2002-07

A High-Recall Self-Improving Web Crawler That Finds Images Using Captions

Rowe, Neil C.

IEEE Intelligent Systems, July/August2002
https://hdl.handle.net/10945/35973

A High-Recall Self-Improving Web Crawler That Finds Images Using Captions

Neil C. Rowe

U.S. Naval Postgraduate School
Code CS/Rp, 833 Dyer Road
Monterey, CA 93943
USA
(831) 656-2462, ncrowe@nps.navy.mil

ABSTRACT

Finding multimedia objects to meet some need is considerably harder on the World Wide Web than finding text because content-based retrieval of multimedia is much harder than text retrieval and caption text is inconsistently placed. We describe a Web "crawler" and caption filter MARIE-4 we have developed that searches the Web to find text likely to be image captions and its associated image objects. Rather than examining a few features like existing systems, it uses broad set of criteria including some novel ones to yield higher recall than competing systems, which generally focus on high precision. We tested these criteria in careful experiments that extracted 8140 caption candidates for 4585 representative images, and quantified for the first time the relative value of several kinds of clues for captions. The crawler is self-improving in that it obtains from experience further statistics as positive and negative clues. We index the results found by the crawler and provide a user interface. We have done a demonstration implementation of a Web search engine for all 667,573 publicly-accessible U.S. Navy Web images.

Keywords: Images, captions, World Wide Web, software agents, data mining, digital libraries, information filtering, keywords, parsing, image processing, probabilistic reasoning, servlets.

To appear in *IEEE Intelligent Systems,* July/August2002.

1. INTRODUCTION

We are building intelligent software agents to find information on the World Wide Web ("crawlers" or “spiders”). Images are among the most valuable assets of the Web, permitting it to be an extensive virtual picture library. But finding the images on the Web that match a query is quite difficult: Typically only a small fraction of the text on pages describes associated images, and images are not captioned consistently. Progress is being made with content-based image retrieval systems that analyze the images themselves [1] but the systems require considerable image preprocessing time. Furthermore, surveys of users attempting image retrieval show they are much more interested in the identification of objects and actions depicted by images than in only the color, shape, and other visual properties that most content-based retrieval provides [2]. Since object and action information is much more easily obtained from captions, caption-based retrieval appears the only hope for broadly useful image retrieval from the Web [3].

Commercial tools like the "Image Search" in the AltaVista search engine achieve respectable precision (the fraction of correct answers retrieved in all answers retrieved) by indexing only "easy" pages, like those from photograph libraries where images are one to a page and captions are easy to identify. Recall (the fraction of correct answers retrieved of all correct answers) is often equally or more important, but users often do not realize how poor it is for their queries. In experiments with ten representative phrases, we found Alta Vista Image Search had a precision of 0.46 and recall of 0.10, using inspection of pages retrieved by traditional keyword-based Alta Vista search to calculate recall. Higher recall than that requires dealing with a large variety of page layout formats and styles of captioning.

Recent work has made important progress on general image indexing from the Web by intelligent "information filtering" of Web text. By looking for the right clues, large amounts of the text on a Web page can be excluded as captions for any given image, and good guesses as to the captions in the remaining text can be inferred [4, 5, 6]. Clues can include wording of the caption candidate; HTML constructs around the candidate; distance from the associated image; words of the image file name; and properties of the associated image. These clues reduce the amount of text that needs to be examined to find captions, and the reduced text can be indexed and used for keyword-based retrieval. But so far the selection of these clues has been intuitive, and there has been no careful study of the relative values of clues.

This paper reports on MARIE-4, our latest in a series of caption-based image-retrieval systems [7]. MARIE-4 uses a wide range of clues, broader than any system we know about, to locate image-caption pairs in HTML web pages. It is in part an expert system in which the knowledge used is not especially novel in itself, but the synergy of a variety of knowledge working together provides surprisingly good performance. Unlike [3] and previous MARIE systems which required an image database with captions already extracted, MARIE-4 is a "Web crawler" that autonomously searches the Web, locates captions using intelligent reasoning, and indexes them. MARIE-4 does not attempt full natural-language processing and does not require the elaborate lexicon information of the earlier prototypes, so it is more flexible.

Figure 1 shows a block diagram of MARIE-4, built in Java 2. A crawler searches the World Wide Web from a given starting page. It locates all images on each page and good candidate captions for them. This information is passed to a caption rater that assigns a likelihood to each image-caption pair based on a weighted sum of factors, and (for selected images) to a caption tagger that permits a user to confirm captions manually for training and testing. The indexer indexes the inferred caption words, and the Web-based query interface uses them to answer queries in the form of keywords by providing images that match those keywords, sorted in order of decreasing likelihood of match. We also developed a discriminator for photographs from graphics, but experiments showed it did not help much for caption extraction and we do not use it in the final system.

