IntelliSearch: Intelligent Search for Images and Text on the Web

Epimenides Voutsakis[1], Euripides G.M. Petrakis[1], and Evangelos Milios[2]

[1] Dept. of Electronic and Computer Engineering
Technical University of Crete (TUC)
Chania, Crete, GR-73100, Greece
pimenas@softnet.tuc.gr, petrakis@intelligence.tuc.gr
[2] Faculty of Computer Science, Dalhousie University
Halifax, Nova Scotia
B3H 1W5, Canada
eem@cs.dal.ca

Abstract. *IntelliSearch* is a complete and fully automated information retrieval system for the Web. It supports fast and accurate responses to queries addressing text and images in Web pages by incorporating state-of-the-art text and Web link information indexing and rertieval methods in conjunction with efficient ranking of Web pages and images by importance (authority). Searching by semantic similarity for discovering information related to user's requests (but not explicitly specified in the queries) is a distinguishing feature of the system. *IntelliSearch* stores a crawl of the Web with more than 1,5 million Web pages with images and is accessible on the Internet[3]. It offers an ideal test-bed for experimentation and training and serves as a framework for a realistic evaluation of many Web image retrieval methods.

1 Introduction

Searching for effective methods to retrieve information from the Web 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 [2]. Image retrieval on the Web, in particular, is a very important problem in itself [3]. The relevant technology has also evolved significantly propelled by advances in image database research [4].

Several approaches to the problem of content-based image retrieval on the Web have been proposed and some have been implemented on research proto-types (e.g., ImageRover [5],WebSEEK [6], Diogenis [7]) and commercial systems. The last category of systems, includes general purpose image search engines (e.g.,

---

[3] http://www.intelligence.tuc.gr/intellisearch

Google Image Search [4], Yahoo [5], Altavista [6] Ditto [7]) as well as systems providing specific services to users such as detection of unauthorized use of images, Web and e-mail content filters (e.g., Cobion [8]), image authentication, licensing and advertising (e.g., Corbis [9]).

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 [8].

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 typical 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., [5, 6, 9]).

Hybrid retrieval systems combining the above approaches such as systems using image analysis features in conjunction with text and attributes (e.g., [7, 10, 11]).

The problem of how to select authority Web pages and images on the topic of the query has not been addressed by any of the above methods, which focus mainly on image and text content. In Web text retrieval, link analysis methods such as HITS [12] and PageRank [13] have been applied to estimate the quality of Web pages and the topic relevance between the Web pages and the query.

---

[4] `http://www.google.com/imghp`
[5] `http://images.search.yahoo.com`
[6] `http://www.altavista.com/image`
[7] `http://www.ditto.com`
[8] `http://www.cobion.com`
[9] `http://pro.corbis.com`

Incorporating page content within link analysis has also been proposed [14]. Extending these ideas to image retrieval on the Web is the natural next step. Building upon HITS, PicASHOW [15] shows how to handle pages that link to images and pages that contain images. WPicASHOW [11] shows how to handle image content in conjunction with link information.

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. However, 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. Typically, images are retrieved by addressing text associated with them (e.g., captions) in Web pages [8]. This is the stat-of-the-art approach for achieving consistency of representation and high accuracy results.

IntelliSearch is motivated by these ideas. The link analysis and text retrieval methods referred to above are implemented and integrated into *IntelliSearch*. The resulting system provides an ideal test-bed for experimentation and training and also a framework for a realistic evaluation of state-of-the-art Web image retrieval methods. An analysis of the performance of all these methods is presented in [11, 16]. The main points of this analysis are also discussed in this work. Furthermore, *IntelliSearch* supports fast and accurate responses to queries addressing Web pages or images by incorporating efficient indexing of text information extracted from Web pages. The system stores a crawl of the Web with 1,5 million Web pages with images.

2   Information Retrieval in *IntelliSearch*

IntelliSearch 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. Typically, images are described by text surrounding them in the Web pages (e.g., caption, title) [8]. The following image descriptors are derived from Web pages 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.

