Weighted Link Analysis for Logo and Trademark Image Retrieval on the Web

Epimenides Voutsakis, Euripides G.M. Petrakis
Dept. of Electronic and Comp. Engineering
Technical University of Crete (TUC)
Chania, Crete, Greece
pimenas@softnet.tuc.gr, petrakis@ced.tuc.gr

Evangelos Milios
Faculty of Comp. Science
Dalhousie University
Halifax, Nova Scotia, Canada
eem@cs.dal.ca

Abstract

Image retrieval on the Web requires that important (authoritative) images satisfying the query selection criteria are assigned higher ranking over other relevant images. PicASHOW [5] achieves this goal using link information alone. This work proposes WPicASHOW (Weighted PicASHOW) a weighted scheme for co-citation analysis that incorporates within the link analysis method of PicASHOW the text and image content of the queries and of the Web pages. WPicASHOW is implemented and integrated into a fully automated Web retrieval system for logo and trademark images.

1. Introduction

Several approaches to the problem of content-based image retrieval on the Web have been proposed and some have been implemented on research prototypes (e.g., WebSEEK [7], Diogenis [1]) and commercial systems. The last category of systems, includes general purpose image search engines such as Google Image Search [1] and Yahoo [2], as well as systems providing specific services to users such as unauthorized use of images (e.g., ImageLock [3]) and Web and e-mail content filters (e.g., Cobion [4]). Focusing mainly on image and text content, the work referred to above does not show how to select good quality Web pages and images on the topic of the query. Link analysis methods such as HITS [4] estimate the quality of Web pages and the topic relevance between the Web pages and the query. Building upon HITS, PicASHOW [5] shows how to handle pages that link to images and pages that contain images. However, PicASHOW does not show how to handle image content and queries by image example. This is exactly the focus of this work.

The contributions of the proposed work are summarized in the following:

- WPicASHOW, a weighted scheme for link analysis is proposed. It relies on the combination link information with text and image content for regulating the influence of links between pages and for retrieving relevant images from high quality Web pages.
- A complete and fully automated system is proposed and implemented for retrieving logo and trademark images on the Web. The system supports indexing and storage for Web pages and information extracted from Web pages (e.g., image descriptions, URLs, link information).

Extraction of image descriptions from Web pages and image similarity measures are discussed in Sec. 2. WPicASHOW, is presented in Sec. 3. A content-based retrieval of logo and trademark images that integrates the above ideas is presented in Sec. 4. Experimental results are presented and discussed in Sec. 5 followed by conclusions in Sec. 6.

2. Image Content Representation

We choose the problem of logo and trademark images as a case study for the evaluation of the proposed methodology. Because images are not properly categorized on the Web, filters based on learning by decision trees for selecting Web pages with logo and trademark images are designed and implemented.

2.1. Text Description

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

[1] http://www.google.com/imghp
[2] http://images.search.yahoo.com
[3] www.imagelock.com
[4] www.cobion.com

Image Filename: The URL entry 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 if the image fails to load. This attribute is optional (i.e., is not always present).

Page Title: 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). In either case, the caption is limited to 30 words before or after the reference to the image file. If neither case applies, the caption is considered to be empty.

All descriptions are syntactically analyzed and reduced into vectors of stemmed terms (nouns). Similarly, text queries are also transformed to term vectors and matched against image term vectors according to the Vector Space Model (VSM) [6]. The text similarity between the query Q and the image I is computed as

$$S_{text}(Q,I) = S_{file\_name}(Q,I) + S_{alternate\_text}(Q,I) + S_{page\_title}(Q,I) + S_{caption}(Q,I). \quad (1)$$

Each S term is computed according to the Vector Space Model as a distance between vectors of tf-idf weights without normalizing by query term frequencies (not required for short queries).

2.2. Image Content Descriptions

Because the same logo or trademark image may appear as color or grey scale image in different Web pages, color information is not useful in content representations. All images are converted to grey scale. For logo and trademark images the following features are computed:

Intensity Histogram: Shows the distribution of intensities over the whole range of intensity values.

Energy Spectrum [8]: Describes the image by its frequency content. It is computed as a histogram showing the distribution of average energy over 256 co-centric rings (with the largest ring fitting the largest inscribed circle of the DFT spectrum).

Moment Invariants [8]: Describes the images by its spatial arrangement of intensities. It is a vector of 7 moment coefficients.

