: Pages which contain high-quality images are called *image containers*, while pages which point to good image containers are called *image hubs*. Thus, image hubs are once removed from the authoritative images themselves, which are contained (as the name implies) in the image containers.

Our co-citation based image retrieval schemes can find both image containers and image hubs, either separately or in some mixed manner. The outcome depends on the type of adjacency matrix used to describe the collection $\mathcal{IC}$,

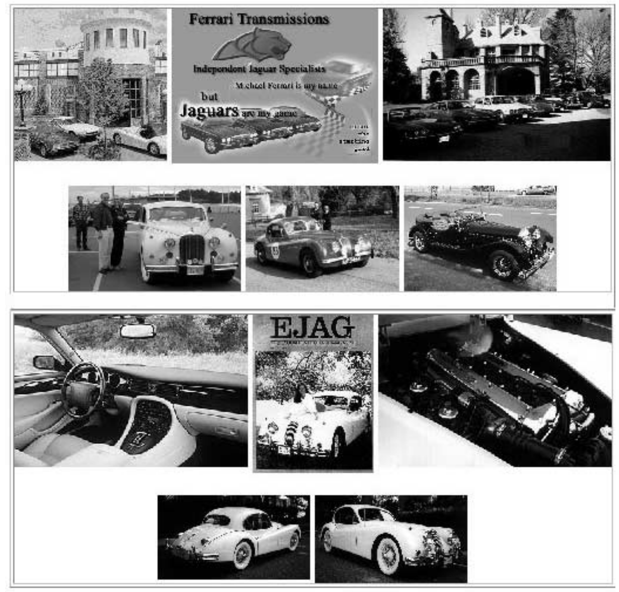


Figure 8: PicASHOW/Lycos images for the query *Jaguar car*

which, in turn, implies the bibliographic coupling matrix which governs the ranks of pages as image hubs/containers (the technical details of how hub ranks are derived from the adjacency matrix were given in section 2.2).

When using the adjacency matrix M (or M_R), the p'th row describes which images are contained in page p. The pages whose coordinates will stand out in the principal eigenvector of the matrix MM^T will, accordingly, form a community of image containers. Image hubs are likely to be found when using the adjacency matrix WM, whose p'th row describes which images are contained in pages to which p links (the corresponding bibliographic coupling matrix will be WMM^TW^T). Co-citation analysis using the matrix $(W + I_{|\mathcal{P}|})M$ allows us to find communities of pages which both contain images and which point to other image containers, since the p'th row there details the images which are contained either in p or in pages to which p links. In general, when using the adjacency matrix $A_{\mathcal{IC}}(k)$, high values of k shift the analysis towards image hubs, while lower values of k accentuate image containers.

One of the most striking image containers found throughout our experiments with PicASHOW is the fractal image container `http://sprott.physics.wisc.edu/fractals.htm`. The page contains over a hundred images of fractals. Figure 9 shows a few of those striking images.

Figure 9: Images of fractals contained in http://sprott.physics.wisc.edu/fractals.htm

Another example is the image container for the query *Magritte*, `http://www.xs4all.nl/~renebos/magritte.html`. The page itself contains just 5 words: Rene Magritte : 10 Genius Paintings. Underneath that phrase are displayed 10 of Magritte's masterpieces, five of which are shown in Figure 10.

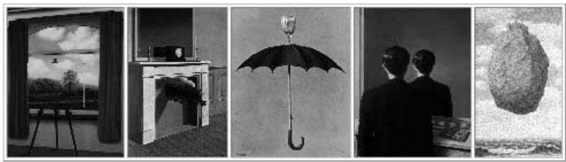


Figure 10: Magritte paintings contained in http://www.xs4all.nl/ renebos/magritte.html

Here are a couple of examples for image hubs:

1. For the query *pyramids*, many relevant images are just a click away from the following URLs: (a) The Ancient Egypt page, pyramids section at `http://members.aol.com/TeacherNet/AncientEgypt.html` and (b) The Art History Resources: Part 2 Ancient Art page, Egypt section at `http://witcombe.sbc.edu/ARTHancient.html#AncEgypt`

2. For the query *Yosemite*, "Photos by Rick Ellis – Yosemite" (`http://www.fnet.net/~ellis/photo/yosemite.html`) is both an image container and an image hub. Actually, it is a Yosemite hub in general, since it points to many valuable resources on Yosemite National Park. Many Yosemite images are accessible from `http://home.earthlink.net/~mgordon324/sierra_new.htm`, Michael Gordon's High Sierra page, which is an image container and hub for the Sierra Nevada mountain range (which includes the Yosemite National Park).