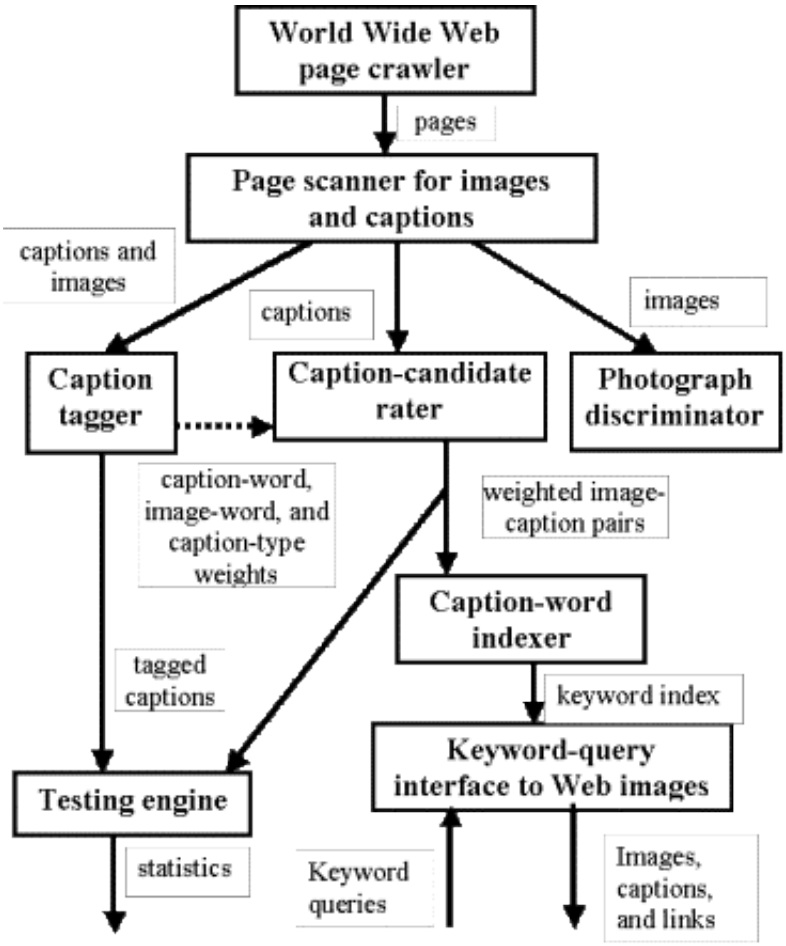

Figure 1: MARIE-4 organization.

2. THE WEB CRAWLER AND PAGE SCANNER

The MARIE-4 Web crawler searches thoroughly using a rule-based expert system to ensure high recall of captions. It fetches the HTML source code for a given page and scans it for image references. It also finds the links to other pages (“HREF”, “FRAME”, “AREA”, and certain Javascript constructs) which it puts into a queue; pages are subsequently considered in queue order to give a breadth-first search. To localize the search, it only examines pages with the same last K words in their site name as the initial page, where K and the initial page are specified by the user. So if K=2 and the starting site is nps.navy.mil, cs.nps.navy.mil and www.navy.mil would be considered but not www.army.mil. A site-URL hash table prevents revisiting the same page and a page-content hash table prevents visiting a page with the same content.

Image references in HTML are both "IMG" constructs and “HREF” links to files with image extensions like ".GIF" and ".JPEG". The page scanner searches for captions near each image reference. The kinds of captions considered are: (1) the “filename" or words (with punctuation and low-information "stop words" removed) of the full path to the image file; (2) any "ALT" string associated with the image, which represents associated text; (3) clickable text that retrieves the image; (4) text delineated by HTML constructs for fonts, italics, boldface, centering, table cells and rows, and explicit captions (the latter quite rare); (5) the title and nearest-above headings on the page (but not "meta" constructs since we found them often unreliable); (6) unterminated or unbegun paragraph ("P") constructs; and (7) specific word patterns of image reference (e.g. “Figure 5.1”, “in the photo above”, and “view at the right”), as found by partial parsing using a context-free grammar of image references and then checking consistency of reference direction (e.g. “above” should refer to an image above the caption). The first four categories were used in [4] but were improved here; the fifth was used in [5]; and the sixth and seventh are apparently new with our work here. We found that identifying these specific types of captions was considerably more precise and successful than just assigning a word weight which was a decreasing function of distance from the image reference [6].

The fourth, sixth, and seventh types require the caption candidate be within image-distance bounds which we set, based on experiments, to be 800 characters of an image reference above or below, provided the intervening characters do not contain a structural boundary, or 1500 if the candidate's construct surrounds the image reference. After experiments, we determined that a structural boundary can be usually another image construct, end of a table row, a horizontal line, a beginning or end of a paragraph when searching for weaker constructs, and an opposite of a sought construct (e.g. if we encounter the end of italics when we are searching left for the start of italics). The rules for scope that we inferred are complex to handle all the special cases, and require a carefully designed rule-based system.

