A Relevance Model for Web Image Search*

Cheng Thao
Ethan V. Munson

Department of EECS
University of Wisconsin-Milwaukee
Milwaukee, WI 53211 USA
E-mail: {chengt,munson}@cs.uwm.edu

Abstract

This article describes the construction of a relevance model for Web image search. Using custom image retrieval software, 24 textual queries were used to retrieve over 5800 images. Each image's relevance to its query was evaluated by human raters. The Web documents containing these images were analyzed for the presence of text matching the query in each of 53 HTML features. Finally, logistic regression was used to construct the relevance model that best predicted the human ratings from the presence of matching text in HTML features. The resulting relevance model has a precision of over 65% when applied to our entire sample. It uses a total of thirteen HTML features with image filename and document title being the most important. A number of methodologic issues are discussed and suggestions for future research are made.

1 Introduction

The World Wide Web (Web) is a virtual information space that connects millions of computers around the globe. Many Web pages use images for communicating content, for page layout, for navigation, or for decoration. In the year 2000, the Web was estimated to contain 489 million images with new images being added at the rate of one million a day [2]. So, it is natural to view the Web as a very large, unstructured database of images.

It is also natural to search for and retrieve images from the Web. Several commercial image search engines have been developed. Google, for example, has indexed over 425 million images and allows users to search this index using textual queries. The site's frequently asked questions give a brief description of how Google find images that match users' queries:

> Google analyzes the text on the page adjacent to the image, the image caption and dozens of other factors to determine the image content. Google also uses sophisticated algorithms to remove duplicates and ensure that the highest quality images are presented. [5]

We are interested in understanding effective mechanisms for image retrieval on the Web, particularly those that exploit the content and structure of HTML, rather than the pixels of the images themselves. Our earlier research [10] showed that text-based search can be quite precise, but also suggested that structural features of the HTML source were not very useful in determining image relevance. That study was quite small and did not attempt to combine multiple features into a single model of image relevance.

In this article, we present the results of a larger study that does combine multiple features of Web pages into a single model of image relevance. We built software tools to support image search and used them to gather image search results for twenty-four two-word queries. The images returned were rated for relevance by three human raters and the pages containing the images were analyzed for the presence of text matching the query in any of 53 different HTML features. Then, we used the statistical technique of logistic regression to create a relevance model that uses the presence of matching text in HTML features to predict the human ratings of relevance.

2 Background

Extensive image retrieval research has been performed with local databases, where the number of images ranges from the thousands to a few millions. This work can be broadly classified into two categories: text-based [3] and content-based [11].

*This material is based upon work supported by the U. S. Department of Defense and by the National Science Foundation under Grant No. 9734102.

In traditional text-based image retrieval [3], images are annotated with text that describes the semantics of the image. Annotations can describe various metadata information, such as the subject, location, and time of the image, using vocabularies and formats that generally depend on the image domain. Images are retrieved by comparing textual queries to the annotations. For Web image retrieval, the critical problem with manual annotation is scale. The Web is too large and its management is too chaotic for manual annotation of images to be feasible.

Content-based image retrieval (CBIR) focuses on automated indexing of image content rather than relying on manual annotation. In CBIR, images are analyzed for various low level visual features, such as color, shape, and texture. A user can construct a CBIR query in several ways: by specifying visual features directly, by submitting an example of the type of image desired, or by sketching an approximation of the desired image. All CBIR query approaches share the problem that human concepts do not map neatly to low level image features. Thus, while CBIR is effective at finding images that have shared appearance, it is less effective at finding images with shared meaning. For Web image search, CBIR techniques must also cope with the scale of the Web. The image processing techniques used by CBIR require substantial computation. The construction of a feature-based index of the entire Web would be a daunting task and efficient matching of queries against such an index appears to be an open research question.

An important quality of Web images is that they are always accompanied by HTML source code. Often, this source code can function like manual annotations, because it describes the very concepts that are shown in the images. This fact has led several researchers to exploit HTML source as a resource for image retrieval. WebSeer [4] and Diogenes [1] combined text-based and content-based approaches to retrieve Web images containing human faces. All three systems used a small set of textual features that were combined with heuristically assigned weights to identify candidate images and then used a face detector to identify images of faces among the candidates. Shen *et al* [8] used four HTML features to identify likely image content: the image filename, the value of the ALT attribute, the page title, and surrounding text. They combined them using the Weighted ChainNet approach, which is taken from natural language processing research, and uses heuristically assigned weights. The system most similar to the work presented here is MARIE-4 [7], which uses a statistically-derived relevance model to decide whether an HTML feature is an image caption. Once a caption is identified, its text is used to construct entries for the image in a search index.

3 Method

This study was conducted in four phases: query selection, query result downloading, relevance rating, and document analysis.