5. DISCUSSION AND FUTURE WORK

The examples showcased throughout the paper show that PicASHOW, while relying on very little besides link analysis, demonstrates retrieval abilities comparable to those of available WWW image search engines. In addition, PicASHOW's retrieval of image containers and hubs is a natural and useful extension of the image search paradigm, which, to the best of our knowledge, has not been pursued before.

We have shown results for queries from diverse fields such as science, art, geographical landmarks, transportation and celebrities. The following list gives more details on the ranking scheme that was applied in each example, and highlights other noteworthy points which follow from the example. The reader should recall the definition of the ranking schemes from section 3.3.

1. Figure 1 brings PicASHOW's results for the query *"Michael Jordan"*. The image collection consisted of 1083 images, and the results shown were derived by applying the PageRank-influenced ranking scheme. The URLs of these images (see Table 4) exemplify the practice of image replication among Web authors. Note that two of those images are animated GIFs.

2. Figure 2 brings PicASHOW's results for the query *"Vincent Van Gogh"*. The image collection consisted of 582 images, and the results shown were derived by applying Kleinberg's algorithm using the adjacency matrix $(W + I_{|\mathcal{P}|})M$.

3. Figure 7 brings our results for the query *"Solar System"*. The image collection consisted of 682 images, and the images were ranked by weighted in-degree according to the relation $\mathcal{E}$.

4. Figure 8 has our results for the query *"Jaguar car"*. The image collection in this case was very small, and consisted of just 67 images. The ranking scheme was Kleinberg's algorithm, using the adjacency matrix M. Note that while PicASHOWs results all come from different servers, most of the Lycos images are from a single server (see Table 1).

5. Figure 11 displays PicASHOW's results for the query *"Kilimanjaro"*. The 309 images of this example were ranked by SALSA, using the adjacency matrix M. Note how some of the image names do not contain anything resembling the query, but rather the name *"Kibo"*, which is the name of one of Mt. Kilimanjaro's peaks (see Table 2).

Figure 11: PicASHOW images for the query *Kilimanjaro*

Our ranking techniques require that the search topic be one of wide interest on the Web, where image replication is likely to occur. For example, people are likely to display replications of publicly released images of celebrities. Natural landmarks may also induce image replications. However, a query such as *Paris* is less likely to achieve quality results, since people will often display images of their own vacation in Paris, rather than replicated versions of the city's landmarks. Also, as in the case of link analysis in Web page search, queries on obscure topics of little interest will most likely fail to produce quality results.

Since PicASHOW performs no image analysis whatsoever, it cannot handle queries that contain image qualifiers such as color, orientation, and other specific features. For example, PicASHOW can retrieve images of Michael Jordan, but not of Michael Jordan *wearing a suit*. It can find images of Jaguar cars, but not of *red* Jaguar cars, and while it will rank nicely images of Mount Kilimanjaro, one cannot ask for those images *not to contain trekkers*, or to be taken *from below*. Note however, that link-analysis based techniques could still be of value for such queries. For example, PicASHOW could be used as an initial filter to find candidate images on the topic of interest (e.g., Jaguar cars). Some form of image analysis could then be performed on these candidate images in order to select those that further satisfy the image qualifications (e.g. which of the resulting Jaguar cars are red).

Several interesting extensions of the image search schemes are feasible:

- Kleinberg demonstrated in [23] that non-principal communities of authoritative pages can distinguish between topics in multi-topic collections, and between pages which prescribe to opposing views on polarized topics (such as the *pro-life* and *pro-choice* views on *abortion*). It thus seems interesting to investigate the non-principal communities of images which arise from the various proposed co-citation measures.

- Most CBIR systems support image similarity queries, and extending our approach to support such queries would enhance its appeal. Many algorithms have been proposed for finding pages related to a sample page (or set of pages) on the Web by link analysis ([8],[1]). The main idea is to grow a Web-graph around the given seed pages, and then find the dominant authorities in that graph. It seems possible to adapt PicASHOW in the same fashion to support similar-image queries. The input to such queries can be either URLs of sample images, or URLs of sample image containers.