As an example, consider a Web page with HTML source text "<title>Sea Otters</title><h2>The California Sea Otter</h2><img src="images/smallotter.gif" alt='Pair of sea otters' <I>Click on the above to see a larger picture.</I> <hr>Go to home page". This is a page with a small image "smallotter.gif" that when clicked retrieves a larger image "otter.jpeg". Four caption candidates are found for both images: a title of "Sea Otters", heading-font text of "The California Sea Otter", an "alt" string of "Pair of sea otters", and italicized centered text of "Click on the above to see a larger picture." In addition, the larger image has a filename caption candidate of "images otter jpeg" and the smaller has "images otterf18 gif". "Go to home page" is not a candidate since it is separated by a horizontal line ("hr") from the image reference.

Several criteria prune candidate captions. Captions on images not retrievable from the Web (incorrect links or those removed since using the page scanner) are excluded by testing the links. HTML and Javascript syntax is removed from the candidates, and subsequent null captions are eliminated. Small images or those not reasonably square are more likely to be graphics and hence unlikely to have captions; so we require that width and height be greater than 80 pixels and that the length-to-width ratio be less than 3. (Image-file sizes are retrieved from the Web to estimate image sizes not specified on the Web page.) Images appearing more than once on a page, and images appearing on three or more different pages, are eliminated from consideration for captions since such images are almost always iconic and uncaptionable. We also eliminate duplicate captions, and only quoted constants are examined within Javascript code (since full analysis would require implementation of a nondeterministic interpreter). These criteria were derived from experiments in which we found thresholds that eliminated less than 1% of the correct candidates.

2.1. Testing the page scanner

A training set was used of 3945 caption candidates from 14 representative sites with images (the first 300 candidates found there, or all

if the site had fewer than 300), each of which the author manually inspected to confirm it not only was a caption but captioned the referenced image. Assessment of a caption is necessarily subjective, but the principle applied was that a caption should describe the image objects, their properties, and/or their relationships. 1077 caption candidates were recognized as captions for a precision of 0.273 at perfect recall. For comparison, [4] reports a precision of 0.014 for text queries to a standard browser that tried to find pages with images matching particular words.

Recall is harder to estimate, but we got 0.97 in a manual inspection of 20 random ones of these Web pages; we defined recall as the fraction of the image-describing text on these pages that was found by our page scanner. The missed caption text was in paragraphs insufficiently related to corresponding images. Our program labeled 6.28% of the total characters in the training set as part of captions, thus reducing the data by a ratio of 16 to 1 while only hurting caption recall by 3%. 24.7% of the images had at least one proposed caption; 37.3% of the images were excluded on size or aspect ratio, 5.7% were excluded because they appeared on three or more Web pages, 3.5% were excluded because they appeared twice on the same Web page, and the remaining 28.8% had no qualifying captions. Only around 1% of all descriptively captioned images were incorrectly excluded by these three criteria, so recall for these filters was 99%. As for precision, 69% of the images proposed in image-caption pairs were confirmed as having at least one caption. Execution time for the crawler and filter averaged around five seconds per page on a 300 megahertz Pentium PC, but this varied widely with site.

3. INCREASING THE PRECISION OF THE CRAWLER OUTPUT

The caption-candidate filtering described above only eliminates the obvious noncaptions. To obtain better precision and to enable ranking of caption candidates in answers to user queries, we need to assign likelihoods to candidates. We use a simple neural network with carefully-chosen factors.

3.1 Modeling the effect of caption clues

We used the training set, in which all captions are tagged, to identify positive and negative clues for captions. Clues can be the occurrence of specific words and caption attributes. The strength associated with clue i is the conditional probability that the clue occurs in a caption when it occurs, estimated by $r_{ci}/(r_{ci} + r_{wi})$ where r_{ci} is the number of captions containing clue i and r_{wi} is the number of noncaptions containing clue i in the training set. A clue can also be negative, so that its absence from a caption is a clue that we have a caption, but we did not find this generally helpful. Clue occurrence can be modeled as a binomial process, and our approach was to say a clue is statistically significant if it exceeds the binomial distribution prediction by more than one standard

deviation in either direction, or (where n_c is the number of captions and n_u is the number of noncaptions):

$$| r_{ci} - ((n_c / n_u) r_{ui}) | > \sqrt{r_{ui} r_{ci} / (r_{ci} + r_{ui})}$$

Nonlinear functions were applied to the factors so that their median value was roughly 0.5 and standard deviation was roughly 0.15. For a total caption rating from a set of clues we use a linear model where we take a weighted sum of the adjusted likelihoods of all clues. Linear models can be contrasted with Naïve-Bayes and association-rule methods, and are appropriate when clues are strongly correlated [8] as are many caption word clues. Linear models also are preferable here to decision trees since there are unlikely to be complex logical relationships between clues, and are preferable to case-based reasoning since there is no small set of "ideal" captions.