The similarity between a query Q and an image I is computed as a summaiton of similarities between the query and the above image descriptors:

$$S_{image}(Q, I) = S_{file\_name}(Q, I) + S_{alternate\_text}(Q, I) + \qquad (1)$$
$$S_{page\_title}(Q, I) + S_{image\_caption}(Q, I).$$

For queries addressing the text content of Web pages, the similarity between a query Q and a Web page W is computed as the sum of the similarities of the query with the text descriptions obtained from the the entire Web page and its title (if exists):

$$S_{text}(Q, W) = S_{page\_title}(Q, W) + S_{page\_text}(Q, W). \qquad (2)$$

IntelliSearch implements the following two methods for computing similarity S.

2.1 Vector Space Model (VSM) [17]

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. The weight of a term is computed as a function of its frequency of occurrence in the document collection and can be defined in many different ways. The term frequency - inverse document frequency model [17] is used for computing the 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) [17] as the cosine of the inner product between their vector representations

$$S(Q, D) = \frac{\sum_i q_i d_i}{\sqrt{\sum_i q_i^2} \sqrt{\sum_i d_i^2}}, \qquad (3)$$

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

2.2 Semantic Similarity Retrieval Model (SSRM) [16]

The lack of common terms in two documents does not necessarily mean that the documents are unrelated. Similarly, relevant text may not contain the same terms. Semantically similar concepts may be expressed in different words in the documents and the queries, and direct comparison by word-based VSM is not effective. For example, VSM will not recognize synonyms or semantically similar terms (e.g., "car", "automobile").

SSRM works by discovering semantically similar terms using WordNet [10] to estimate the similarity between different terms. The similarity between an expanded and re-weighted query q and a text d is computed as

$$S(Q, D) = \frac{\sum_i \sum_j q_i d_j \text{sim}(i, j)}{\sum_i \sum_j q_i d_j}, \tag{4}$$

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 [18]. 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).

The method, although slow (due to its quadratic time complexity and extensive searches for terms over WordNet and computing their semantic similarity) has been demonstrated to outperform VSM for text and image queries [16].

2.3 HITS [12]

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 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. 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

---

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

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.

2.4 PicASHOW [15]

Building upon HITS, PicASHOW shows how to handle pages that link to images and to pages that contain images. PicASHOW [15] 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 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.
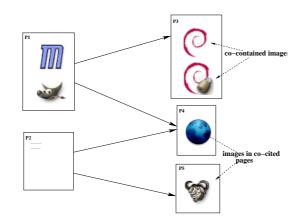

Fig. 1. The focused graph corresponding to query "Debian logo".

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.

Similarly to HITS, $PicASHOW$ defined the so called image *co-citation* and *bibliographic* matrices $[(\mathcal{W}+\mathcal{I})\mathcal{M}]^{\mathcal{T}} \cdot (\mathcal{M}+\mathcal{I})\mathcal{W}$ and $(\mathcal{W}+\mathcal{I})\mathcal{M} \cdot [(\mathcal{W}+\mathcal{I})\mathcal{M}]^{\mathcal{T}}$ respectively. The ij-th entry of the image co-citation matric is the number of pages that jointly point to images with indices i and j. The ij-th entry of the image bibliographic matric is the number of images jointly referred to by pages i and j. Fig. 2 illustrates the adjacency and bibliographic matrices for the the focused graph of Fig. 1.

	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

	img1	img2	img3	img4	img5	img6
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

	img1	img2	img3	img4	img5	img6
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. 2. $\mathcal{W}$, $\mathcal{M}$ and $(\mathcal{W}+\mathcal{I})\mathcal{M}$ matrices computed by PicASHOW for the focused graph of Fig. 1.

Image Authority and Hub 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 $(\mathcal{W}+\mathcal{I})\mathcal{M} \cdot [(\mathcal{W}+\mathcal{I})\mathcal{M}]^{\mathcal{T}}$ 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 [14]:

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.