- In our current prototype, we define the images in an image collection $\mathcal{IC} = (\mathcal{P}, \mathcal{I}, \mathcal{L}, \mathcal{E})$ to be the images contained in the set of pages $\mathcal{P}$. Recall that $\mathcal{P}$ was actually a neighborhood of pages around a set of root pages S on some topic t. Consider the set of root images $\mathcal{I}_S$, defined as the set of images which are contained in the pages of S. We currently expand $\mathcal{I}_S$ into the final set of images $\mathcal{I}$ by following page-to-page links, in other words by expansion on the *page plane*. However, $\mathcal{I}_S$ may additionally be expanded by adding images that are contained in other pages containing replications of the root images. The premise is that if pages, which are not linked to the root set of pages S, contain replications of images from $\mathcal{I}_S$, they may also contain other images of relevance to t. We term as *image plane expansion* the process of adding those pages, and their contained images, to $\mathcal{IC}$. Technically, such expansion requires information of the form "which pages contain the following (signature of an) image". This information is not currently available on the Web, but search engines which will collect page-to-image connectivity information can support such queries, and thus enable image plane expansion as well.

Acknowledgments

We would like to thank Shlomo Moran and Yoelle Maarek for useful discussions on the ideas presented in this paper.

Vitae

Ronny Lempel is a Ph.D. Student in the Faculty of Computer Science, Technion, Haifa, Israel, focusing on WWW link-structure analysis. He received his B.Sc. and M.Sc. from the same faculty in 1997 and 1999, respectively.

Aya Soffer is a Research Staff Member in the Information Retrieval and Organization group at the IBM Research Lab in Haifa. Her research interests include pictorial information systems, information retrieval, and non-traditional database systems. She received her MS and PhD degrees in Computer science from the University of Maryland in 1992 and 1995, respectively.

6. REFERENCES

[1] Y. Aridor, D. Carmel, R. Lempel, Y. S. Maarek, and A. Soffer. Knowledge agents on the web. *Proc. 4th International Workshop on Cooperative Information Agents*, 2000.

[2] S. Brin and L. Page. The anatomy of a large-scale hypertextual web search engine. *Proc. 7th International WWW Conference*, 1998.

[3] M. La Cascia, S. Sethi, and S. Sclaroff. Combining textual and visual cues for content-based image retrieval on the world wide web. *Proc. IEEE Workshop on Content-based Access of Image and Video Libraries*, 1998.

[4] IBM Almaden Research Center. CLEVER. http://www.almaden.ibm.com/cs/k53.clever.html.

[5] S. Chakrabarti, B. Dom, D. Gibson, J. Kleinberg, P. Raghavan, and S. Rajagopalan. Automatic resource list compilation by analyzing hyperlink structure and associated text. *Proc. 7th International WWW Conference*, 1998.

[6] S. Chakrabarti, B. Dom, D. Gibson, J. Kleinberg, S.R. Kumar, P. Raghavan, S. Rajagopalan, and A. Tomkins. Hypersearching the web. *Scientific American*, June 1999.

[7] AltaVista Company. http://www.altavista.com/.

[8] J. Dean and M. R. Henzinger. Finding related pages in the world wide web. *Proc. 8'th International World Wide Web Conference*, 1999.

[9] ditto.com visual search engine. http://www.ditto.com/.

[10] Columbia University Center for Telecommunications Research. webSEEk. http://www.ctr.columbia.edu/webseek/.

[11] C. Frankel, M. Swain, and V. Athitsos. WebSeer: An image search engine for the world wide web. Technical Report TR-96-14, Computer Science Department, University of Chicago, 1996.

[12] T. Gevers and A. W. M. Smeulders. The PicToSeek WWW image search system. *Proc. IEEE International Conference on Multimedia Computing and Systems*, 1:264–269, 1999.

[13] V. Harmandas, M. Sanderson, and M.D. Dunlop. Image retrieval by hypertext links. *Proc. 20'th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 296–303, 1997.

[14] Corbis Inc. Corbis images. http://www.corbis.com/.

[15] Getty Images Inc. http://www.getty-images.com/.

[16] Google Inc. Google search engine. http://www.google.com/.

[17] Lycos Inc. Lycos internet guide. http://multimedia.lycos.com/.

[18] Lycos Inc. Lycos internet guide. http://www.lycos.com/.

[19] Scour Inc. http://www.scour.com/.

[20] Yahoo! Inc. http://www.yahoo.com/.

[21] K. Kanth, D. Agrawal, and A. Singh. Dimensionality reduction for similarity searching in dynamic databases. In *Proc. ACM SIGMOD Int. Conf. on Management of Data*, pages 166–176, 1998.

[22] M.M. Kessler. Bibliographic coupling between

scientific papers. *American Documentation*, 14:10–25, 1963.

[23] J. Kleinberg. Authoritative sources in a hyperlinked environment. *J. ACM*, 46:5:604–632, 1999.