3.2 Clues from specific words in the caption

Good clues as to whether a candidate is a caption are from the occurrence of specific words in the candidate string. So we tabulated word counts for the training set and calculated the associated conditional probabilities. 0.273 was the expected value in the training set, so only words deviating more than one standard deviation from this value in either direction were used. Some word clues found in the training set suggested were valid for the Web in general (like "gif", "center", and "photograph") but others reflected unrepresentative phenomenon in a small sample of the Web (like "child" and "destroyer") and needed to be diluted by more data. The total assessment of the word clues in a caption was $\exp((\sum_{i=1}^{M} (q_i - \overline{q})) / M)$ where M is the number of word clues, q_i is the conditional probability for word i of the caption, $\overline{q}$ is the fraction of captions in the training set, and exponentiation used to keep the result positive.

Destemming words first is important for word clues because related forms often occur in natural languages, like "picture", "pictures", "pictured", "picturing", and "picturedly" in English. We developed a destemmer using details from [9] but improved to cover important cases it missed like "ier", "edly", "ity", and "tionism" endings, and the necessary irregular forms (422 words and 1002 intermediate forms) that it did not enumerate. We improved it using a Unix "spell" utility dictionary of 28,806 common English words, mapping them first through the Wordnet thesaurus system to eliminate around 4,000, and then manually inspecting fifty separate classes of endings to eliminate around another 4,000 words. This gave 19,549 words which we supplemented with 674 technical words from computer-science papers and words from the training set that were incorrectly destemmed (261, mostly proper nouns ending in "s"). The final lexicon was 20,223 words.

3.3 Other text clues

The type of the caption is a good clue, both positively and negatively. Table 1 shows the statistics on the training set. No types are certain to be captions; even "alt" strings can just be a useless word like "photo".

Table 1: Statistics showing likelihood of a caption given candidate type in the training set.

Caption candidate type	Number that were captions	Number in training set	Probability	Significant?
i (italics)	2	5	0.40	no
b (boldface)	24	67	0.36	no
em (emphasis)	0	1	0.00	no
strong	1	15	0.07	no
big	1	4	0.25	no
font	45	120	0.37	yes
center	4	63	0.06	yes
td (table datum)	90	193	0.47	yes
tr (table row)	141	352	0.40	yes
caption	0	0	--	--
object	0	0	--	--
h1 (heading font 1)	5	15	0.33	no
h2	63	129	0.49	yes
h3	2	39	0.05	yes
h4	0	2	0.00	no
h5	0	7	0.00	no
h6	0	1	0.00	no
title	320	936	0.34	yes
alt (substitute text)	119	481	0.25	no
a (dynamic link)	97	149	0.65	yes
filename (of image)	42	1143	0.04	yes
wording	21	45	0.47	yes

The words of the image file path also furnish clues; 67 negative (e.g. “button”) and 10 positive (e.g. “media”) such clues were found. Powerful clues are the occurrence of the same word in both the caption and image file name, as for the image "http://www.nps.navy.mil/hermann_hall.gif" and caption "View of Hermann Hall". Another good clue is the format of the image referred to by a caption, since 53.6% of candidate captions on JPEG images were valid in the training set and 16.2% of those on GIF images. Other useful clues are digits in the image filename (images important enough to be captioned are often numbered), sentence length, and distance of the caption from its associated image. We also explored several formulations of a “template-fit” clue that measured how common that kind of caption and its placement (above or below the image) were for other pages on its site, but it was not sufficiently reliable to help.

3.4 Deciding whether an image is a photograph or graphics

[4] and [6] suggest it helps to know if an image is a photograph or graphics: A sample of our training set showed that 95% of the photographs had captions whereas 10% of the nonphotographs had captions. Both can be stored in similar image formats, so some content analysis is necessary to confirm photographs. We followed the linear model of MARIE-3 but with parameters the size of the image as measured by the length of the diagonal; the number of color bins having at least one associated pixel, for 256 bins evenly distributed in intensity-hue-saturation space; the count in the color bin having the most associated pixels; the average "saturation"; the average color variation between neighboring pixels as measured in intensity-hue-saturation space; and the average brightness variation between neighboring pixels. Again, nonlinear functions were applied to the factors to adjust their scales.

Figure 2 shows precision versus recall for discriminating photographs, for the six factors and their weighted average on the 648 photographs and 309 nonphotographs in the training set (excluding those that had become unavailable). We optimized to find the best weightings of the factors. The fifth and sixth factors are clearly negative influences, but we could get no improvement by assigning them negative weights. So the "Best weighted average" in Figure 2 represents the best weighted sum with the first four factors, for which we got 93.4% precision at 50% recall on the training-set images, insufficiently better than using the size factor alone. In addition, size is easier to compute than the other image properties since it can be extracted from the image-file header without any image processing. So we used only the size factor in subsequent assessment of caption candidates.