3.1 Query selection

A relevance model that would be effective for Web image search in general must be suitable for a wide range of realistic queries. We identified eight query categories: famous people, non-famous people, famous places, less-famous places, holidays, concepts, phenomena, and landmarks. Then, for each category, we selected three queries for a total of twenty-four queries (see Thao [9] for details.) Three queries contained only one word, while the remainder contained two words.

3.2 Query result downloading

Using new image retrieval tools, each query string was sent to Google, which returned the 1000 pages that best matched the query string (according to Google's proprietary relevance model.) For each query, one hundred of these pages were selected at random. When Google returned a bad link, a replacement page was selected randomly. A total of 2400 Web pages and all the images that they referenced were downloaded. Images that were small (typical of decorative images, such as bullets) or had extreme aspect ratios (typical of advertisements) were discarded. A total of 5806 images were retained. The pages were modified so that their image references pointed to the downloaded copies of the images. The original URLs of the pages and images were recorded in a database along with other study data.

3.3 Relevance rating

Each of the 5806 images was examined by three trained raters, who each chose one of three values for an image. An image was rated *Relevant* if it was a picture of the idea represented by the query string. An image was rated *Somewhat Relevant* if it was not Relevant but did show content directly related to the query. Otherwise, the image was *Not Relevant*. For example, if the query was "bill gates," then a picture of a person named "Bill Gates" was Relevant, a picture of Bill Gates's home was Somewhat Relevant, and a picture of Donald Knuth was Not Relevant.

For this study, we have simplified these ratings by converting the three point scale to a two-point scale (Relevant vs. Not Relevant). An image is considered relevant if two of the three raters gave it a Relevant rating.

3.4 Document analysis

Finally, the HTML source code of the Web pages was analyzed for the presence of text matching the query in any of 53 different HTML features. The 53 HTML features included page-level features (URL, title, metadata), image element features (URL, identifier, alternative text), and the textual content of links, objects, related parts of tables, nearby headings, and elements that emphasize text in various ways.

When multiple words were present in a query string, there were several alternatives for defining a match. In general, we used a form of "phrase matching," requiring that matching words appear in the same order as in the query and were not separated by more than twenty characters. An exception to this rule was that only one matching word was required to appear in a URL.

4 Results

In evaluating any information retrieval study, it is useful to consider the classic information retrieval statistics of recall and precision. *Recall* is the proportion of all relevant items in a collection that a retrieval technique returns. *Precision* is the proportion of returned items that are relevant. In general, the scale of the Web is so enormous that precision is of far more practical importance than recall.

Of the 5806 images in this study, 1447 or 24.9% were rated Relevant. This is an important point of reference because it means that a completely random relevance model can have a precision of 24.9%.

We used logistic regression to construct our relevance model. Logistic regression is a non-linear regression technique that is well-suited to categorical or binary data [6]. It yields a formula whose value can be used to estimate the probability that an event will occur. In traditional regression terminology, the event is the dependent variable and the model's predictors are the independent variables. Logistic regression models the probability of an event based on k independent variables, $X_1 \ldots X_k$ as

$$P_{event} = \frac{e^Z}{1+e^Z} = \frac{1}{1+e^{-Z}}$$

$$\text{where } Z = B_0 + B_1X_1 + B_2X_2 + \ldots + B_kX_k.$$

It finds a set of values for the coefficients $B_1 \ldots B_k$ using the maximum-likelihood method, which maximizes the probability of the observed results occurring. The solution is found by an iterative approximation algorithm.

In this study, the dependent variable was the human relevance rating, with a rating of Relevant having a value of one and a rating of Not Relevant having a value of zero. Each of the 53 potential independent variables was a binary value that was true (one) when a particular HTML feature contained matching text and was false (zero) otherwise. The 53 potential independent variables were reduced to a total of 13 variables in the relevance model by using the forward stepwise procedure to restrict the set of independent variables to contain only those variables whose coefficient had a Wald statistic that was significant at the $p = .05$ level. The thirteen HTML features in the model are shown in Table 1 along with their coefficients in the model, Wald statistic values and significance levels. Positive coefficients indicate that matching text in a feature increases the probability of relevance while negative coefficients indicate reduced probability of relevance.

Feature	B	Wald	sig.
Constant	-2.317	1137.236	.000
Image filename	1.886	299.599	.000
Page title	1.092	177.592	.000
Page filename	.867	96.862	.000
ALT attribute	1.076	43.774	.000
Image path	1.060	36.816	.000
Page path	-.787	18.030	.000
Cell below	.709	11.509	.001
Meta description	1.092	9.615	.002
Cell above	.664	9.164	.002
Other body text	-.222	9.153	.002
Anchor text	2.023	8.580	.003
Cell right	.331	4.690	.030
Cell left	.370	3.830	.050

Table 1. Relevance model computed by logistic regression analysis.