[24] R. Lempel and S. Moran. The stochastic approach for link-structure analysis (SALSA) and the TKC effect. *Proc. 9th International WWW Conference*, 2000.

[25] D. Rafiei and A.O. Mendelzon. What is this page known for? computing web page reputations. *Proc. 9th International WWW Conference*, 2000.

[26] H. Small. Co-citation in the scientific literature: A new measure of the relationship between two documents. *J. American Soc. Info. Sci.*, 24:265–269, 1973.

[27] J. R. Smith and S.-F. Chang. Searching for images and videos on the world-wide web,. Technical Report 459-96-25, Columbia University/CTR, 1996.

[28] C.J. van Rijsbergen. *Information Retrieval*. Butterworths, 1979.

APPENDIX

A. URLS OF IMAGES

We bring here the URLs of the showcased images. For each result set, the URLs are given in clockwise order, starting from the upper left corner. For PicASHOW's results, multiple URLs are given for replications of the same image, when applicable.

URLs of PicASHOW's *Jaguar Car* images
http://www.classicar.com/museums/welshjag/outside.gif
http://www.ferrari-transmissions.co.uk/home2.jpg [5]
http://www.jtc-nj.com/Doylestowncrowd.jpg
http://www.jaguar-association.de/images/verkaufsbilder/12-00-tesch/ss100s-lg.jpg
http://www.j-c-c.org.uk/images/drive.jpg
http://www.seattlejagclub.org/IMAGES/picyzk.jpg
URLs of the Lycos *Jaguar car* images
http://www.auto.com/art/reviews/98_jaguar_xjr/98_Jaguar_XJR_Interior.jpg
http://highway-one.com/Images/Photos/Jaguar/LaGrassaJaguar4.jpg
http://highway-one.com/Images/Photos/Jaguar/LaGrassaJaguar2.jpg
http://highway-one.com/Images/Photos/Jaguar/LaGrassaJaguar.jpg
http://highway-one.com/Images/Photos/Jaguar/LaGrassaJaguar3.jpg from

Table 1: URLs of *Jaguar car* images

URLs of PicASHOW's *Kilimanjaro* images
http://www.calle.com/ carl/brett.kili.jpg http://www.premier.org.uk/graphics/programmes/kili001.jpg
http://www.sfusd.edu/cj/kibo.jpg http://nisus.sfusd.k12.ca.us/cj/kibo.jpg
http://www.geocities.com/Yosemite/1015/kili1.jpg
http://seclab.cs.ucdavis.edu/ wee/images/kili-summit.gif
http://www.geocities.com/Yosemite/1015/kili2.jpg
http://www.picton-castle.com/jpg/Kilimanjaro_masai_T.jpg
http://www.adventure.co.za/1STPAGEOFKIBO.jpg

Table 2: URLs of *Kilimanjaro* images

[5]When enlarged, this image reads "Michael Ferrari is my name, but Jaguars are my game". Mr. Ferrari claims to be an independent Jaguar transmission specialist...

[6]Although two of the URLs have a suffix of *.gif*, all three contain the same *jpeg* image. The suffix, in these cases, does not describe the file type correctly.

URLs of PicASHOW's *Van Gogh* images
http://www.vangoghgallery.com/images/small/0612.jpg http://www-scf.usc.edu/ wrivera/vangogh.jpg http://www.openface.ca/ vangogh/images/small/0612.jpg
http://www.vangoghgallery.com/images/small/0627.jpg http://www.sd104.s-cook.k12.il.us/rhauser/vangoghsel.jpg http://www.openface.ca/ vangogh/images/small/0627.jpg
http://www.vangoghgallery.com/images/intro/1530.jpg http://www.openface.ca/ vangogh/images/intro/1530.jpg
http://www.bc.edu/bc_org/avp/cas/fnart/art/19th/vangogh/vangoghself3.jpg
http://sunsite.unc.edu/wm/paint/auth/gogh/entrance.jpg http://www.ibiblio.org/wm/paint/auth/gogh/entrance.jpg http://www.southern.com/wm/paint/auth/gogh/entrance.jpg
http://www.bc.edu/bc_org/avp/cas/fnart/art/19th/vangogh/vangogh_starry1.jpg
URLs of AltaVista's *Vincent Van Gogh* images
http://www.ElectronicPostcards.com/pc/pics/van12b.jpg
http://www.ElectronicPostcards.com/pc/pics/van5b.jpg
http://www.ElectronicPostcards.com/pc/pics/van1b.jpg
http://www.culturekiosque.com/images5/van.jpg
http://www.ElectronicPostcards.com/pc/pics/van6b.jpg
http://www.ElectronicPostcards.com/pc/pics/van2b.jpg