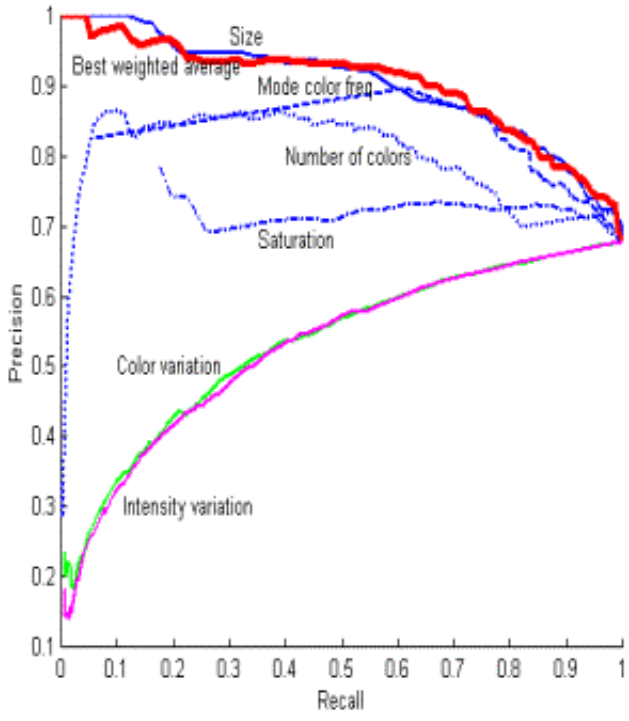

Figure 2: Precision versus recall on the training set for photograph discrimination using six factors.

3.5 Putting all the caption clues together

Finally, we implemented our linear model (a simple neural network) and added all the caption clue strengths with that of the image-size clue, and rated the likelihood of being a caption for each candidate in the training set. It appears eight of the nine factors are helpful (see Figure 3), with the exception of the distance between the caption and the image. (Recall in Figure 3 is for only the results of caption-candidate rater, and should be multiplied by 0.97 to get the total recall.) We obtained weightings for the eight factors by both least-squares linear regression and steepest-ascent optimization on the training set, and the latter weightings were better, but only 2% better than weights of 0.1 except for 0.3 for the caption-word factor. This unimpressive improvement suggests a danger of overtraining and argues against use of a more complex neural network.

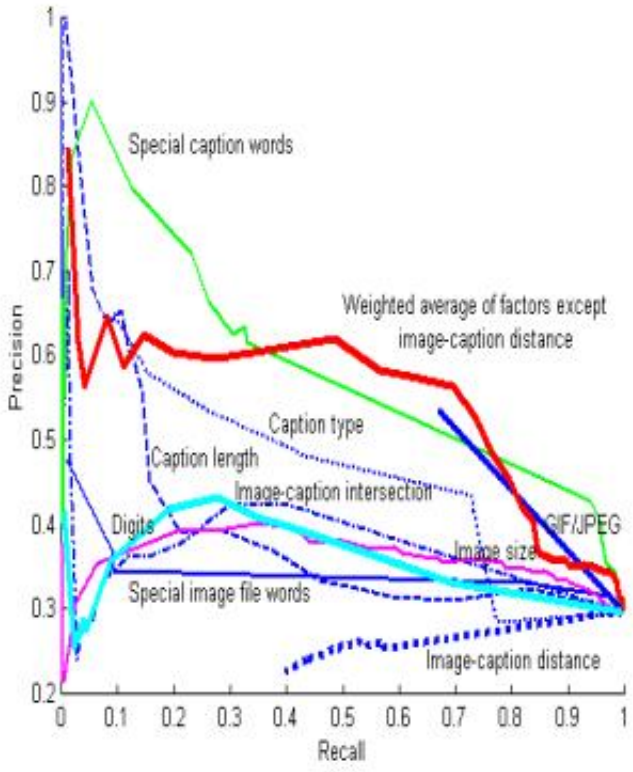

Figure 3: Precision versus recall on the training set for caption discrimination using nine factors.

Surprisingly, image size did not have much effect. Here we are assessing caption candidates, not captionable images, and other factors matter more; and we have excluded the smallest and most asymmetric images by our earlier expert system, something not done for the Figure 2 experiments. The "special image words" factor appears unhelpful, but this is misleading: Only a few image file names had

special word clues, but when they occurred, there was a clear advantage to exploiting them. On the other hand, the distance between the caption and image is clearly unhelpful due to the many "filename" and "alt" candidates at distance 0 that are not captions; this questions the reliance on it in [6].