The purpose of this type of representations is twofold:

Logo-Trademark Detection: A five-dimensional vector is formed: Each image is specified by the mean and variance of its intensity and energy spectra plus a count of the number of distinct intensities per image. A set of 1,000 image examples is formed consisting of 500 logo-trademark images and 500 images of other types. Their feature vectors are fed into a decision-tree [9] which is trained to detect logo and trademark images. The estimated classification accuracy by the algorithm is 85%. For each image, the decision tree computes an estimate of its likelihood of being logo or trademark or "Logo-Trademark Probability".

Logo-Trademark Similarity: The image similarity between a query image Q and a Web image I is computed as

$$S_{image}(Q,I) = S_{intensity\_spectrum}(Q,I) + S_{energy\_spectrum}(Q,I) + S_{moments}(Q,I). \quad (2)$$

The similarity between histograms is computed by their intersection whereas the similarity between their moment invariants is computed by subtracting the Euclidean vector distance from its maximum value.

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

$$S(Q,I) = S_{image}(Q,I) + S_{text}(Q,I). \quad (3)$$

All measures are normalized to lie in the interval [0,1].

3. Weighted PicASHOW (WPicASHOW)

PicASHOW [5] 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. PicASHOW computes authority and hub values by link analysis on the *query focused graph* $\mathcal{F}$ (i.e., a set of pages formed by initial text query results expanded by backward and forward links).

The following matrices are defined on $\mathcal{F}$:

$\mathcal{W}$**:** The page to page adjacency matrix (as in HITS) relating each page in $\mathcal{F}$ with the pages it points to.

$\mathcal{M}$**:** The page to image adjacency matrix relating each page in $\mathcal{F}$ with the images it contains.

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

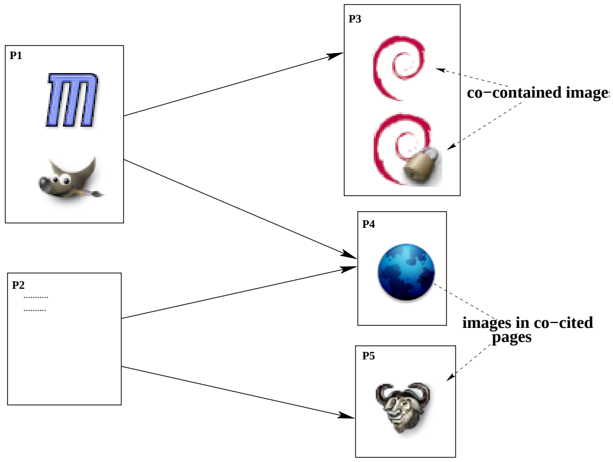

Figure 1. A query focused graph.

PicASHOW computes image authority and hub values of images 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, cannot answer queries by image example and similarly to HITS, it suffers from the following problems [3]:

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. Then, the highest authority and hub pages tend not to be related to the topic of the query.

WPicASHOW handles all these issues:

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 according to the Vector Space Model 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 depending on query type: For text (e.g., keyword) queries the weights are computed according to Eq. 1 (as the similarity between the query and the descriptive text of an image). For queries combining text and image example, the weights are computed according to Eq. 3 (as the average of similarities between the text and image contents of the query and the image respectively).

In WPicASHOW the query focused graph $\mathcal{F}$ is formulated as follows:

- An initial set $\mathcal{F}$ of images is retrieved. These are images contained or pointed-to by pages matching the text query according to Eq. 1.
- Non-informative images such as banners, bars, buttons or mail-boxes are excluded from $\mathcal{F}$ utilizing simple heuristics (e.g., very small file size). Images with logo-trademark probability less than 0.5 are excluded as well. At most T images are retained and this limits the size of $\mathcal{F}$. In this work $T = 10,000$.
- $\mathcal{F}$ is expanded with pages pointing to images in $\mathcal{F}$.
- $\mathcal{F}$ is further expanded to include pages and images that point to pages or images already in $\mathcal{F}$. To limit the influence of popular sites, for each page in $\mathcal{F}$, at most t new pages are included ($t = 100$ in this work).
- The last two steps are repeated until $\mathcal{F}$ contains T pages and images.

WPicASHOW computes the answer to the query by ranking the elements of the principal eigenvector of the image co-citation matrix $[(\mathcal{W}+\mathcal{I})\mathcal{M}]^{\mathcal{T}} \cdot (\mathcal{W}+\mathcal{I})\mathcal{M}$ by their authority values.