2.5 Weighted PicASHOW (WPicASHOW) [11]

PicASHOW does not show how to handle image content or image text context. This problem is addressed by WPicASHOW (or Weighted PicASHOW) [11], a weighted scheme for co-citation analysis is proposed. 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 [11].

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. 2 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 (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. 2.
- 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 ($t = 100$ in this work) 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. 3 illustrates these matrices for the example set R of Fig. 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).

Fig. 4 illustrates authority and hub values computed by WPicASHOW in response to query "Debian logo". The answers to the query are ranked by authority values. Notice the high authority scores of pages showing logo or trademark images of "Debian Linux".

2.6   Weighted HITS [14]

Similarly to WPicASHOW, WHITS (weighted HITS) a weighting link analysis scheme for retrieval of Web pages is also implemented. HITS uses link information between pages (does not consider links to images or to pages containing images). Links are weighted by their text similarity (as computed by VSM).

	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. 3. $\mathcal{W}$, $\mathcal{M}$ and $(\mathcal{W} + \mathcal{I})\mathcal{M}$ matrices computed by WPicASHOW for the focused graph of Fig. 1.

Image						
Authority Values	.751	.657	.0418	.0418	.008	0

Page	P_1	P_2	P_3	P_4	P_5
Hub Values	.519	.0001	.854	.001	0

Fig. 4. Image Authority (left) and Hub values (right) computed by WPicASHOW in response to query "Debian logo".

3 *IntelliSearch* Architecture

A complete prototype Web image retrieval system is developed and is accessible on the Web [11]. The system is implemented in Java. The architecture of *IntelliSearch* is illustrated in Fig. 5. It consists of several modules, the most important of them being the following:
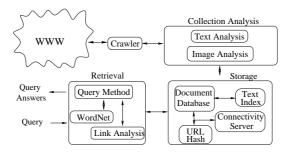

Fig. 5. *IntelliSearch* Architecture.

Crawler module: Implemented based upon Larbin [12], 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

---

[11] http://www.intelligence.tuc.gr/intellisearch
[12] http://larbin.sourceforge.net

and Linux products. The crawler worked recursively in breadth-first order and visited pages up to depth 5 links from each origin.

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.
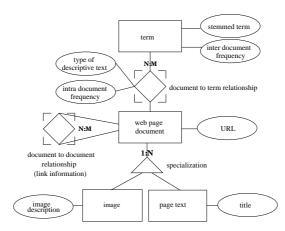

Fig. 6. The Entity Relational Diagram (ERd) of the database.

The Entity Relationship Diagram (ERD) of the database in Fig. 6 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.)

10

The database schema is implemented in BerkeleyDB [13] Java Edition. BerkeleyDB is an embedded database engine providing a simple Application Programming Interface (API) supporting efficient storage and retrieval of Java objects. The mapping of the ERD of Fig. 6 to database files (Java objects) was implemented using the Java Collections-style interface. Apache Lucene[14] 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 [19] 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.

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). All methods in Sec. 2 are implemented.

4   Conclusions

IntelliSearch[15], is a complete and fully automated system for retrieving text pages and images on the Web. It supports retrieval of important (authoritative) Web pages and images (by incorporating link analysis into its search and retrieval methods) as well as, searching by semantic similarity for discovering information related to the needs of the users (even if it is not explicitly specified in the queries). Retrievals are speeded-up by indexing text and link information specific to Web pages and images.

The results in [11, 16] indicate that text searching methods like VSM and SSRM are far more effective than link analysis methods (text is a very effective descriptor of Web content itself). However, text similarity methods tend to assign higher ranking even to Web pages and images pointed to by very low quality pages such as pages created by individuals or small companies. Between the two, SSRM demonstrated promising performance improvements over VSM.

Link information alone (e.g., as in HITS and PicASHOW) is not an effective descriptor for Web pages and images. 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.

IntelliSearch is currently being expanded to support queries by image content (e.g., queries by image example). This requires that image analysis methods be applied and appropriate image content representations extracted from images and used in retrievals. Future work includes also experimentation with larger data sets and more image types (i.e., video and graphics).