3.6 Testing the caption rater

To test the caption rater we modified the crawler to more randomly sample Web pages. This is harder than it might seem, as there are now an estimated three billion Web pages, and the loose organization of the Web precludes any easy way to choose a "random" page. So we started with 10 representative pages (not necessarily those with many images as with the training set) and performed a random search to retrieve 600 pages starting from each of them. For a more depth-first search (1) only two random links on each page (not necessarily links to its site) were used to find new pages, and (2) one random caption-image pair for each page was selected. This encouraged exploration as the search starting with a metallurgy-journal site spent much time on country-music sites and the search starting with a fashion site spent much time on sports-news sites. But this did bias search toward sites with many links to them, which raises ethical questions like those of the popularity-weighting Google search engine.

Testing found 2,024 caption-image pairs for 1,876 images. The number found per Web page varied from 0.17 to 16.71 over the ten runs. Captionable images were fewer than small graphical icons; captions themselves are not routine even on captionable images. The caption-image pairs were then tagged by the author as to whether they were captions. The fraction of captioned images per site varied widely, from 0.020 (www.nytimes.com) to 0.260 (www.amazon.com) to 0.464 (www.arabfund.org) to 0.843 (www.charteralaska.net) to 0.857 (www.kepnerfamily.com). The clues proposed were confirmed as helpful for this test set; other clues found were whether the page name ended in "/" (negative) and whether the site name ended in ".mil" (positive). Example confirmed clue words in image-file names were "logo", "icon", "adobe", and "service" (negative), and "people", "library", and "photo" (positive). Example confirmed word clues in captions with their caption probabilities were "update" (0.000), "thumbnail" (0.000), "download" (0.029), "customer" (0.038), "week" (0.780), "forward" (0.875), and "photographer's" (0.960).

To test whether our caption rater could learn from experience using statistics on caption words, image-file words, and caption types, we rather further tests with a second tagged test set of 2148 caption candidates on 1577 images obtained from the crawler by the random search starting on 16 additional sites. We then rated the captions using four sets of probabilities obtained from statistics. Version 1 used no statistics but did use the image-clue words from MARIE-3; version 2 used statistics from just the training set; version 3 used statistics from both the training set and the first (2024-pair) test set, with 75% of the filename and title candidates eliminated to provide a better balance among caption types; and version 4 used artificially tagged data discussed below. Figure 4 shows the results for precision versus recall, demonstrating a clear advantage of more knowledge, excepting greater random fluctuations at low values of recall (with small sample sizes). A healthy steady increase in precision occurred as recall decreased, and

no significant differences in the shape of the curve were observed on any major subsets of the test data.

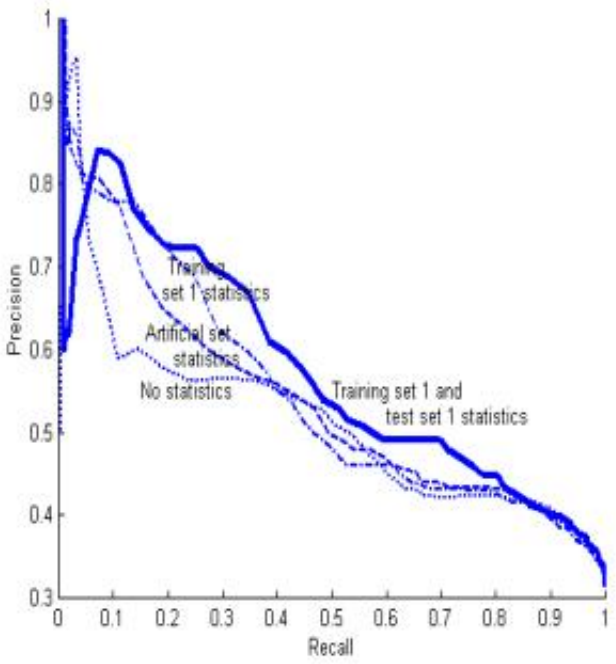

Figure 4: Precision versus recall for four versions of the caption rater, illustrating its learning from experience.

Ultimately the system should tag obvious captions itself to provide further statistics. The dashed line in Figure 4 illustrates this on 40,239 candidates from the crawler, a superset from which we derived the training and test sets. We rated these candidates using statistics of the training and first test sets, assumed the top 10% were captions to derive a new set of statistics, then reran the second test set with the guidance of the new statistics. (The top 10% gave 80% precision on the second test set, so the new statistics should be roughly 80% correct.) Although performance was not as good as for the manual-tag statistics, this approach can be improved with smarter tagging.