Table 3: URLs of *Vincent Van Gogh* images

URLs of PicASHOW's *Michael Jordan* images
http://views.vcu.edu/ absiddiq/jordan01.gif http://www.geocities.com/Colosseum/Sideline/1534/jordan01.gif http://scnc.sps.k12.mi.us/ powers/jordan01.gif http://www.eng.fsu.edu/ toliver/jordanmovie.gif http://www2.gvsu.edu/%7Ejirtlee/Bettermovingdunk.gif http://sesd.sk.ca/scp/images/AIR_JORDAN.gif hspace0.3cm(animated gif)
http://www.geocities.com/Colosseum/Sideline/1534/jumper.jpg http://scnc.sps.k12.mi.us/ powers/jumper.jpg http://icdweb.cc.purdue.edu/ fultona/MJ11.jpg
http://www.engin.umd.umich.edu/ jafreema/mj/mjpics/jordan5-e.gif
http://www.angelfire.com/ny/Aaronskickasspage/images/1-11.JPG http://homepages.eu.rmit.edu.au/dskiba/mjsmile.jpg http://scnc.sps.k12.mi.us/ woolwor1/jordan.jpg http://www.fidelweb.com/graphic/jordan4.jpg
http://www.engin.umd.umich.edu/ jafreema/mj/mjpics/jordan10-e.jpg
http://www.engin.umd.umich.edu/ jafreema/pictures/1991.gif http://metal.chungnam.ac.kr/ myoungho/1991.gif hspace0.3cm(animated gif)
URLs of Scour's *Michael Jordan* images
http://www.geocities.com/SunsetStrip/2546/mjjuwan.jpg
http://www.geocities.com/SunsetStrip/2546/jordanmvp2.jpg
http://www.unc.edu/ lbrooks2/jordan2.jpg
http://www.geocities.com/Colosseum/Track/7823/JORDAN_ALLSTAR_1.JPG
http://www.big.du.se/ joke/f1-96/pics/car/jordan96_car.jpg
http://www.unc.edu/ lbrooks2/mjbugs.jpg

Table 4: URLs of *Michael Jordan* images

URLs of PicASHOW's *Solar System* images
http://oposite.stsci.edu/pubinfo/jpeg/M16Full.jpg http://www.geosci.unc.edu/classes/Geo120/SNsmall.gif http://www.geosci.unc.edu/classes/Geo15/SNsmall.gif [6]
http://nssdc.gsfc.nasa.gov/image/planetary/solar_system/family_portraits.jpg
http://www.seds.org/nineplanets/nineplanets/gif/SmallWorlds.gif http://www4.netease.com/ chwu/images/solar_system/nineplanets/SmallWorlds.gif http://www.physics.louisville.edu/tnp/gif/SmallWorlds.gif http://img.iln.net/images/main/astronomy/gif/SmallWorlds.gif http://www-hpcc.astro.washington.edu/mirrors/nineplanets/gif/SmallWorlds.gif http://seds.lpl.arizona.edu/nineplanets/nineplanets/gif/SmallWorlds.gif
http://www.seds.org/nineplanets/nineplanets/NinePlanets.jpg http://kiss.uni-lj.si/ k4fg0152/devetplanetov/xslike/9planetov_x.jpg http://www.physics.louisville.edu/tnp/NinePlanets.jpg http://img.iln.net/images/main/astronomy/NinePlanets.jpg http://www-hpcc.astro.washington.edu/mirrors/nineplanets/NinePlanets.jpg http://seds.lpl.arizona.edu/nineplanets/nineplanets/NinePlanets.jpg
http://www.solarviews.com/images/rocketvision.gif (animated gif)
http://nssdc.gsfc.nasa.gov/image/planetary/solar_system/solar_family.jpg
URLs of Ditto's *Solar System* images
http://www.festivale.webcentral.com.au/shopping/art_com/SYST.jpg
http://www.coseti.org/images/12358.jpg
http://www.greenbuilder.com/sourcebook/SourcebookGifs/HeatCoolSolar.2.GIF
http://www.astro.ufl.edu/aac/icons/solsyt.gif
http://connect.ccsn.edu/edu/shs/grant/solar_system.gif
http://www.bonus.com/bonus/card/solarsystembrowser/solarsystembrowser.jpg

Table 5: URLs of *Solar System* images