4. System Architecture

A complete prototype Web image retrieval system is developed as part of this work. Figure 2 illustrates the architecture of the proposed system. The system consists of several modules the most important of them being the:

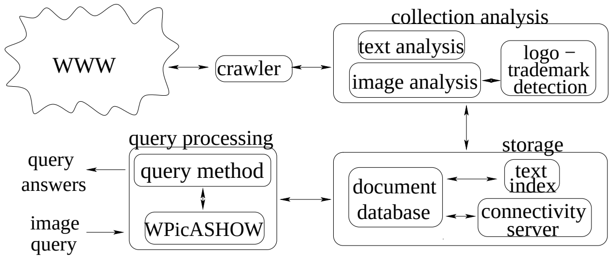

Figure 2. System Architecture.

Crawler module: Implemented based upon Larbin [5], the crawler assembled locally a collection of 250,000 pages and a similar number of 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 [6] to 20 queries on various topics (e.g., topics related to Linux and software 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. For each image, its text and image description and its logo-trademark probability are computed (Sec. 2.2).

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, 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 and logo-trademark probabilities.

Query Processing module: Queries are issued by keywords or free text or as a combination of text and image example. Implements WPicASHOW (Sec. 3).

The database is implemented in BerkeleyDB [7]. Two inverted files implement the connectivity server [2] and provide fast access to linkage information between pages (backward and forward links) and two inverted files associate terms with their intra and inter document frequencies and allow for fast computation of term vectors.

5. Experiments

The system is implemented under Linux in Perl, Java and C/C++ and the database stores 250,000 Web pages with images. The following methods are evaluated:

PicASHOW [5]: Ranks Web images by exploiting co-citation information only. It can answer text queries by explpoiting text content and co-citation information.

WPicASHOW: The proposed method. It can answer both text and image queries. It exploits co-citation, image and text content information.

Vector Space Model (VSM) [6]: The state-of-the-art text retrieval method. Text queries are processed by searching the database according to Eq. 1. It exploits only text content (no co-citation or image content information). Queries by text and image example are processed by searching the database according to Eq. 3.

For the evaluations, 20 queries are created on topics related mainly to Linux and software. The evaluation is based on human relevance judgments.

Each method is represented by a *precision-recall* curve. Each query retrieves the best 30 answers and each point in a curve is the average precision and recall over 20 queries. A method is better than another if it achieves better precision and recall.

5.1. Text Queries

All queries specified the term “logo”. The query and a Web image are similar if they are on the same topic. Query “Linux logo” may retrieve the logo image of any Linux distribution (e.g., “Debian Linux”).

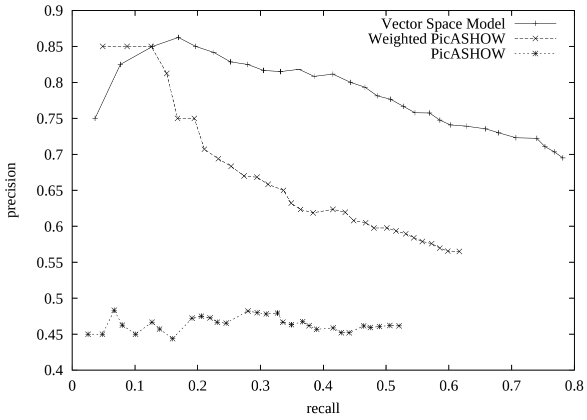

Figure 3. Precision-recall for text queries.

Fig. 3 illustrates that PicASHOW is the worst method. This result indicates that link information alone is not an effective descriptor for image content. PicASHOW suffers from topic drift and retrieved many irrelevant images which either coexisted within the same pages with other relevant images, or are pointed to by high quality pages (e.g., pages of software companies). WPicASHOW is more effective than PicASHOW, handled topic drift and assigned higher ranking to images whose text is more relevant to the topic of the query. VSM is more effective (except from the first 3 answers) but retrieved many low quality images (from sites created by small companies and individuals).

[5] http://larbin.sourceforge.net

[6] http://www.google.com/imghp

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

5.2. Text Queries with Image Example

Each keyword query is augmented by a logo image. An answer is similar to the query if it is on the same topic and shows a similar image. Query “Linux logo” with a penguin logo is similar to answers showing a Linux penguin logo.