	Model		Percent
Raters	**Relevant**	**Not**	**Correct**
Relevant	531	916	36.7
Not	274	4085	93.7

Table 2. Classification table comparing human ratings and the relevance model.

The practical significance of a logistic regression model can be evaluated in two ways. First, there are two approximations to the traditional R^2 measure of *variance accounted for* for regression. For the image relevance model, the Cox & Snell R^2 is .223 and the Nagelkerke R^2 is .330. Thus, the model accounts for somewhere between 22% and 33% of the variance in image relevance.

The second approach is to compare the relevance classification produced by the model when using a particular cut

point to that of the human raters in a classification table, as shown in Table 2. For the purposes of this table, images were classified as relevant by the model if their estimated probability under the model was greater than 0.5. The classification table shows that 93.7% of the Not Relevant images and 36.7% of the Relevant images were correctly classified. The latter number is the recall statistic. The model's precision is 66.0%.

5 Discussion

The results described in the previous section show that it is possible to construct a useful relevance model for image search on the Web. While neither the 66% precision statistic or the 37% recall statistic are particularly high, the precision level is a substantial improvement over the 25% chance of choosing a relevant image from our sample by random chance and should be high enough to be acceptable in practice.

It is possible to make some interesting observations about the HTML features that play a role in the model. First, these results confirm our earlier study [10], which found that the two most important HTML features are the filename under which the image is stored and the Web page's TITLE element. Our earlier study also found the ALT attribute value of the image element to be useful and this result is confirmed here.

Most of the other features with strong influence are also parts of the image or page URL. Interestingly, the path portion of the page's URL has a negative relationship with relevance. This suggests that when the page's path contains matching text, then the matching text is widely used for file and directory names on the site, and thus conveys less meaning than is normally the case.

Many of the features that have weaker contributions to the model are parts of tables. Because tables are widely used for layout in HTML, they are often used to place captions near images.

Finally, no features based on text emphasis, the use of heading elements, or ID, NAME, or TITLE attributes of elements appear in the model. We have examined the overall frequency of appearance of certain features. The advent of CSS appears to be eliminating the use of HTML's heading elements (H1 . . . H6), especially when authoring tools are used to create Web pages. To a lesser extent, the same is true of the text emphasis elements, such as the B, IT, and EM elements. The ID, NAME, and TITLE attributes are also hardly ever used.

6 Future Work

We plan to continue our analysis of the data set from which we constructed this relevance model with particular attention to differences between query types. Early analysis results suggest that queries based on proper names may have different characteristics than other queries. If this is indeed the case, it should be possible to use lightweight natural language techniques to identify proper name queries and choose an appropriate relevance model.

The growing use of CSS to control formatting means that HTML elements convey even less semantics than they once did. So, we want to explore how analysis of CSS style sheets can help produce an enhanced relevance model.

References

[1] Y. A. Aslandogan and C. T. Yu. Evaluating strategies and systems for content based indexing of person images on the Web. In *Proceedings of the eighth ACM international conference on Multimedia*, pages 313–321. ACM Press, 2000.

[2] Censorware Project. Size of the Web: A dynamic essay for a dynamic medium. http://censorware.org/web_size.

[3] S.-K. Chang and A. Hsu. Image information systems: Where do we go from here? *IEEE Transactions on Knowledge and Data Engineering*, 1992.

[4] C. Frankel, M. Swain, and V. Athitsos. WebSeer: An image search engine for the World Wide Web. Technical Report 96-14, University of Chicago, Department of Computer Science, July 1996.

[5] Google, Inc. Google frequently asked questions: Image search. http://www.google.com/help/faq_images.html, Apr. 2003.

[6] M. J. Norušis. *SPSS Regression Models 10.0*. SPSS, 1999.

[7] N. C. Rowe. Marie-4: a high-recall, self-improving web crawler that finds images using captions. *IEEE Intelligent Systems*, 2002.

[8] H. T. Shen, B. C. Ooi, and K.-L. Tan. Giving meanings to WWW images. In *Proceedings of the eighth ACM international conference on Multimedia*, pages 39–47. ACM Press, 2000.

[9] C. Thao. A relevance model for Web image search. Master's thesis, University of Wisconsin-Milwaukee, Aug. 2003. In preparation.

[10] Y. Tsymbalenko and E. V. Munson. Using HTML metadata to find relevant images on the Web. In *Proceedings of Internet Computing 2001, Volume II, Las Vegas*, pages 842–848. CSREA Press, June 2001.

[11] R. Veltkamp and M. Tanase. Content-based image retrieval systems: A survey. Technical Report UU-CS-2000-34, Dept. of Computing Science, Utrecht University, 2000.