---

[13] http://www.sleepycat.com
[14] http://lucene.apache.org
[15] http://www.intelligence.tuc.gr/intellisearch

References

1. Arasu, A., Cho, J., Garcia-Molina, H., Paepke, A., Raghavan, S.: Searching the Web. ACM Transactions on Internet Technology **1**(1) (2001) 2–43
2. R. Baeza-Yates, E.: Modern Information Retrieval. Addison Wesley (1999)
3. Kherfi, M., Ziou, D., Bernardi, A.: Image Retrieval from the World Wide Web: Issues, Techniques, and Systems. ACM Computing Surveys **36**(1) (2004) 35–67
4. Smeulders, A.W., Worring, M., Santini, S., Gupta, A., Jain, R.: Content-Based Image Retrieval at the End of the Early Years. IEEE Transactions on Pattern Analysis and Machine Intelligence **22**(11) (2000) 1349–1380
5. Taycher, L., Cascia, M., Sclaroff, S.: Image Digestion and Relevance Feedback in the ImageRover WWW Search Engine. In: 2nd Intern. Conf. on Visual Information Systems, San Diego (1997) 85–94
6. Smith, J., Chang, S.F.: Visually Searching the Web for Content. IEEE Multimedia **4**(3) (1997) 12–20
7. Aslandongan, Y., Yu, C.: Evaluating Strategies and Systems for Content-Based Indexing of Person Images on the Web. In: 8th Intern. Conf. on Multimedia, Marina del Rey, CA (2000) 313–321
8. Shen, H.T., Ooi, B.C., Tan, K.L.: Giving Meanings to WWW Images. In: 8th Intern. Conf. on Multimedia, Marina del Rey, CA (2000) 39–47
9. Gevers, T., Smeulders, A.: The PicToSeek WWW Image Search Engine. In: IEEE ICMS. (1999)
10. Zhao, R., Grosky, W.: Narrowing the Semantic GapImproved Text-Based Web Document Retrieval Using Visual Features. IEEE Transactions on Multimedia (2) (2002) 189–200
11. Voutsakis, E., Petrakis, E., Milios, E.: Weighted link analysis for logo and trademark image retrieval on the web. (In: IEEE/WIC/ACM International Conference on Web Intelligence (WI2005)) 581–585
12. Kleinberg, J.M.: Authoritative Sources in a Hyperlinked Environment. Journal of the ACM **46**(5) (1999) 604–632
13. Page, L., Brin, S., Motwani, R., Winograd, T.: The PageRank Citation Ranking: Bringing Order to the Web. Technical report, Computer Systems Laboratory, Stanford Univ., CA (1998)
14. Bharat, K., Henzinger, M.R.: Improved Algorithms for Topic Distillation in a Hyperlinked Environment. In: Proc. of SIGIR-98, Melbourne (1998) 104–111
15. Lempel, R., Soffer, A.: PicASHOW: Pictorial Authority Search by Hyperlinks on the Web. ACM Transactions on Information Systems **20**(1) (2002) 1–24
16. Varelas, G., Voutsakis, E., Raftopoulou, P., Petrakis, E., Milios, E.: Semantic Similarity Methods in WordNet and their Application to Information Retrieval on the Web. In: 7th ACM International Workshop on Web Information and Data Management (WIDM 2005), Bremen, Germany (2005) 10–16
17. Salton, G.: Automatic Text Processing: The Transformation Analysis and Retrieval of Information by Computer. Addison-Wesley (1989)
18. Li, Y., Bandar, Z.A., McLean, D.: An Approach for Measuring Semantic Similarity between Words Using Multiple Information Sources. IEEE Trans. on Knowledge and Data Engineering **15**(4) (2003) 871–882
19. Bharat, K., Broder, A., Henzinger, M.R., Kumar, P., Venkatasubramanian, S.: The Connectivity server: Fast access to Linkage Information on the Web. In: Proceedings of the 7th International World Wide Web Conference (WWW-7), Brisbane, Australia (1998) 469–477