4. THE QUERY INTERFACE

The words of all proposed captions found by the page scanner are indexed. The index is used by keyword-lookup Java servlets that run on our Web site. The user enters keywords for the images which they seek and specifies how many answers they want. The servlet destems the user's keywords, uses its index of destemmed words to find images matching at least one keyword, ranks the matches, and displays the images and captions of the best matches. The user can click on links to go to the source Web pages. Figure 5 shows example output for a database of all images on Web pages at our school. Table 2 shows some statistics on a more ambitious project analyzing and indexing the 667,573 images we found on all 574,887 publicly-accessible U.S. Navy (or “navy.mil”) sites using our single 300-megahertz PC. The servlets takes around 15 seconds to load on a Unix server machine, and 10 to 90 seconds to answer

a typical three-word query. The servlets are accessible at http://triton.cs.nps.navy.mil:8080/rowe/rowedemos.html , including a companion servlet built with many of the same principles that indexes all Navy audio and video clips.

Match ranking exploits additional factors besides the overall caption likelihood. Following the recommendations of [10] for short queries and short independent documents like captions, we should add weights for all matching keywords, and an inverse document-frequency factor should ensure a higher weight on rarer keywords. In addition, for a random sample of 363 true captions from our full training set, the probability of a keyword being depicted in the image decreased steadily from 0.87 for 3-word captions to 0.24 for 90-word captions; by "depict" we mean correspond to some area of the image. (Note that length has the opposite effect for keyword-match rating than it does for ascertaining caption likelihood.) The probability of a word being depicted in the image also steadily decreased as a function of relative position in the caption, from 0.68 for words in the first 10% of the caption to 0.15 for words in the last 10%. This is because long captions tend to include background material toward the end, more so than other kinds of text. So we did least-squares fitting for these factors from the sample.

For the overall weight, we use predominantly a Naive-Bayes approach (since the factors are close to independent) where we sum the products of the factors for each keyword (since we expect the keywords to be correlated). Minor factors for capitalization matching and keyword adjacency in the caption were added to the total with small-scale factors as with many current Web search engines. So the weight on caption-image pair i is:

$$(c_i((-0.176*\ln(k))+0.968)\sum_{j=1}^{m}[\ln(N/n_j)*((-2.33*p_j)+2.717)])+(0.1*c_i)+(0.1*a_i)+(0.05*b_i)$$

where c_i is the likelihood the caption describes the image, k is the number of non-stopwords in the caption, j is the index number of a keyword, m is the number of keywords, N is the number of captions, n_j is the number of captions containing keyword j after destemming, p_j is the fraction of the distance through the caption that keyword j first appears, c_i is the number of capitalized keywords that exactly match capitalized caption words, a_i is the number of keywords that appear adjacently in the caption, and b_i is the number of keywords that appear separated by a single word in the caption. To test the formula, we generated 32 three-keyword queries by choosing 150 random caption candidates and picking three representative keywords from each of those that were true captions. In 22 of the 32 cases, the above formula gave better answers than a control formula using only the caption-likelihood and document-frequency factors; in 9 cases the answers were the same; and in 1 case they were worse.