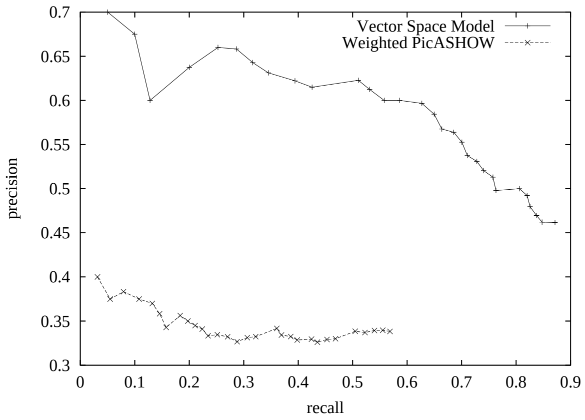

Figure 4. Precision-recall for image queries.

PicASHOW cannot answer such queries. Fig. 4 illustrates that VSM is more effective than WPicASHOW. A closer look into the answers reveals that WPicASHOW assigned higher ranking to images in Web pages with more general content on the topic of the query. The reason for this behavior is that Web pages with more general content are more strongly connected than pages with more specific topic. In this experiment, with the addition of logo image, the queries become more specific than before and WPicASHOW assigned higher ranking to more general but less similar images.

5.3. Discussion

The experimental results indicate that surrounding text is a very effective descriptor of the image itself. VSM will therefore retrieve the most relevant images regardless of quality. However, this is not desirable in Web searches. PicASHOW, as any other link analysis method would do, assign higher ranking to higher quality but not necessarilly relevant images. Therefore, a link analysis method alone cannot be more accurate than VSM. WPicASHOW attempted to compromise between the two. In any case, WPicASHOW is more effective than PicASHOW, the state-of-the-art method for image retrieval on the WEB.

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

6. Conclusions

WPicASHOW, an approach that incorporates text and image content into the analysis of Web link structure is proposed. A complete prototype Web retrieval system for logo and trademark images is also proposed and implemented. WPicASHOW allows for more sophisticated image queries such as queries by example image in addition to text queries. The results demonstrated that WPicASHOW is far more effective than PicASHOW [5] using link information alone. WPicASHOW improves the quality of the results but not necessarily their accuracy (at least for data sets smaller than the whole Web).

The analysis reveals that content relevance and searching for authoritative answers can be traded-off against each other: Giving higher ranking to important pages seems to reduce the accuracy of the results. Given the results, it may be desirable to allow the user to adjust the tradeoff between text, link and image similarity contributions in ranking the results, something that WPicASHOW allows.

References

[1] Y. Aslandongan and C. Yu. Evaluating Strategies and Systems for Content-Based Indexing of Person Images on the Web. In 8th *Intern. Conf. on Multimedia*, pages 313–321, Marina del Rey, CA, 2000.

[2] K. Bharat, A. Broder, M. R. Henzinger, P. Kumar, and S. Venkatasubramanian. The Connectivity server: Fast access to Linkage Information on the Web. In *Proceedings of the 7th International World Wide Web Conference (WWW-7)*, pages 469–477, Brisbane, Australia, 1998.

[3] K. Bharat and M. R. Henzinger. Improved Algorithms for Topic Distillation in a Hyperlinked Environment. In *Proc. of SIGIR-98*, pages 104–111, Melbourne, 1998.

[4] J. M. Kleinberg. Authoritative Sources in a Hyperlinked Environment. *Journal of the ACM*, 46(5):604–632, 1999.

[5] R. Lempel and A. Soffer. PicASHOW: Pictorial Authority Search by Hyperlinks on the Web. *ACM Trans. on Info. Systems*, 20(1):1–24, Jan. 2002.

[6] G. Salton and C. Buckley. Term-weighting approaches in automatic text retrieval. *Information Processing and Management*, 24(5):513–523, 1988.

[7] J. Smith and S.-F. Chang. Visually Searching the Web for Content. *IEEE Multimedia*, 4(3):12–20, July-Sept. 1997.

[8] M. Sonka, V. Hlavec, and R. Boyle. *Image Processing Analysis, and Machine Vision*, chapter 6 & 14. PWS Publishing, 1999.

[9] I. H. Witten and E. Frank. *Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations*, chapter 4. Morgan Kaufmann, Academic Press, 2000.