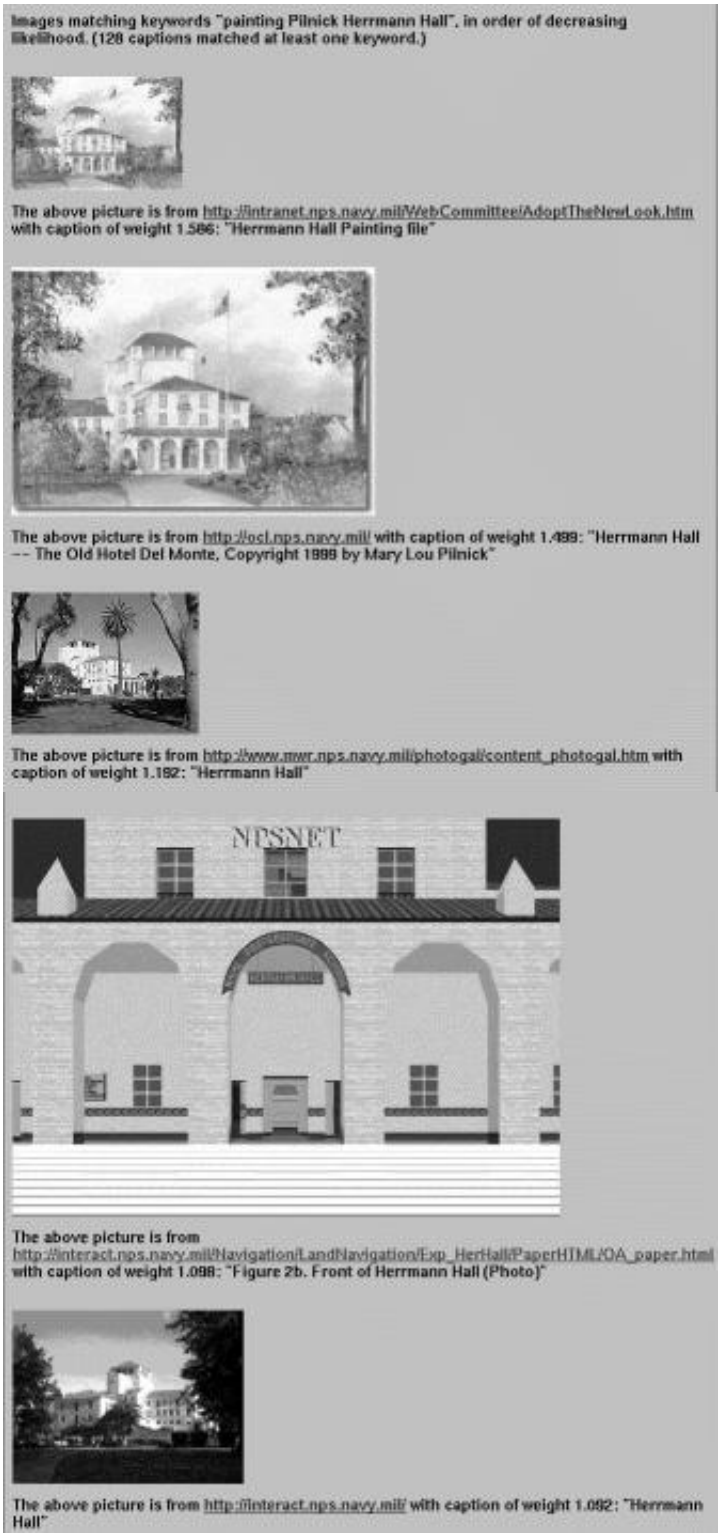

Figure 5: Example use of the query interface, showing the best five candidates found for the query "painting Pilnick Herrmann Hall".

Table 2: Statistics on the building of the MARIE-4 servlet that indexes all images at navy.mil sites.

Number of items found	Megabytes for result	Computation real time (minutes)	Description
6,002,295	1468.4	circa 13,000	Initial page scan (in which 574,887 Web pages were examined)
2,198,549	582.4	860	Checking for the existence of image files, retrieving the size of those not described on the Web page, excluding captions on too-small images, and removing images with too many references
2,198,549	897.1	130	Rating of caption candidates
211,398	462.0	197	Indexing of caption candidates (by root words)
211,398	3.8	--	Main-memory hash table for the servlet to the secondary-storage index
85,124	5.5	--	Text of all distinct Web-page links for captions
667,573	67.4	--	Text of all distinct image-file links for captions
2,193,792	124.0	--	Text of all distinct captions

5. CONCLUSIONS

The diversity of the Web requires automated tools to find useful information. But this very diversity means the tools must have some intelligence to cope with all the different formats they find. We have shown that the seeming wide diversity of image formats on the Web can be substantially indexed with our tool. Careful tests on 8140 caption candidates for 4585 representative images have confirmed the factors we use and how they are combined. But this comprehensive approach does require a spectrum of methods be used, not just one method, and learning from experience must play an important role.

ACKNOWLEDGEMENTS

Jorge Alves, Vanessa Ong, and Nickolaos Tsardas built prototypes of several software modules.

REFERENCES

[1] Atsuo Yoshitaka and Tadao Ichikawa, "A Survey on Content-Based Retrieval for Multimedia Databases," *IEEE Transactions on*

Knowledge and Data Engineering, vol. 11, no. 1, pp. 81-93, January/February 1999.

[2] C. Jorgensen, "Attributes of images in describing tasks," *Information Processing and Management*, vol. 34, no. 2/3, pp. 161-174, 1998.

[3] R. K. Srihari, "Use of captions and other collateral text in understanding photographs," *Artificial Intelligence Review*, vol. 8, no. 5-6, pp. 409-430, 1995.

[4] N. C. Rowe and B. Frew, "Automatic caption localization for photographs on World Wide Web pages," *Information Processing and Management*, vol. 34, no. 1, pp. 95-107, 1998.

[5] Sougata Mukherjea and Junghoo Cho, "Automatically determining semantics for World Wide Web multimedia information retrieval," *Journal of Visual Languages and Computing*, vol. 10, pp. 585-606, 1999.

[6] S. Sclaroff, M. La Cascia, S. Sethi, and L. Taycher, "Unifying textual and visual cues for content-based image retrieval on the World Wide Web," *Computer Vision and Image Understanding*, vol. 75, Nos. 1/2, pp. 86-98, July/August 1999.

[7] N. C. Rowe, "Precise and efficient retrieval of captioned images: The MARIE project," *Library Trends*, vol. 48, no. 2, pp. 475-495, Fall 1999.

[8] I. Witten and E. Frank, *Data Mining: Practical Machine Learning with Java Implementations*. San Francisco, CA: Morgan Kaufmann, 2000.

[9] M. F. Porter, "An algorithm for suffix stripping," *Program*, vol. 14, pp. 130-137, 1980.

[10] G. Salton and C. Buckley, "Term-weighting approaches in automatic text retrieval," *Information Processing and Management*, vol. 24, pp. 513-523, 1988.